Limits: Journal of Mathematics and Its Applications
E-ISSN: 2579-8936
P-ISSN: 1829-605X
Vol. 22, No. 3, Nopember 2025, 135-151
DOI: https://doi.org/10.12962/limits.v22i3.3118

Evaluasi Kinerja Metode CLARA dan FCM dalam Analisis Gerombol untuk Data Berjumlah Besar dengan Pencilan

Indahwati [1]*, Intan Juliana Panjaitan [2], Farit Mochamad Afendi [3]
[1,2,3]Program Studi Statistika dan Sains Data; IPB University, Indonesia
e-mail: indahwati@apps.ipb.ac.id

Diajukan: 14 Maret 2025, Diperbaiki: 29 April 2025, Diterima: 15 Juli 2025

Abstrak

Analisis gerombol adalah suatu metode statistika yang mengidentifikasi gerombol objek berdasarkan karakteristik serupa. Masalah yang sering terjadi dalam analisis gerombol adalah keberadaan data pencilan. Keberadaan pencilan dapat mengakibatkan output yang tidak sesuai dengan gambaran yang sebenarnya, sehingga gerombol yang dihasilkan tidak merepresentasikan objek dengan tepat. Masalah lain yang dapat muncul dalam analisis gerombol adalah besarnya jumlah amatan, sehingga diperlukan metode analisis yang efisien dalam penggerombolan. Penelitian ini juga memperdalam tentang kinerja keduanya terhadap jarak antara pusat gerombol dan kondisi penggerombolan melalui kajian simulasi, dimana masing-masing faktor terdiri dari tiga level yang diobservasi. Metode *Clustering Large Applications* (CLARA) dan *Fuzzy C-Means* (FCM) adalah metode yang kekar (*robust*) terhadap pencilan dan mampu menganalisis dataset besar. Metode FCM menggunakan nilai pembobot (w) yang optimal agar kekar terhadap pencilan. Metode CLARA memiliki sifat kekar dikarenakan menggunakan medoid sebagai pusat gerombol dan penggunaan jarak Manhattan dalam perhitungan jarak antara objek dan pusat gerombol. Metode tersebut akan dievaluasi menggunakan beberapa kriteria evaluasi kebaikan yaitu berdasarkan akurasi penggerombolan serta rasio simpangan baku dalam gerombol dan antar gerombol. Hasil analisis menunjukkan pengaruh signifikan pada masing-masing faktor dan interaksi antar faktor. Visualisasi menunjukkan bahwa peningkatan persentase pencilan mengurangi akurasi penggerombolan, sementara jumlah data yang lebih besar meningkatkan akurasi. Jarak yang lebih besar antara pusat gerombol dan kondisi gerombol yang terpisah menghasilkan rasio simpangan baku gerombol yang lebih kecil. Hasil penelitian menunjukkan bahwa metode FCM lebih efektif dalam menangani data dengan variasi yang signifikan.
Kata Kunci: Gerombol, CLARA, FCM, Pencilan

Abstract

Cluster analysis is a statistical method that identifies clusters of objects based on similar characteristics. A common issue in cluster analysis is the presence of outliers. Outliers can lead to outputs that do not reflect the actual data distribution, resulting in clusters that fail to represent the objects accurately. Another issue in cluster analysis is the large number of observations, necessitating efficient methods for clustering. This study also explores the performance of both methods in terms of the distance between cluster centers and clustering conditions through simulation studies, where each factor comprises three observed levels. The Clustering Large Applications (CLARA) and Fuzzy C-Means (FCM) methods are robust against outliers and capable of analyzing large datasets. The FCM method uses an optimal weighting value (w) to achieve robustness against outliers. The CLARA method is robust because it employs medoids as cluster centers and utilizes Manhattan distance to calculate the distance between objects and cluster centers. These methods are evaluated using several goodness-of-fit namely based on clustering accuracy as well as the ratio of standard deviation within and between clusters. The analysis results show significant effects for each factor and interactions between factors. Visualization reveals that increasing the percentage of outliers reduces clustering accuracy, whereas a larger dataset size improves accuracy. Greater distances between cluster centers and more separated cluster conditions result in lower within-

cluster standard deviation ratios. The study concludes that the FCM method is more effective in handling data with significant variation.
***Keywords**: Clustering, CLARA, FCM, Outliers*

1 Pendahuluan

Analisis gerombol adalah suatu metode statistika yang mengidentifikasi gerombol objek berdasarkan karakteristik serupa. Analisis ini menggerombolkan elemen mirip sebagai objek penelitian. Elemen-elemen tersebut memiliki tingkat homegenitas yang tinggi antar objek dan menjadi gerombol yang berbeda dengan tingkat heterogenitas tinggi antar gerombol [1]. Terdapat dua pendekatan dalam analisis gerombol, yaitu pendekatan hirarki dan pendekatan tak berhirarki (partisi). Pendekatan hirarki menggerombolkan objek pengamatan secara terstruktur berdasarkan sifat kemiripan objek, sedangkan pendekatan tak berhirarki menggerombolkan objek ke dalam gerombol-gerombol yang sudah ditentukan sebelumnya oleh peneliti [2].

Kendala yang sering terjadi saat analisis gerombol adalah keberadaan pencilan dalam data. Istilah pencilan mengacu pada nilai amatan data yang menyimpang secara signifikan dari nilai umum amatan lainnya [3]. Keberadaan pencilan tersebut dapat menyebabkan hasil analisis yang tidak sesuai dengan gambaran sebenarnya, sehingga gerombol yang dihasilkan tidak merepresentasikan objek dengan tepat [4]. Penanganan kendala yang paling mudah adalah dengan membuang data pencilan tersebut, tetapi hal itu bukanlah solusi terbaik. Penanganan lain diperlukan untuk mencari metode alternatif dalam mengatasi keberadaan pencilan tanpa harus membuangnya. Metode Clustering Large Application (CLARA) memiliki sifat kekar (*robust)* sehingga dapat mengurangi dampak dari keberadaan pencilan [5]. Metode Fuzzy C-Means (FCM*)* juga merupakan metode yang kekar terhadap keberasaan pencilan dengan menggunakan nilai pembobot (w) yang optimal [6]–[8]. Selain itu, metode FCM terbukti lebih unggul dalam mengatasi data pencilan dibandingkan metode K-Means [9], [10].

Kendala lain yang dapat muncul dalam analisis gerombol selain keberadaan pencilan adalah banyaknya jumlah amatan yang jikalau tidak menggunakan metode yang tepat, membuat tidak efisien dalam proses waktu komputasi, sehingga diperlukan metode analisis yang efisien dalam penggerombolan. Metode FCM dapat diimplementasikan untuk menggerombolkan dataset dengan jumlah amatan yang relatif banyak [11]. Metode CLARA dan FCM juga mampu menganalisis dataset dengan jumlah amatan yang relatif banyak [12]. CLARA adalah metode tak berhirarki yang merupakan pengembangan dari metode PAM (Partitioning Around Medoid) yang memiliki sifat kekar terhadap keberadaan pencilan dan dapat digunakan pada data jumlah besar. CLARA merupakan metode yang berbasis *sampling* yang memanfaatkan algoritma K-Medoid pada

beberapa data sampel sehingga waktu komputasi semakin efisien. Metode FCM adalah metode tak berhirarki di mana keberadaan setiap amatan dalam suatu gerombol ditentukan oleh derajat keanggotaan. Metode FCM sering digunakan untuk melakukan penggerombolan karena mampu menangani dataset dengan jumlah amatan yang relatif banyak, kekar terhadap pencilan, serta memberikan hasil yang lebih tinggi dibandingkan dengan metode K-Means [13]. Selain itu, metode ini juga menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan metode K-Means [14].

Penelitian ini mengevaluasi kinerja analisis metode CLARA dan FCM dalam menggerombolkan data jumlah amatan banyak dan menggandung pencilan. Meskipun FCM dan CLARA merupakan metode yang sudah lama ada, mereka masih tetap relevan dan penting untuk dibandingkan karena masing-masing menawarkan kelebihan dalam menangani permasalahan dalam analisis gerombol. Kestabilan yang baik, kemudahan implementasi dan interpretasi hasil, serta kinerja metode yang baik menjadi alasan umum mengapa peneliti dan praktisi tetap memilih metode-metode ini untuk melakukan analisis gerombol. Kinerja kedua metode tersebut nantinya akan di evaluasi menggunakan beberapa kriteria evaluasi kebaikan yakni rasio simpangan baku dalam dan antar gerombol. Semakin kecil nilai rasio simpangan bakunya, maka semakin baik hasil analisis dari metode tersebut. Kriteria evaluasi kebaikan selanjutnya adalah akurasi metode terhadap gerombol awal yang keanggotaan gerombolnya sudah ditentukan. Evaluasi tersebut bertujuan untuk melihat apakah rasio terkecil yang diperoleh juga memberikan keakuratan terbesar.

2 Metode Penelitian

2.1 Data

Data yang digunakan pada penelitian ini merupakan data bangkitan dengan skenario yang telah dirancang. Proses simulasi juga dilakukan dengan menggunakan jumlah data sebanyak 100 amatan, 500 amatan, dan 1000 amatan. Pemilihan nilai-nilai tersebut secara berturut-turut dimaksudkan untuk mewakili ukuran data kecil, sedang, dan besar [15]. Persentase pencilan yang digunakan sebesar 0%, 10%, dan 20% [16]. Penetapan nilai-nilai persentase pencilan tersebut dipilih dengan alasan bahwa skenario persentase 0% mewakili kondisi tidak ada pencilan, persentase sebesar 10% mewakili persentase pencilan sedang, dan 20% mewakili persentase pencilan tinggi Kontaminasi pencilan nantinya akan diterapkan pada satu peubah saja.

Gerombol akan dibangkitkan dalam tiga kondisi, yaitu ketiga gerombol saling terpisah, satu gerombol terpisah dan dua gerombol tumpang tindih, serta ketiga gerombol saling tumpang tindih. Membangkitkan banyaknya pengamatan yang tumpang tindih dicobakan dengan tiga jenis ukuran

jarak antara dua nilai tengah (pusat) gerombol, yang disesuaikan dengan jauh dekatnya jarak antara vektor rataan gerombol (*d*). Hal ini didasari pemikiran bahwa semakin dekat jarak antara kedua pusat gerombol, semakin banyak pengamatan yang tumpang tindih. Sebaliknya semakin jauh jarak antara kedua pusat gerombol, semakin sedikit pengamatan yang tumpang tindih. Skenario jarak antar pusat gerombol akan dibangkitkan menjadi jarak dekat, sedang, dan jauh sebagai berikut:

jarak dekat	jarak sedang	jarak jauh
$\mu_1 = (4\ 1\ 3)$	$\mu_1 = (5\ 2\ 5)$	$\mu_1 = (6\ 1\ 5)$
$\mu_2 = (4\ 5\ 3)$	$\mu_2 = (5\ 7\ 4)$	$\mu_2 = (6\ 7\ 3)$
$\mu_3 = (4\ 5\ 7)$	$\mu_3 = (5\ 7\ 10)$	$\mu_3 = (6\ 7\ 11)$

Pemilihan kondisi tumpang tindih, jarak, nilai tengah, dan ragam pada proses simulasi ini didasarkan dari ide penelitian yang dilakukan Timbul Pardede pada tahun 2012, dengan modifikasi skenario adanya pencilan dan jumlah amatan. Ragam untuk skenario ketiga gerombol saling tumpang tindih adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, untuk skenario ketiga gerombol saling terpisah yaitu $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$, dan untuk skenario satu gerombol terpisah dan dua gerombol tumpang tindih adalah $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$. Total kombinasi skenario pada tahap simulasi yang akan dibangkitkan adalah sebanyak 81 kombinasi dan akan dilakukan pengulangan sebanyak 100 kali untuk semua skenario bangkitan data, sehingga akan diperoleh 100 nilai rasio simpangan baku untuk masing-masing skenario. 100 nilai rasio simpangan baku tersebut akan dihitung kembali nilai rata-ratanya sebagai perbandingan kinerja metode.

2.2 Tahapan Penelitian

Tahapan penelitian pada data simulasi yaitu menentukan jumlah gerombol (*k)* yang terbentuk sebanyak 3 gerombol. Selanjutnya membangkitkan data menyebar normal, 3 peubah dengan 3 gerombol yang memiliki 4 skenario simulasi. Lakukan pengulangan membangkitkan data sebanyak 100 kali dan hitung rata-rata nilai rasio simpangan baku dari 100 kali ulangan. Melakukan penggerombolan menggunakan metode CLARA dan FCM menggunakan jarak Manhattan. Algoritme *Fuzzy C-Means* dijelaskan pada uraian berikut :

1. Menentukan jumlah gerombol yang akan dibentuk (*k*) dengan syarat (*k*≥2), nilai pembobot (*w*) dengan syarat *w* >1.
2. Membangkitkan bilangan acak dengan syarat pada persamaan berikut :

$$\sum_{p=1}^{k} u_{ip} = 1 \quad (1)$$

dimana u_{ip} adalah bilangan acak ke-*i* pada gerombol ke-*p*

3. Menghitung pusat gerombol untuk setiap gerombol dengan persamaan

$$c_{pj} = \frac{\sum_{i=1}^{n} (u_{ip}{}^{w}\, x_{ij})}{\sum_{i=1}^{n} (u_{ip})^{w}} \quad (2)$$

dimana, c_{pj} adalah pusat gerombol ke-*p* pada peubah ke-*j,* u_{ip} adalah bilangan acak ke-*i* pada gerombol ke-*p,* x_{ij} adalah objek ke-*i* yang telah distandarisasikan pada peubah ke-*j,* dan w adalah pangkat pembobot.

4. Menghitung nilai derajat keanggotaan setiap objek pada setiap gerombol dengan persamaan berikut:

$$u_{ip} = \left[\frac{d(x_i,c_p)^{\frac{-2}{w-1}}}{\sum_{p=1}^{k} d(x_i,c_p)^{\frac{-2}{w-1}}}\right] \quad (3)$$

dengan

$$d(x_i,c_p) = \left[\sum_{j=1}^{l} d(x_{ij}-c_{pj})^2\right]^{\frac{1}{2}}$$

Dimana, c_{pj} adalah pusat gerombol ke-*p* pada peubah ke-*j,* u_{ip} adalah nilai derajat keanggotaan objek ke-*i* dari gerombol ke-*p,* x_{ij} adalah objek ke-*i* yang telah distandarisasikan pada peubah ke-*j,* w adalah pangkat pembobot, dan $d(x_i,\ c_p)$ adalah jarak antara objek ke-*i* dengan pusat gerombol ke-*p.*

5. Menempatkan objek ke pusat gerombol yang ada dengan melihat jarak terdekat berdasarkan persamaan berikut:

$$w_{ip} = \begin{cases} 1, & u_{ip} = \max(u_{i1}, u_{i2}, \dots, u_{ip}) \\ 0, & lainnya \end{cases} \quad (4)$$

6. Kembali ke langkah 3, 4, dan 5 sampai pusat gerombol tidak berubah lagi dan tidak ada anggota yang berpindah ke gerombol lainnya [17].

Iterasi dihentikan ketika $max_{ip}\{|u_{ip}^{k+1} - u_{ip}^{k}|\} < \varepsilon$, dengan nilai ε biasanya ε=1×10^{-3} [18].

Algoritma CLARA dijelaskan pada uraian berikut :

1. Menentukan jumlah gerombol (k)
2. Mengambil sampel dengan ukuran 40 + 2k secara acak dari dataset dan sampel tersebut kemudian digerombolkan menjadi k gerombol menggunakan metode K-Medoid yang juga memberikan k objek perwakilan gerombol (Medoid)
3. Menghitung jarak setiap objek terhadap medoid awal
4. Mengalokasikan setiap objek ke suatu gerombol terhadap medoid terdekat
5. Menentukan calon medoid baru pada setiap gerombol secara acak
6. Menghitung jarak setiap objek pada setiap gerombol dengan calon medoid baru
7. Mengalokasikan setiap objek ke suatu gerombol terhadap calon medoid baru terdekat
8. Menghitung simpangan (S) dengan menentukan selisih antara nilai total jarak objek ke calon medoid baru dan total jarak objek ke medoid awal. Jika diperoleh $S < 0$, maka ganti medoid awal dengan calon medoid baru sebagai medoid baru

9. Mengulangi langkah 5 hingga 8, dengan sedemikian sehingga diperoleh $S > 0$ atau medoid tidak mengalami perubahan, sehingga diperoleh gerombol beserta anggota gerombol masing-masing

Selanjutnya melakukan analisis interaksi antar faktor. Terakhir, membandingkan hasil penggerombolan metode CLARA dan FCM dengan membandingkan nilai rasio simpangan baku dalam gerombol (S_w) dan antar gerombol (S_b) terkecil serta akurasi terbesar. Simpangan baku dalam gerombol (S_w) dinyatakan dalam persamaan

$$S_w = \frac{1}{k}\sum_{p=1}^{k} S_p \tag{5}$$

dengan

$$S_p = \sqrt{\frac{\sum_{i=1}^{n_p}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2}{n_p - 1}}$$

dimana, S_w adalah simpangan baku didalam gerombol, S_p adalah simpangan baku gerombol ke-p, x_{ij} adalah objek pengamatan ke-i pada peubah ke-j, $\bar{x}_j$ adalah rata-rata objek pada peubah ke-j, dan n_p adalah banyaknya objek pada gerombol ke-p.

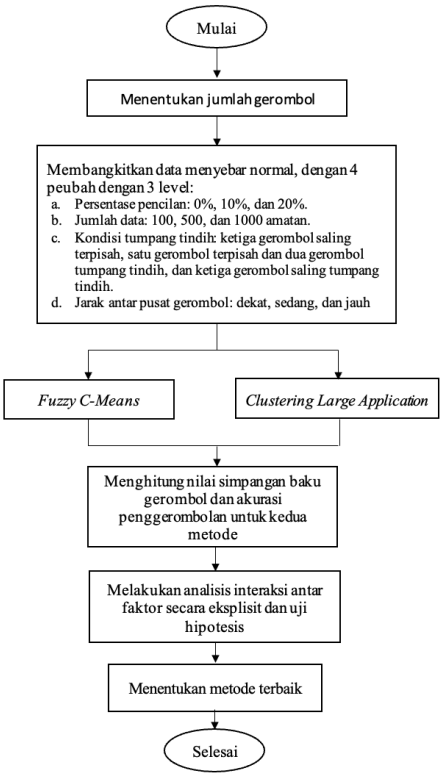
Simpangan baku antar gerombol (S_b) dinyatakan dalam persamaan berikut:

$$S_b = \sqrt{\frac{1}{k-1}\sum_{p=1}^{k}\left(\bar{x}_p - \bar{x}\right)^2} \tag{6}$$

dengan

$$\bar{x} = \frac{1}{k}\sum_{p=1}^{k} \bar{x}_p$$

dimana, k adalah banyaknya gerombol, S_b adalah simpangan baku antar gerombol, $\bar{x}_p$ adalah rata-rata gerombol ke-p, dan $\bar{x}$ adalah rata-rata objek pada keseluruhan gerombol. Tahapan pada penelitian ini secara singkatnya dapat dilihat pada Gambar 1.



Gambar 1. *Flowchart*

3 Hasil dan Pembahasan

3.1 Eksplorasi Data Simulasi

Eksplorasi terhadap data simulasi dilakukan untuk memastikan data bangkitan sesuai dengan skenario yang telah dirancang. Data yang dibangkitkan terdiri dari 81 kombinasi, dengan setiap kombinasi terdiri dari tiga gerombol. Kombinasi data simulasi tersebut dibedakan berdasarkan jumlah amatan (n), persentase pencilan ($\boldsymbol{\delta}$), jarak antar pusat gerombol, dan kondisi tumpang tindih antar gerombol. Visualisasi lengkap data yang telah dibangkitkan dapat dilihat pada lampiran 1. Hasil bangkitan data akan digunakan sebagai data awal, dengan asumsi bahwa penggerombolan dari skenario tersebut merupakan penggerombolan yang optimal dengan nilai rasio simpangan baku minimum. Data awal ini akan digunakan sebagai acuan untuk menghitung nilai akurasi dari hasil penggerombolan yang dilakukan oleh metode CLARA dan FCM. Hasil penggerombolan masing-masing metode dibandingkan berdasarkan rasio simpangan baku dan akurasi penggerombolan, yang secara lengkap dapat dilihat pada lampiran 2.

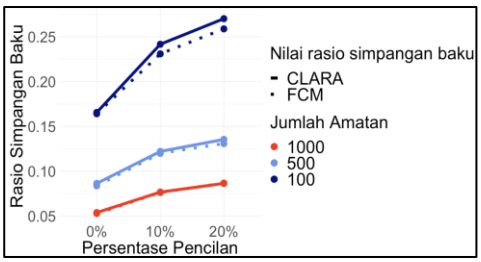
3.2 Evaluasi Interaksi Kondisi Gerombol

Hasil awal penggerombolan metode CLARA dan FCM dibandingkan berdasarkan rasio simpangan baku dan akurasi penggerombolan. Evaluasi ini bersifat eksploratif, bertujuan untuk

mengamati pengaruh interaksi antar faktor yaitu faktor persentase pencilan, jumlah amatan. jarak antar pusat gerombol, serta kondisi tumpang tindih antar gerombol

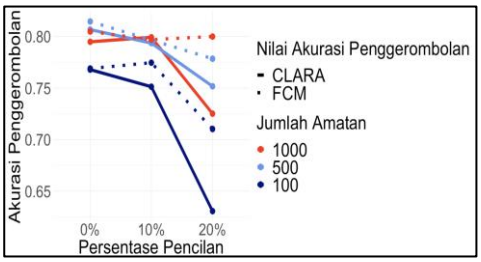
3.2.1 Perbandingan Skenario Persentase Pencilan dan Jumlah Amatan

Visualisasi nilai rasio simpangan baku terhadap skenario persentase pencilan dan jumlah amatan dapat dilihat pada Gambar 2. Semakin besar persentase pencilan maka nilai rasio simpangan baku juga semakin besar. Hal ini menunjukkan bahwa pencilan cenderung mengganggu pembentukan gerombol yang homogen. Pencilan ini menyebabkan gerombol menjadi lebih tidak stabil dan kurang jelas batasannya, karena data yang menyimpang dari pola umum membuat gerombol lebih kabur. Semakin banyak jumlah amatan maka semakin kecil rasio simpangan baku. Pada kombinasi jumlah amatan 1000, terlihat kedua metode beririsan, yang menandakan bahwa kedua metode menghasilkan nilai rasio simpangan baku yang tidak berbeda secara signifikan. Hal ini menunjukkan bahwa lebih banyak data membantu membentuk gerombol yang lebih stabil dan jelas. Banyaknya amatan membuat variasi dalam data dapat lebih mudah diidentifikasi, dan gerombol yang dihasilkan cenderung lebih akurat dalam mencerminkan pola-pola dalam data. Hal yang cukup menarik adalah semakin sedikit jumlah amatan dan semakin besar persentase pencilan, maka selisih perbandingan nilai rasio simpangan baku antar kedua metode akan semakin besar. Semakin besar jumlah amatan dan semakin kecil persentase pencilan, maka selisih perbandingan nilai rasio simpangan baku antar kedua metode akan semakin tidak jauh berbeda. Hal ini menunjukkan bahwa kedua metode mungkin berbeda dalam cara mereka menangani situasi ini.



Gambar 2. Plot perbandingan nilai rasio simpangan baku untuk skenario pencilan dan jumlah amatan terhadap metode FCM dan CLARA

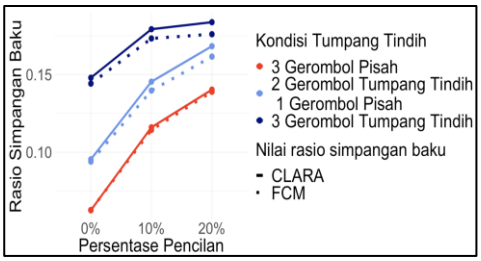
Visualisasi nilai akurasi penggerombolan dengan skenario persentase pencilan dan jumlah amatan dapat dilihat pada Gambar 3. Gambar 3 menunjukkan bahwa semakin besar persentase pencilan, maka nilai akurasi penggerombolan semakin kecil. Pada kombinasi jumlah amatan 1000 dan 500, nilai akurasi penggerombolan tidak terlalu berbeda secara signifikan pada persentase pencilan 10%. Nilai akurasi penggerombolan yang dihasilkan untuk jumlah amatan 100 berbeda signifikan antara kedua metode.



Gambar 3. Plot perbandingan nilai akurasi penggerombolan untuk skenario pencilan dan jumlah amatan terhadap metode FCM dan CLARA

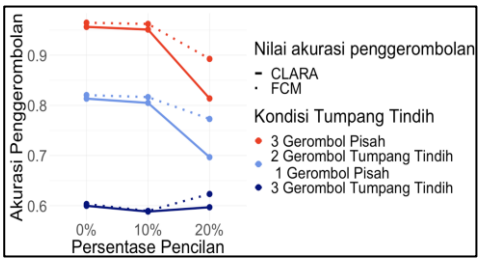
3.2.2 Perbandingan Skenario Pencilan dan Kondisi Tumpang Tindih

Visualisasi nilai rasio simpangan baku terhadap skenario persentase pencilan dan kondisi tumpang tindih dapat dilihat pada Gambar 4. Semakin besar persentase pencilan, maka nilai rasio simpangan baku juga semakin besar. Nilai rasio simpangan baku semakin kecil jikalau semakin terpisah gerombol. Hal yang menarik, untuk kondisi tiga gerombol pisah dengan nilai persentase pencilan 0% nilai rasio simpangan baku antar kedua metode beririsan dan sangat sedikit perbedaanya dibandingkan persentase pencilan 10% dan 20%.



Gambar 4. Plot perbandingan nilai rasio simpangan baku untuk skenario pencilan dan kondisi tumpang tindih terhadap metode FCM dan CLARA

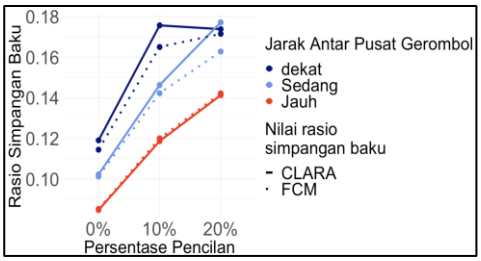
Visualisasi nilai akurasi penggerombolan dengan skenario persentase pencilan dan kondisi tumpang tindih dapat dilihat pada Gambar 5. Semakin besar persentase pencilan maka nilai akurasi penggerombolan semakin kecil terkecuali pada kondisi 3 gerombol saling tumpang tindih, sementara untuk skenario kondisi tumpang tindih, semakin terpisah gerombol maka semakin besar akurasi penggerombolan. Hal yang menarik adalah untuk persentase pencilan 0% dan 10%, perbedaan nilai akurasi penggerombolan antara kedua metode hampir berimpit untuk setiap kondisi tumpang tindih, yang mengartikan bahwa nilai perbedaan akurasi penggerombolan kedua metode tersebut tidak jauh berbeda. Perbedaan nilai akurasi pengerombolan cukup signifikan untuk nilai persentase pencilan 20%. Secara keseluruhan, untuk skenario kondisi tumpang tindih dan persentase pencilan, dapat dilihat bahwa metode FCM melakukan penggerombolan lebih baik, di mana setiap skenario menghasilkan nilai akurasi penggerombolan yang lebih tinggi dibandingkan dengan metode CLARA.



Gambar 5. Plot perbandingan nilai akurasi penggerombolan untuk skenario pencilan dan kondisi tumpang tindih terhadap metode FCM dan CLARA

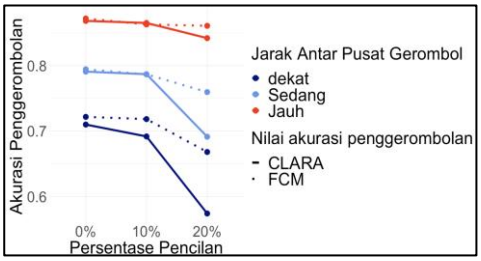
3.2.3 Perbandingan Skenario Pencilan dan Jarak Antar Pusat Geerombol

Visualisasi nilai rasio simpangan baku gerombol dengan skenario persentase pencilan dan jarak antar pusat gerombol dapat dilihat pada Gambar 6. Semakin besar persentase pencilan, maka nilai rasio simpangan baku semakin besar. Semakin jauh jarak antar pusat gerombol maka semakin kecil rasio simpangan baku. Hal yang menarik, untuk pencilan 20% pada jarak pusat gerombol sedang, perbedaan nilai rasio simpangan baku antar dua metode cukup berbeda, tidak seperti kondisi jarak dekat dan jauh pada persentase pencilan 20%.



Gambar 6. Plot perbandingan nilai rasio simpangan baku untuk skenario pencilan dan jarak antar pusat gerombol terhadap metode FCM dan CLARA

Visualisasi nilai akurasi penggerombolan dengan skenario persentase pencilan dan jarak antar pusat gerombol dapat dilihat pada Gambar 7. Semakin besar persentase pencilan, maka nilai akurasi penggerombolan semakin kecil. Semakin jauh jarak antar pusat gerombol maka semakin besar akurasi penggerombolannya. Hal menarik adalah pada persentase pencilan 20%, perbedaan nilai akurasi penggerombolan antara dua metode cukup signifikan berbeda, tidak seperti persentase pencilan 0% dan 10% yang mana perbedaan nilai akurasi penggerombolan tidak terlalu signifikan berbeda.



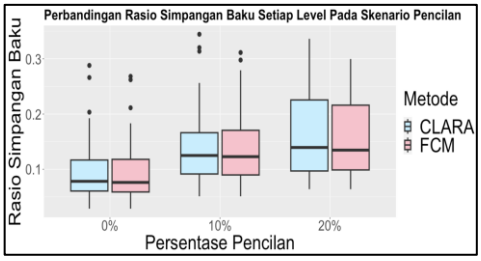
Gambar 7. Plot perbandingan nilai akurasi penggerombolan untuk skenario pencilan dan jarak antar pusat gerombol terhadap metode FCM dan CLARA

3.3 Analisis Hasil Penggerombolan Tiap Skenario

Setelah melihat perbandingan pengaruh antar skenario yang telah dirancang, peneliti tertarik untuk melihat nilai rasio simpangan baku antar gerombol dan akurasi penggerombolan untuk masing-masing skenario itu tanpa mempertimbangkan pengaruh skenario lainnya.

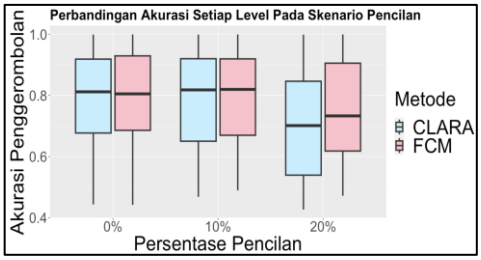
3.3.1 Pencilan

Gambar 8 menunjukkan perbandingan nilai rasio simpangan baku menggunakan *box plot* untuk setiap level pada skenario pencilan dengan menggunakan metode CLARA dan FCM. Nilai rasio simpangan baku untuk level persentase pencilan 0% lebih kecil dibandingkan dengan level persentase pencilan 10% dan 20%. Hal ini menunjukkan bahwa semakin sedikit adanya pencilan dalam suatu data, maka nilai rasio simpangan baku semakin kecil. Sebaran nilai rasio simpangan baku yang dihasilkan metode FCM lebih kecil dibandingkan dengan metode CLARA.



Gambar 8. Perbandingan nilai rasio simpangan baku antara metode CLARA dan FCM berdasarkan persentase pencilan

Gambar 9 memperlihatkan perbandingan akurasi penggerombolan menggunakan *box plot* untuk setiap level skenario pencilan dengan menggunakan metode CLARA dan FCM. Level persentase pencilan 0% menghasilkan nilai akurasi penggerombolan yang lebih tinggi. Perbandingan nilai akurasi penggerombolan metode FCM lebih tinggi dibandingkan dengan metode CLARA, terutama terlihat jelas pada level persentase pencilan 20%.

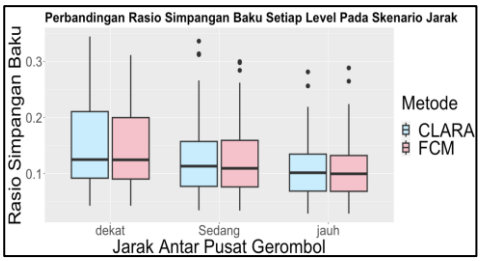


Gambar 9. Perbandingan nilai akurasi penggerombolan antara metode CLARA dan FCM berdasarkan persentase pencilan

3.3.2 Jarak Antar Pusat Gerombol

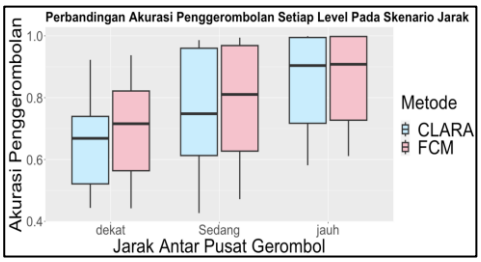
Gambar 10 menampilkan perbandingan nilai rasio simpangan baku menggunakan *box plot* untuk setiap level pada skenario jarak antar pusat gerombol dengan menggunakan metode CLARA dan FCM. Nilai rasio simpangan baku lebih rendah pada level dengan jarak antar pusat gerombol yang jauh dibandingkan dengan level dengan jarak sedang dan dekat. Hal ini menunjukkan bahwa

semakin jauh jarak antar pusat gerombol, maka nilai rasio simpangan baku semakin kecil. Sebaran nilai rasio simpangan baku yang dihasilkan metode FCM lebih kecil dibandingkan dengan metode CLARA.



Gambar 10. Perbandingan nilai rasio simpangan baku antara metode CLARA dan FCM berdasarkan jarak antar pusat gerombol

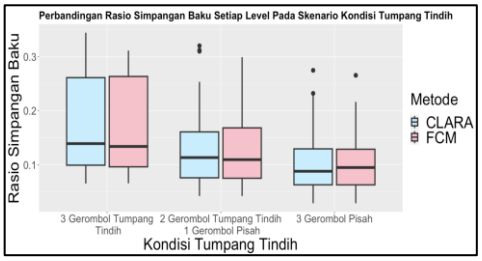
Gambar 11 menunjukkan perbandingan akurasi penggerombolan menggunakan *box plot* untuk setiap level skenario jarak antar pusat gerombol dengan menggunakan metode CLARA dan FCM. Nilai akurasi penggerombolan lebih tinggi pada level dengan jarak antar pusat gerombol yang jauh. Perbandingan nilai akurasi penggerombolan metode FCM lebih tinggi dibandingkan dengan metode CLARA.



Gambar 11. Perbandingan nilai akurasi penggerombolan antara metode CLARA dan FCM berdasarkan jarak antar pusat gerombol

3.3.3 Kondisi Tumpang Tindih

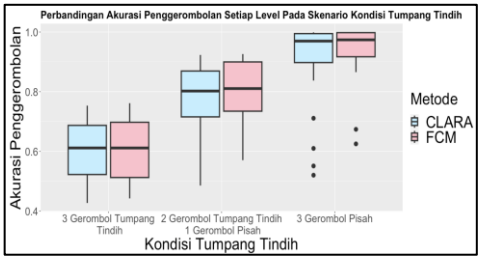
Gambar 12 menunjukkan perbandingan nilai rasio simpangan baku menggunakan *box plot* untuk setiap level pada skenario kondisi tumpang tindih dengan menggunakan metode CLARA dan FCM. Nilai rasio simpangan baku lebih rendah pada level ketiga gerombol saling terpisah dibandingkan dengan level satu gerombol terpisah dua gerombol tumpang tindih dan ketiga gerombol saling tumpang tindih. Hal ini menunjukkan bahwa semakin terpisah penggerombolan, maka nilai rasio simpangan baku semakin kecil. Sebaran nilai rasio simpangan baku yang dihasilkan metode FCM lebih kecil dibandingkan dengan metode CLARA.



Gambar 12. Perbandingan nilai rasio simpangan baku antara metode CLARA dan FCM

berdasarkan kondisi tumpang tindih

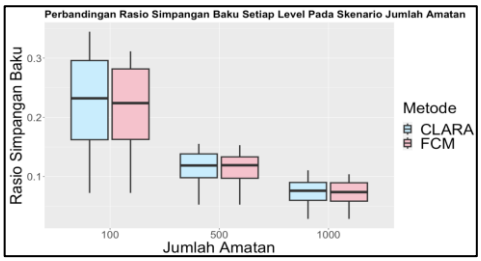
Gambar 13 menunjukkan perbandingan akurasi penggerombolan menggunakan *box plot* untuk setiap level skenario kondisi tumpang tindih dengan menggunakan metode CLARA dan FCM. Nilai akurasi penggerombolan lebih tinggi pada level ketiga gerombol saling terpisah. Perbandingan nilai akurasi penggerombolan metode FCM lebih tinggi dibandingkan dengan metode CLARA.



Gambar 13. Perbandingan nilai akurasi penggerombolan antara metode CLARA dan FCM berdasarkan kondisi tumpang tindih

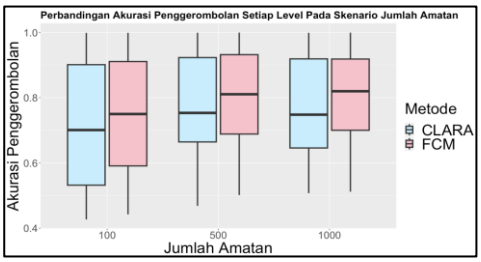
3.3.4 Jumlah Amatan

Gambar 14 menunjukkan perbandingan nilai rasio simpangan baku menggunakan *box plot* untuk setiap level pada skenario jumlah amatan dengan menggunakan metode CLARA dan FCM. Nilai rasio simpangan baku lebih rendah pada level 1000 amatan dibandingkan dengan level 500 dan 100 amatan. Hal ini menunjukkan bahwa semakin besar jumlah amatan, maka nilai rasio simpangan baku semakin kecil. Sebaran nilai rasio simpangan baku yang dihasilkan metode FCM lebih kecil dibandingkan dengan metode CLARA.



Gambar 14. Perbandingan nilai rasio simpangan baku antara metode CLARA dan FCM berdasarkan jumlah amatan

Gambar 15 menunjukkan perbandingan akurasi penggerombolan menggunakan *box plot* untuk setiap level skenario jumlah amatan dengan menggunakan metode CLARA dan FCM. Nilai akurasi penggerombolan lebih lebar pada level 100 amatan dibandingkan dengan level 500 dan 1000 amatan. Perbandingan nilai akurasi penggerombolan metode FCM lebih tinggi dibandingkan dengan metode CLARA.



Gambar 15. Perbandingan nilai akurasi penggerombolan antara metode CLARA dan FCM berdasarkan jumlah amatan

3.4 Uji Konfirmatori (Uji Hipotesis)

Uji ANOVA digunakan untuk mengkonfirmasi pengaruh masing-masing faktor dan interaksinya terhadap rasio simpangan baku dan akurasi penggerombolan. Secara umum, Tabel 1 dan 2 menunjukkan bahwa sebagian besar faktor dan interaksinya berpengaruh signifikan terhadap hasil penggerombolan metode FCM, yang ditunjukkan oleh nilai p-value yang sangat kecil. Namun, terdapat pengecualian, yaitu pada interaksi Jumlah amatan*Jarak antar pusat gerombol* Kondisi tumpang tindih , yang memiliki p-value sebesar 0,2590 (> 0,05), sehingga tidak signifikan secara statistik. Temuan ini sebagian besar mendukung hasil analisis eksploratif pada Subbab 3.3, meskipun tidak semua interaksi memberikan pengaruh signifikan. Hasil untuk metode CLARA dapat dilihat di Lampiran 3.

Tabel 1. Hasil ANOVA Terhadap Rasio Simpangan Baku Metode FCM

	Est.	**df1**	**df2**	***p-value***
Pencilan	3226,97	1,96	1937,99	0,0000
Jumlah amatan	20195,11	1,24	1937,99	0,0000
Jarak antar pusat gerombol	1062,56	1,90	1937,99	0,0000
Kondisi tumpang tindih	3104,35	1,83	1937,99	0,0000
Pencilan*Jumlah amatan	318,77	2,43	1937,99	0,0000
Pencilan*Jarak antar pusat gerombol	28,36	3,68	1937,99	0,0000
Jumlah amatan*Jarak antar pusat gerombol	205,24	2,36	1937,99	0,0000
Pencilan*Kondisi tumpang tindih	163,46	3,47	1937,99	0,0000
Jumlah amatan*Kondisi tumpang tindih	847,05	2,25	1937,99	0,0000
Jarak antar pusat gerombol* Kondisi tumpang tindih	35,10	3,43	1937,99	0,0000
Pencilan*Jumlah amatan*Jarak antar pusat gerombol	6,60	4,55	1937,99	0,0000
Pencilan*Jumlah amatan*Kondisi tumpang tindih	26,11	4,29	1937,99	0,0000
Pencilan*Jarak antar pusat gerombol* Kondisi tumpang tindih	17,98	6,37	1937,99	0,0000
Jumlah amatan*Jarak antar pusat gerombol* Kondisi tumpang tindih	22,12	4,23	1937,99	0,0000
Pencilan*Jumlah amatan*Jarak antar pusat gerombol* Kondisi tumpang tindih	8,65	7,81	1937,99	0,0000

Tabel 2. Hasil ANOVA Terhadap Akurasi Penggerombolan Metode FCM

	Est.	**df1**	**df2**	***p-value***
Pencilan	22,41	2,00	5434,90	0,0000
Jumlah amatan	54,43	2,00	5434,90	0,0000
Jarak antar pusat gerombol	479,79	2,00	5434,90	0,0000
Kondisi tumpang tindih	2046,32	1,73	5434,90	0,0000
Pencilan*Jumlah amatan	8,30	3,99	5434,90	0,0000
Pencilan*Jarak antar pusat gerombol	4,19	3,97	5434,90	0,0022
Jumlah amatan*Jarak antar pusat gerombol	5,61	3,99	5434,90	0,0002
Pencilan*Kondisi tumpang tindih	20,95	3,46	5434,90	0,0000
Jumlah amatan*Kondisi tumpang tindih	16,44	3,46	5434,90	0,0000
Jarak antar pusat gerombol* Kondisi tumpang tindih	32,97	3,40	5434,90	0,0000
Pencilan*Jumlah amatan*Jarak antar pusat gerombol	3,80	7,91	5434,90	0,0002
Pencilan*Jumlah amatan*Kondisi tumpang tindih	2,31	6,91	5434,90	0,0245
Pencilan*Jarak antar pusat gerombol* Kondisi tumpang tindih	38,75	6,75	5434,90	0,0000
Jumlah amatan*Jarak antar pusat gerombol* Kondisi tumpang tindih	1,28	6,79	5434,90	0,2590
Pencilan*Jumlah amatan*Jarak antar pusat gerombol* Kondisi tumpang tindih	2,58	13,44	5434,90	0,0013

4 Simpulan

Penelitian ini menyimpulkan bahwa metode FCM memiliki kinerja dan akurasi lebih baik dibandingkan CLARA, terutama dalam pengelompokan data berukuran besar dan mengandung pencilan. FCM juga lebih stabil dalam menghadapi variasi jumlah amatan, jarak antar pusat gerombol, tumpang tindih gerombol, dan persentase pencilan. Saran untuk penelitian selanjutnya adalah sebaiknya menggunakan metode terbarukan yang dapat menjawab permasalahan ini dan untuk kajian simulasinya diharapkan dapat membangkitkan pencilan secara multivariate. Lampiran analisis dan kode sintaks analisis yang digunakan dalam penelitian ini tersedia secara terbuka di repositori GitHub berikut: https://github.com/Intanjulianapanjaitan/Lampiran-Hasil-Analisis/blob/e196da30e160068f8190bfb05f975e9f93338752/Kode%20Sintax .

5 Daftar Pustaka

[1] C. Hennig, “Cluster validation by measurement of clustering characteristics relevant to the user,” in *Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining*, 2019. doi: 10.1002/9781119597568.ch1.

[2] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, “Comprehensive Review of K-Means Clustering Algorithms,” *Int. J. Adv. Sci. Res. Eng.*, vol. 07, no. 08, 2021, doi:

10.31695/ijasre.2021.34050.

[3] M. Bieber, W. J. C. Verhagen, F. Cosson, and B. F. Santos, "Generic Diagnostic Framework for Anomaly Detection—Application in Satellite and Spacecraft Systems," *Aerospace*, vol. 10, no. 8, 2023, doi: 10.3390/aerospace10080673.

[4] A. Nowak-Brzezińska and W. Łazarz, "Qualitative data clustering to detect outliers," *Entropy*, vol. 23, no. 7, 2021, doi: 10.3390/e23070869.

[5] P. R. Fitrayana and D. R. S. Saputro, "Algoritme Clustering Large Application (CLARA) untuk Menangani Data Outlier," *Prism. Pros. Semin. Nas. Mat.*, vol. 5, pp. 721–725, 2022.

[6] K. L. Wu, "Analysis of parameter selections for fuzzy c-means," *Pattern Recognit.*, vol. 45, no. 1, 2012, doi: 10.1016/j.patcog.2011.07.012.

[7] K. Zhou and S. Yang, "Fuzzifier Selection in Fuzzy C-Means from Cluster Size Distribution Perspective," *Informatica*, vol. 30, no. 3, 2019, doi: 10.15388/informatica.2019.221.

[8] H. Y. Wang, J. S. Wang, and L. F. Zhu, "A new validity function of FCM clustering algorithm based on intra-class compactness and inter-class separation," *J. Intell. Fuzzy Syst.*, vol. 40, no. 6, 2021, doi: 10.3233/JIFS-210555.

[9] C. Ramadhana, Y. D. L. W, and K. D. K. W, "Data Mining dengan Algoritma Fuzzy C-Means Clustering Dalam Kasus Penjualan di PT Sepatu Bata," *Semant. 2013*, vol. 2013, no. November, 2013.

[10] O. N. Kenger, Z. D. Kenger, E. Ozceylan, and B. Mrugalska, "Clustering of Cities Based on Their Smart Performances: A Comparative Approach of Fuzzy C-Means, K-Means, and K-Medoids," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3333753.

[11] B. Choudhary and V. Saxena, "Fuzzy C-Mean Technique for Accessing Large Database of Banking Sector," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 4, 2023.

[12] N. Anand and P. Vikram, "Comprehensive Analysis & Performance Comparison of Clustering Algorithms for Big Data," *Rev. Comput. Eng. Res.*, vol. 4, no. 2, 2017, doi: 10.18488/journal.76.2017.42.54.80.

[13] E. Ahmadov, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms on Demographic Data Using the Pca Method," *Probl. Inf. Technol.*, vol. 14, no. 1, pp. 15–22, 2023, doi: 10.25045/jpit.v14.i1.03.

[14] S. Ghosh and S. Kumar, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 4, 2013, doi: 10.14569/ijacsa.2013.040406.

[15] B. Grün, G. Malsiner-Walli, and S. Frühwirth-Schnatter, "How many data clusters are in the Galaxy data set?: Bayesian cluster analysis in action," *Adv. Data Anal. Classif.*, vol. 16,

no. 2, 2022, doi: 10.1007/s11634-021-00461-8.

[16] N. F. Mohd. Azmi, H. Midi, and N. Fairus Ismail, "The Performance of Clustering Approach with Robust MM–Estimator for Multiple Outlier Detection in Linear Regression," *J. Teknol.*, 2012, doi: 10.11113/jt.v45.320.

[17] Mahmudi, R. Goejantoro, and F. D. T. Amijaya, "Comparison of C-Means and Fuzzy C-Means Methods in the Districts/Cities on the Island of Kalimantan Based on the 2019 HDI Indicators," *J. EKSPONENSIAL*, vol. 12, no. 2, 2021.

[18] R. Babuska, "Fuzzy And Neural Control Disc Course Lecture Notes (October 2001)," *Control*, no. October, 2001.