JPM IAIN Antasari
Vol. 02 No. 1 Juli – Desember 2014, h. 39-62

PERBANDINGAN HASIL PENGGEROMBOLAN K-MEANS, FUZZY K-MEANS, DAN TWO STEP CLUSTERING

Lathifaturrahmah

Abstrak

Analisis gerombol merupakan salah satu metode peubah ganda yang tujuan utamanya adalah mengelompokkan objek berdasarkan kemiripan atau ketidakmiripan karakteristik-karakteristiknya, sehingga objek yang terletak dalam satu gerombol memiliki kemiripan sifat yang lebih besar dibandingkan dengan objek pengamatan yang terletak pada gerombol lain. K-means merupakan salah satu metode penggerombolan tak berhirarki yang paling banyak digunakan, namun karena menggunakan rataan sebagai centroidnya, metode ini lebih sensitif terhadap keberadaan pencilan pada data. Sehingga berkembanglah metode baru, k-medoid, dengan berbasis median sebagai pusat gerombolnya. Penelitian ini bertujuan untuk membandingkan hasil analisis gerombol metode k-means dengan k-medoid baik pada saat data mengandung pencilan maupun tidak. Metode k-medoid diharapkan lebih kekar terhadap pencilan dibandingkan dengan k-means, sehingga dapat memberikan hasil gerombol yang lebih akurat dengan nilai tingkat salah klasifikasi yang lebih kecil. Hasil penggerombolan menunjukkan bahwa metode k-medoid mempunyai nilai rataan tingkat salah klasifikasi yang lebih rendah dan signifikan pada kondisi proporsi pencilan 5%, sedangkan pada kondisi proporsi pencilan 10% dan 15% hasil nilai rataan salah klasifikasinya tidak berbeda signifikan dengan metode k-means.

Kata Kunci: *metode gerombol, k-means, fuzzy k-means, two step cluster*

Pendahuluan

Masalah penggerombolan seringkali ditemui di kehidupan sehari-hari, baik itu terkait dengan bidang sosial, bidang kesehatan, bidang *marketing* maupun bidang akademik. Mendeskripsikan dan memaparkan keunikan proses atau hasil pengelompokan merupakan hal yang menarik dan dapat memberikan ide-ide tertentu. Misalnya saja dalam membuat

segmentasi pemasaran, dengan analisis gerombol dapat dikelompokkan pelanggan atau pembeli berdasarkan manfaat atau keuntungan yang diperoleh dari pembelian barang. Hasil dari penggerombolan ini selanjutnya dapat digunakan dalam pengambilan keputusan untuk strategi pemasaran selanjutnya. Namun jika pengelompokan ini tidak sesuai atau tidak representatif dengan apa yang diharapkan, apalagi menyangkut pengambilan keputusan yang cukup penting akibatnya akan cukup fatal. Oleh karena itu, perlu dilakukan *review* pada proses penggerombolan.

Analisis gerombol adalah salah satu analisis peubah ganda yang digunakan untuk mengelompokkan objek-objek menjadi beberapa gerombol berdasarkan pengukuran kemiripan peubah-peubah yang diamati, sehingga diperoleh kemiripan objek dalam gerombol yang sama dibandingkan antar objek dari gerombol yang berbeda.

Manfaat penggerombolan antara lain adalah untuk eksplorasi data, reduksi data, dan pelapisan data. Dengan eksplorasi data dapat diperoleh informasi yang ada dalam himpunan data, dengan reduksi data dimungkinkan mengambil suatu ringkasan gerombol yang dapat mewakili seluruh anggota tersebut. Penggerombolan dapat digunakan sebagai pelapisan data dalam penarikan contoh atau penggolongan tipe objek.

Dalam penggerombolan objek, untuk menggabungkan dua atau lebih objek menjadi suatu gerombol, biasanya digunakan suatu ukuran kemiripan atau ketidakmiripan. Semakin mirip dua objek semakin tinggi peluang untuk dikelompokkan dalam suatu gerombol. Sebaliknya semakin tidak mirip semakin rendah pula peluang untuk dikelompokkan dalam satu gerombol.

Pada umumnya metode pada analisis gerombol dibedakan menjadi metode berhierarki (*hierarchical clustering methods*) dan metode tak berhierarki (*non hierarchical clustering methods*). Metode berhierarki digunakan bila jumlah gerombol yang diinginkan tidak diketahui, sedangkan metode tak berhierarki digunakan bila jumlah kelompok yang diinginkan telah ditentukan sebelumnya. Contoh dari metode tak

berhierarki yang sering digunakan adalah *k-means* dan *fuzzy k-means* dan kedua metode ini cocok digunakan untuk data berukuran besar yang memiliki tipe peubah kontinu. Namun dewasa ini telah dikembangkan suatu metode untuk jenis data yang berukuran besar, yaitu metode *two step cluster*. Metode ini dikembangkan oleh Chiu *et al.* (2001) yang memungkinkan untuk mengolah data yang memiliki tipe peubah berbeda, yaitu kontinu dan kategorik.

Ketiga metode ini memiliki kelebihan maupun kelemahan. Menurut Serban dan Grigoreta (2006) dalam penelitiannya metode *fuzzy k-means* lebih baik dari pada *k-means* pada aspek *mining.* Kelebihan dari metode *k-means* adalah mampu mengelompokkan data besar dengan sangat cepat, sedangkan kekurangan dari metode *k-means* adalah banyaknya gerombol harus ditentukan sebelumnya (Teknomo 2007). Adapun kelebihan dari *fuzzy k-means* adalah mampu menempatkan suatu data yang terletak diantara dua atau lebih gerombol yang lain pada suatu gerombol, dan menurut Kusumadewi *et al.* (2006) kelemahannya adalah pada partisi *fuzzy* masih belum dapat membedakan apakah suatu data merupakan anggota beberapa gerombol atau merupakan data pencilan. Menurut Kusdiati (2006) dalam penelitiannya menyatakan bahwa persentasi salah klasifikasi dari metode *two step cluster* tidak berbeda nyata dengan yang dihasilkan dari metode *k-means*, jika peubahnya kontinu.

Pada penelitian ini digunakan data Afifi yang diambil dari software SPSS. Dari data yang sama ingin dibandingkan hasil penggerombolan dengan metode *k-means*, metode *fuzzy k-means*, dan metode *two step cluster* yang akan memberikan penggerombolan yang terbaik, yaitu yang mempunyai variansi di dalam yang lebih homogen dan variansi antar gerombol yang lebih heterogen.

Oleh karena itu, tujuan penelitian ini adalah (1) membandingkan hasil penggerombolan metode *k-means*, *fuzzy k-means*, dan *two step cluster* pada data Afifi*;* (2) menentukan jumlah *cluster* yang ideal untuk masing-masing metode tersebut pada data Afifi.

Metode Penelitian

A. Metode *K-means Clustering*

Metode *k-means* pertama kali diperkenalkan oleh MacQueen JB pada tahun 1976. Metode ini adalah salah satu metode *non hierarchi* yang umum digunakan. Metode ini termasuk dalam teknik penyekatan (*partition*) yang membagi atau memisahkan objek ke *k* daerah bagian yang terpisah. Pada *k-means*, setiap objek harus masuk dalam gerombol tertentu, tetapi dalam satu tahapan proses tertentu, objek yang sudah masuk dalam satu gerombol, pada satu tahapan berikutnya objek akan berpindah ke gerombol lain. Untuk itu digunakan algoritma *k-means* yang di dalamnya memuat aturan, yakni (1) jumlah *cluster* yang diinginkan; (2) Hanya memiliki atribut bertipe numerik.

Metode *k-means* berawal dari penentuan jumlah gerombol yang ingin dibentuk, kemudian menentukan objek sebagai *centroid* awal yang biasanya dilakukan secara random, selanjutnya menghitung ukuran jarak dari masing-masing objek ke *centroid*. Setelah objek masuk pada *centroid* terdekat dan membentuk gerombol baru, *centroid* baru ditentukan kembali dengan menghitung rata-rata objek pada *centroid* yang sama. Jika masih ada perbedaan dengan *centroid* yang sudah dibentuk, maka dilakukan perhitungan kembali *centroid* baru.

Hasil *cluster* dengan dengan metode *k-means* sangat bergantung pada nilai pusat gerombol awal yang diberikan. Pemberian nilai awal yang berbeda bisa menghasilkan gerombol yang berbeda. Ada beberapa cara memberi nilai awal misalnya dengan mengambil sampel awal dari objek, lalu mencari nilai pusatnya, memberi nilai awal secara random, menentukan nilai awalnya atau menggunakan hasil dari gerombol hierarki dengan jumlah gerombol yang sesuai (Santosa 2007).

Tujuan dari algoritma *k-means* adalah meminimumkan jarak antara objek dengan *centroid* yang terdekat, yaitu dengan meminimumkan fungsi objektif *J* yang dirumuskan sebagai fungsi dari *U* dan *V* sebagai berikut:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{C} \mu_{ik} d^2(x_k, v_i)$$

dengan:

U : matriks keanggotaan objek ke masing-masing gerombol
V : matriks *centroid* / rata masing-masing gerombol
μ_{ik} : fungsi keanggotaan objek ke-*k* ke gerombol ke-*i*
x_k : objek ke-*k*
v_i : nilai *centroid* gerombol ke-*i*
d : ukuran jarak

B. Metode *Fuzzy k-means*

Konsep dasar *fuzzy k-means* pertama kali adalah menentukan pusat *cluster* pada kondisi awal, pusat *cluster* ini masih belum akurat dan tiap objek memiliki derajat keanggotaan untuk tiap-tiap *cluster* dengan cara memperbaiki pusat *cluster* dan nilai keanggotaan tiap objek secara berulang maka akan dapat dilihat bahwa pusat *cluster* akan bergerak menuju lokasi yang tepat.

Ketika gerombol-gerombol menjadi *overlapping* atau setiap objek memungkinkan termasuk ke beberapa gerombol, maka μ_{ik} dapat diinterpretasikan sebagai fungsi keanggotaan yaitu $\mu_{ik} \in [0,1]$. Maka fungsi objektif *J* yang dirumuskan sebagai fungsi dari *U* dan *V* sebagai berikut:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{C} (\mu_{ik})^m d^2(x_k, v_i)$$

dengan:

U : matriks keanggotaan objek ke masing-masing gerombol
V : matriks *centroid* / rata-rata masing-masing gerombol
m : pembobot eksponen
μ_{ik} : fungsi keanggotaan objek ke-*k* ke gerombol ke-*i*
x_k : objek ke-*k*
v_i : nilai *centroid* ke-*i*
d : ukuran jarak

Untuk menghitung *centroid* (titik pusat) gerombol *V*, untuk setiap gerombol digunakan rumus sebagai berikut:

$$v_{ij} = \frac{\sum_{k=1}^{N} (\mu_{ik})^m x_{kj}}{\sum_{k=1}^{N} (\mu_{ik})^m}$$

dengan:

m : pembobot eksponen
μ_{ik} : fungsi keanggotaan objek ke-k ke gerombol ke-i
x_{kj} : objek ke-k gerombol ke-j

C. Metode *Two Step Clustering*

Metode *two step cluster* adalah metode yang didesain untuk menangani jumlah objek yang besar, terutama pada masalah objek yang mempunyai peubah kontinu dan kategorik. Prosedur penggerombolan dengan metode *two step cluster* mempunyai dua tahapan yaitu tahap *preclustering* (penggerombolan awal) objek ke dalam *subcluster-subcluster* kecil dan tahap penggerombolan akhir.

Langkah 1: Penggerombolan Awal (*Preclustering*)

Menurut Anonimous (2001) tahap penggerombolan awal dilakukan dengan pendekatan sekuensial, yaitu objek diamati satu persatu berdasarkan ukuran jarak yang kemudian ditentukan apakah objek tersebut masuk dalam gerombol yang telah terbentuk atau harus membentuk gerombol baru. Pada langkah ini diimplementasikan dengan pembentukan *cluster features* (CF) *Tree. Cluster future* itu sendiri adalah kesimpulan dari informasi yang di kumpulkan pada suatu *cluster.*

Langkah 2: Penggerombolan akhir

Pada langkah ini, hasil dari *CF Tree* digerombolkan dengan analisis gerombol hierarki dengan metode *agglomerative*, yaitu dimulai dengan *n* gerombol yang masing-masing beranggotakan satu objek, kemudian dua gerombol yang paling dekat digabung dan ditentukan kembali kedekatan

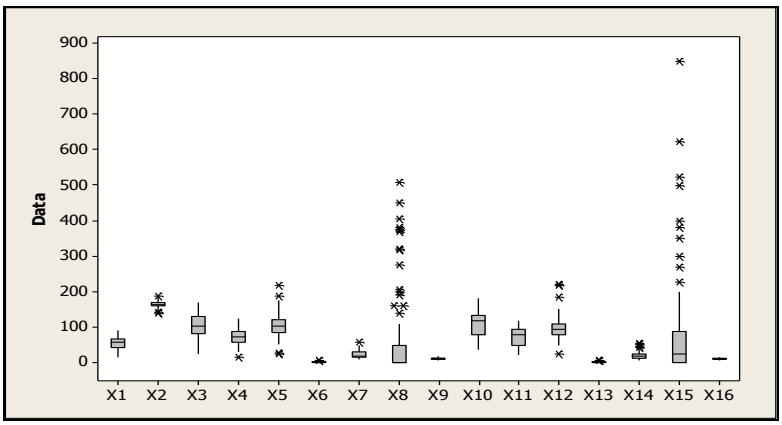
antar gerombol yang baru. Untuk menghitung banyaknya gerombol dapat dilakukan dengan dua tahapan, yang pertama menghitung *schwarz's bayesian criterion* (*BIC*) atau *akaike's information criterion* (*AIC*) untuk tiap gerombol. Rumus *BIC* dan *AIC* untuk gerombol *J* adalah sebagai berikut:

$$BIC(J) = -2\sum_{j=1}^{J} \xi_j + m_j \log(N)$$

$$AIC(J) = -2\sum_{j=1}^{J} \xi_j + 2m_j$$

Hasil dan Pembahasan

Terdapat data hilang pada obyek pengamatan untuk beberapa peubah. Obyek pengamatan yang memiliki data hilang tersebut tidak diikutsertakan dalam analisis. Untuk memberikan gambaran data dari masing-masing peubah maka digunakanlah Boxplot, yang disajikan pada gambar dibawah ini:

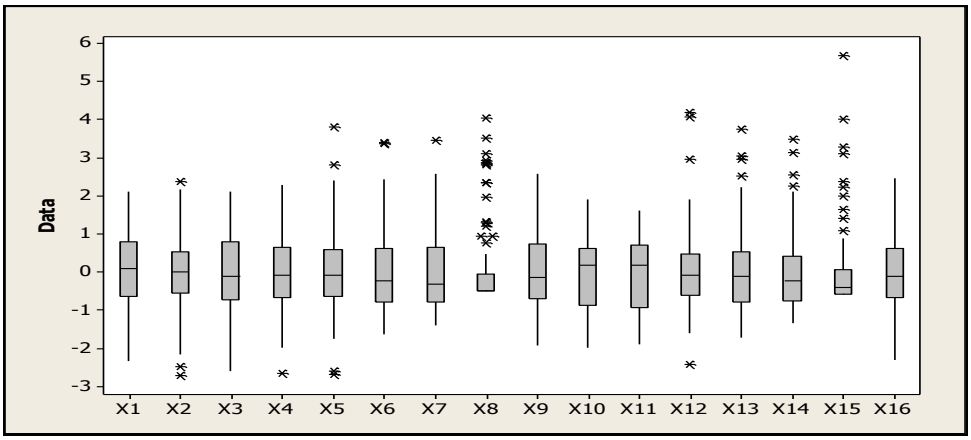


Gambar 1. Boxplot Data Afifi

Keterangan:

X_1= Age
X_2= Height
X_3= Systolic Blood Pressure1
X_4= Mean Arterial Pressure 1
X_5= Heart Rate1
X_6= Cardiac1
X_7 = CTime2
X_8 = Urine 1
X_9= Hemoglobin1
X_{10}= Systolic Blood Pressure2
X_{11}= Mean Arterial Pressure 2
X_{12}= Heart Rate 2
X_{13}= Cardiac 2
X_{14}= CTime 2
X_{15}= Urine 2
X_{16}= Hemoglobin 2

Gambar 1 sebelumnya memperlihatkan bahwa sebaran data untuk peubah 2 sampai peubah 17 cenderung bervariasi dan dapat dilihat bahwa terdapat pencilan untuk peubah X_2, X_4, X_5, X_6, X_8, X_8, X_9, X_{13}, X_{14}, X_{15}, dan X_{16}. Gambar 1 juga memperlihatkan bahwa keragaman peubah X_{15} lebih besar dari keragaman peubah lainnya, sedangkan peubah X_{13} mempunyai keragaman yang paling kecil dibandingkan peubah lainnya. Sedangkan untuk memberikan gambaran data yang sudah distandarisasi, dapat dilihat pada gambar berikut:



Gambar 2. Boxplot Data Afifi Standarisasi

Gambar 2 memperlihatkan bahwa data yang sudah distandarisasi ini mempunyai variansi yang semua peubahnya cenderung relatif lebih homogen.

Karena dalam penggerombolan menggunakan konsep jarak *eulid*, dimana konsep jarak ini mengharuskan tidak adanya korelasi antar peubah, maka terlebih dahulu dilakukan Analisis Komponen Utama (AKU), yang bertujuan untuk memperoleh peubah-peubah yang saling tidak berkorelasi. Hasil Analisis Komponen Utama disajikan pada tabel berikut:

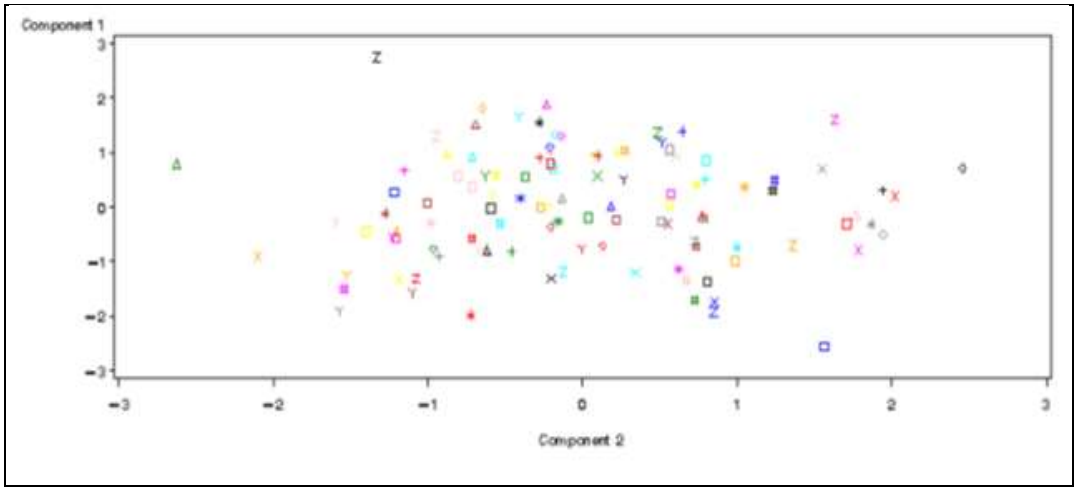
Tabel 1. Koefisien Komponen Utama 1 dan 2

Peubah	Komponen Utama 1	Komponen Utama 2
X_1	-0.2055	0.1417
X_2	0.2239	0.0050

X_3	0.3371	0.1548
X_4	0.3376	0.2173
X_5	-0.0215	0.0765
X_6	0.1763	-0.3690
X_7	-0.2015	0.4052
X_8	0.2015	-0.9954
X_9	-0.0417	0.4142
X_{10}	0.3468	0.2050
X_{11}	0.3662	0.2304
X_{12}	0.2278	0.1716
X_{13}	0.3487	-0.2041
X_{14}	-0.3005	0.2095
X_{15}	0.1470	0.1334
X_{16}	0.0623	0.4391

Sebagai hasil pendekatan yang dilakukan oleh Analisis Komponen Utama pada tabel di atas, dapat dilihat bahwa hanya terdapat 7 komponen utama yang memiliki akar ciri lebih dari 1, ini berarti bahwa ketujuh komponen utama tersebut memberikan kontribusi keragaman yang besar, dan komponen utama yang memiliki akar ciri kurang dari 1 dianggap memiliki kontribusi keragaman yang kurang. Dari tabel di atas, dapat dilihat juga bahwa akar ciri pertama yang memiliki nilai sebesar 4.1284 menjelaskan bahwa komponen utama ke-1 dapat menerangkan keragaman data sebesar 25.80%. Dengan cara yang sama untuk komponen utama selanjutnya sampai komponen ke 16 sebesar 2.87%. Komponen utama ke 1 dan ke 2 memberikan kontribusi keragaman sebesar 25.80% dan 16.73%. Sehingga jika digunakan kedua komponen tersebut, secara kumulatif akan didapatkan keragaman total yang mampu dijelaskan keduanya adalah sebesar 42.53%. Dan dari ketujuh komponen utama tersebut, secara kumulatif memiliki proporsi keragaman sebesar 79.63%, ini berarti bahwa sudah mewakili keragaman total dari seluruh data.

Jika digambarkan nilai kedua skor komponen utama di atas, akan didapatkan gambaran sebagai berikut:



Gambar 3. Plot dua komponen utama pada data Afifi

Gambar 3 memperlihatkan bahwa sebaran data Afifi ini tidak terlihat adanya penggerombolan yang jelas, karena terdapat penggerombolan yang saling tumpang tindih.

Penggerombolan dengan 2 gerombol

Untuk data ini terlebih dahulu ditransformasikan ke dalam bentuk baku sebab adanya perbedaan satuan pengukuran antar peubah. Data yang digunakan untuk pengelompokkan ini adalah data yang mempunyai skala kontinu (interval atau rasio), skala data ini merupakan persyaratan umum digunakannya teknik analisis *cluster*.

Hasil pengelompokkan dengan 2 gerombol untuk metode *k-mean*, *fuzzy k-means* dan *two step cluster* dapat dilihat sebagai berikut :

Tabel 2. Distribusi anggota 2 gerombol

Metode	*k-means*		*fuzzy k-means*		*two step cluster*	
	Jumlah	Persen	Jumlah	Persen	Jumlah	Persen
Gerombol 1	55	51%	56	52%	4	3.8 %
Gerombol 2	53	49%	52	48%	104	96.2%

Tabel 2 memperlihatkan bahwa untuk metode *k-means* dan *fuzzy k-means* penyebaran anggota antara gerombol 1 dan gerombol 2 cenderung hampir sama. Sedangkan distribusi anggota *two step cluster* terlihat jauh

berbeda bila dibandingkan dengan kedua metode tersebut. Selain kesesuaian metode dengan jumlah data yang digunakan, faktor yang menentukan hasil *clustering* ini adalah pemilihan *threshold* atau kriteria penghentian algoritma dari masing-masing metode. Nilai *threshold* ini secara langsung mempengaruhi jumlah *cluster* yang dibentuk. Jika nilai terlalu kecil *cluster-cluster* yang tidak dibutuhkan akan dibuat. Sebaliknya jika nilai terlalu besar maka *cluster-cluster* yang tepat akan diciptakan.

Hasil perbandingan jumlah anggota yang identik dan besarnya persentasi *misclustering* antara metode *k-means*, *fuzzy k-means*, dan *two step cluster* dapat dilihat sebagai berikut:

Tabel 3. Persentasi *misclustering* 2 gerombol hasil antara *k-means* dengan *fuzzy k-means*

Metode		*fuzzy k-means*	
		G1= 56	G2= 52
k-means	G1 = 53	n = 53 100%	n = 0 0%
	G2 = 55	n = 3 5.5%	n = 52 94.5%

Tabel 4. Persentasi *misclustering* 2 gerombol antara metode *k-means* dengan *two step cluster*

Metode		*two step cluster*	
		G1= 104	G2= 4
k-means	G1 = 53	n = 53 100%	n = 0 0%
	G2 = 55	n = 51 92.7%	n = 4 7.3%

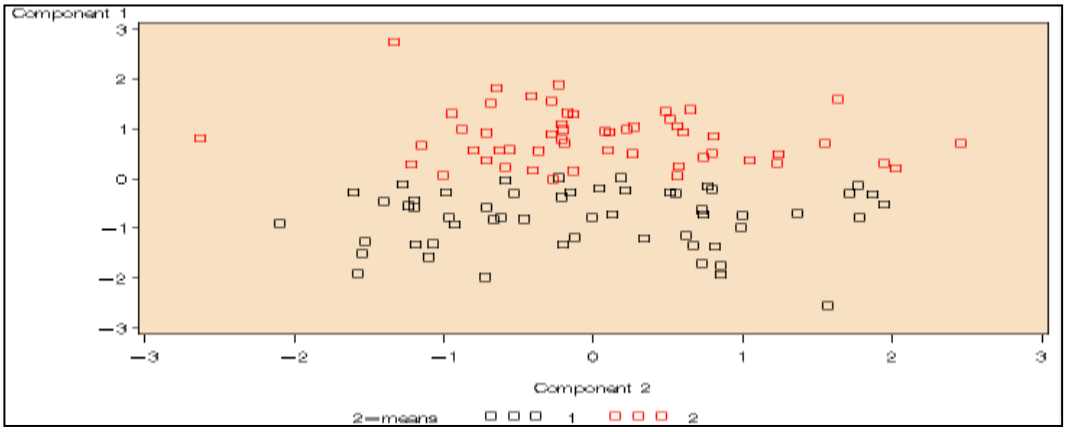
Tabel 5 Persentasi *misclustering* 2 gerombol antara metode *two step cluster* dengan *fuzzy k-means*

Metode		*two step cluster*	
		G1=104	G2= 4
fuzzy k-means	G1= 56	n = 56 100%	n = 0 0%
	G2 = 52	n = 48 92.3%	n = 4 7.7%

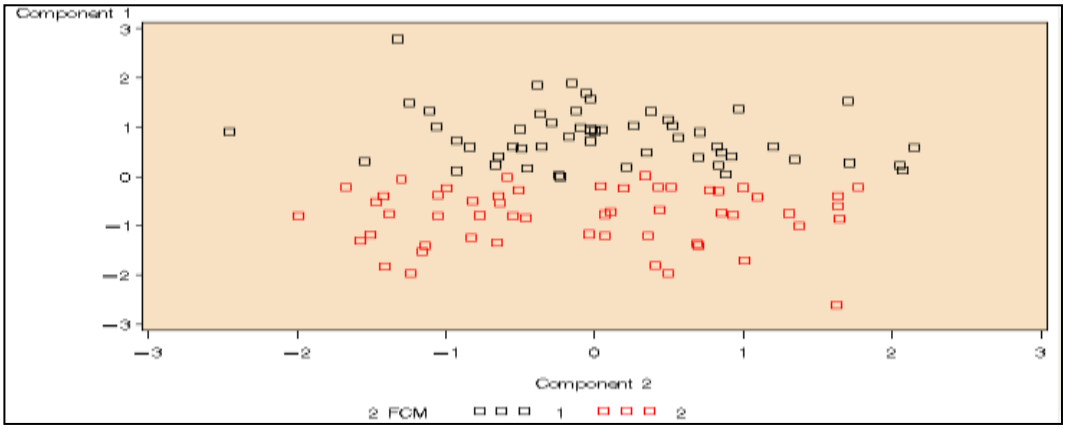
Berdasarkan tabel 3, 4, dan 5 di atas, untuk penggerombolan dengan 2 gerombol banyaknya anggota identik terbesar dimiliki oleh metode *k-means* dengan *fuzzy k-means*. Persentasi salah penggerombolan (*misclustering*) untuk kondisi pada tabel-tabel di atas terlihat bahwa untuk metode penggerombolan yang berbasis *k-means* terhadap *fuzzy k-means* memiliki persentasi salah penggerombolannya paling kecil yaitu sebesar 0%. Ini menunjukkan bahwa penggerombolan antara kedua metode ini memiliki hasil yang tidak jauh berbeda. Sedangkan bila metode *k-means* dibandingkan dengan *two step cluster* terlihat bahwa terdapat nilai *misclustering* yang cukup besar, yaitu mencapai nilai 92.7%, yang berarti bahwa kedua metode ini memiliki hasil yang agak jauh berbeda.

Demikian halnya dengan metode *fuzzy k-means* dengan *two step cluster*, kedua metode ini memiliki nilai *misclustering* yang tidak jauh berbeda dengan metode sebelumnya (*k-means* dengan *two step cluster*) yaitu mencapai nilai sebesar 92.3%. Ini terjadi karena pembentukan *misclustering* ini bergantung pada ketepatan metode dengan besarnya jumlah data yang digunakan.

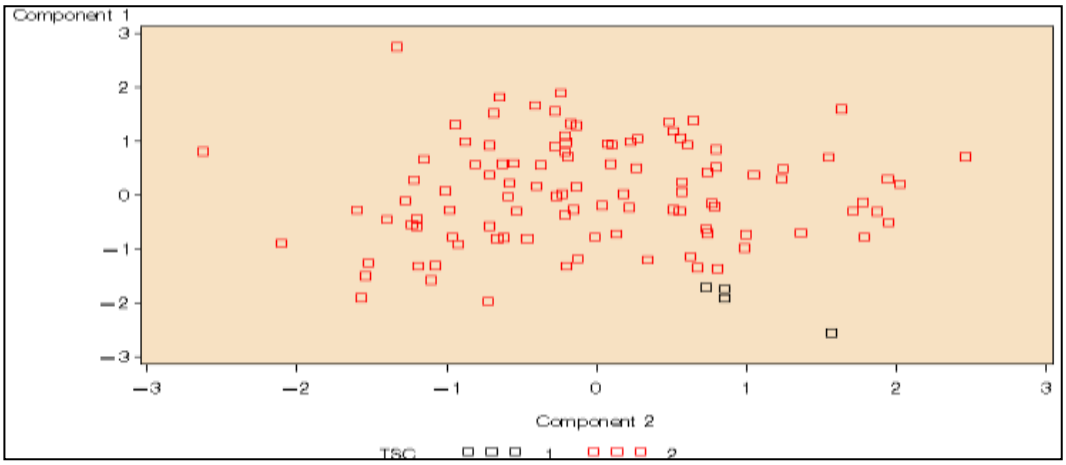
Plot komponen utama dari penggerombolan dengan 2 gerombol untuk masing-masing metode disajikan pada gambar berikut:



Gambar 4. Plot dua komponen utama 2 gerombol pada metode *k-means*



Gambar 5. Plot dua komponen utama 2 gerombol pada metode *fuzzy k-means*



Gambar 6. Plot dua komponen utama 2 gerombol pada metode *two step clustering*

Dari sini dapat disimpulkan bahwa hasil pengelompokkan 2 gerombol antara ketiga metode tersebut yang mempunyai nilai yang tidak jauh berbeda adalah antara metode *k-means* dengan metode *fuzzy k-means*.

Penggerombolan dengan 3 gerombol

Hasil distribusi pengelompokkan dengan 3 gerombol untuk metode *k-means*, *fuzzy k-means* dan *two step cluster* dapat dilihat sebagai berikut:

Tabel 6. Distribusi anggota 3 gerombol

Metode	*k-means*		*fuzzy k-means*		*two step cluster*	
	Jumlah	Persen	Jumlah	Persen	Jumlah	Persen
Gerombol 1	24	22%	56	51.8%	4	3.7%
Gerombol 2	45	42%	2	1.9%	103	95.4%
Gerombol 3	39	36%	50	46.3%	1	0.9%

Dari tabel di atas dapat dilihat bahwa penyebaran anggota untuk pengelompokkan 3 gerombol untuk metode *k-means*, masing-masing gerombolnya tidak merata. Demikian juga dengan metode *fuzzy k-means* dan *two step cluster*, penyebaran anggota pada setiap gerombolnya cenderung jauh berbeda, ini disebabkan karena keragaman antar kelompok yang dimiliki kedua metode ini relatif lebih besar bila dibandingkan dengan *k-means*, jumlah pembentukan banyaknya anggota pada metode ini dipengaruhi oleh besarnya fungsi objektif dan besarnya kriteria penghentian yang dibentuk, serta kesesuaian metode dengan jumlah data yang digunakan.

Hasil perbandingan jumlah anggota yang identik dan besarnya persentasi *misclustering* antara metode *k-means*, *fuzzy k-means*, dan *two step cluster* dapat dilihat sebagai berikut:

Tabel 7. Persentasi *misclustering* 3 gerombol antara metode *k-means* dengan *fuzzy k-means*

Metode		*fuzzy k-means*		
		G1 = 56	G2 = 50	G3=2
k-means	G1=39	n = 38 97.4%	n = 0 0%	n = 1 2.6%
	G2= 45	n=2 4.5%	n =42 93.3%	n = 1 2.2%
	G3= 24	n = 16 66.7%	n = 8 3.3%	n = 0 0%

Tabel 8. Persentasi *misclustering* 3 gerombol antara metode *k-means* dengan *two step cluster*

Metode		*two step cluster*		
		G1 = 103	G2 = 4	G3=1
k-means	G1=24	n = 24 100 %	n = 0 0 %	n = 0 0%
	G2= 45	n = 41 91.1%	n = 4 8.9%	n = 0 0%
	G3= 39	n = 38 97.4%	n = 0 0%	n=1 2.56%

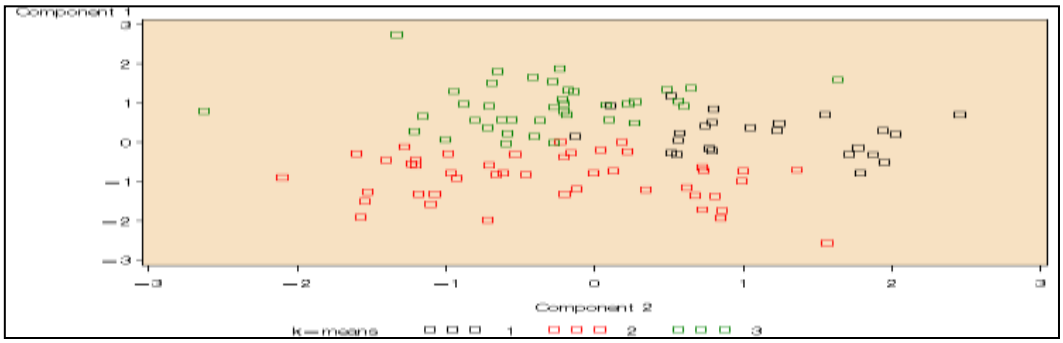
Tabel 9. Persentasi *misclustering* 3 gerombol antara metode *k-means* terhadap *two step cluster*

Metode		*two step cluster*		
		G1 = 103	G2 = 4	G3=1
fuzzy k-means	G1= 2	n = 2 100%	n = 0 0%	n = 0 0%
	G2= 50	n = 46 92.3%	n = 4 7.7%	n = 0 0%
	G3=56	n =55 98.2%	n = 0 0%	n = 1 1.78%

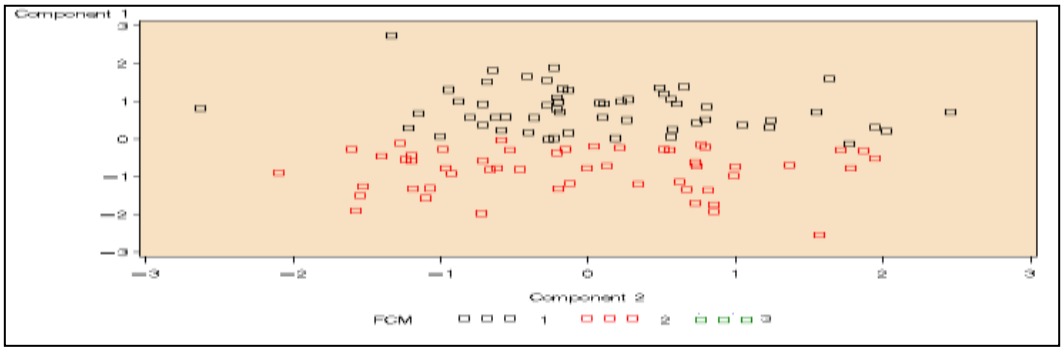
Berdasarkan tabel 7, 8, dan 9 di atas, untuk penggerombolan dengan 3 gerombol diperoleh bahwa persentasi anggota identik yang dimiliki oleh metode *k-means* dengan *fuzzy k-means* lebih besar bila dibandingkan dengan hasil perbandingan antara metode *k-means* dengan *two step cluster.* Ini disebabkan oleh perbedaan keragaman yang besar antara kedua metode tersebut. Sedangkan pada persentasi salah penggerombolan (*misclustering*) untuk kondisi ini, pada tabel-tabel di atas terlihat bahwa untuk metode penggerombolan yang berbasis *k-means* dengan *fuzzy k-means* memiliki persentasi salah penggerombolannya paling kecil yaitu sebesar 0%. Sedangkan *misclustering* yang paling besar dimiliki oleh metode penggerombolan antara *fuzzy k-means* dengan *two step cluster.* Nilai *misclustering* yang paling besar ini mencapai hingga 98.2 %.

Ini menunjukkan bahwa penggerombolan antara metode *k-means* dengan *fuzzy k-means* tidak jauh berbeda, sedangkan antara metode *fuzzy k-means* dengan *two step cluster* hasil penggerombolannya agak jauh berbeda. Ini terjadi karena pada pembentukan *misclustering* ini bergantung pada ketepatan metode dengan besarnya jumlah data yang digunakan.

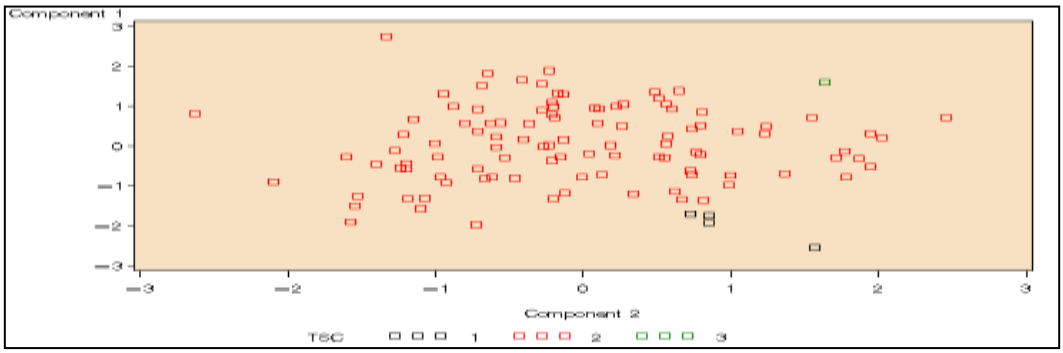
Plot komponen utama dari penggerombolan dengan 3 gerombol untuk masing-masing metode disajikan pada gambar berikut:



Gambar 7. Plot dua komponen utama 3 gerombol pada metode *k-means*



Gambar 8. Plot dua komponen utama 3 gerombol pada metode *fuzzy k-means*



Gambar 9. Plot dua komponen utama 3 gerombol pada metode *two step clustering*

Penggerombolan dengan 4 gerombol

Hasil distribusi pengelompokkan dengan 4 gerombol untuk metode *k-means, fuzzy k-means* dan *two step cluster* dapat dilihat sebagai berikut:

Tabel 10. Distribusi anggota 4 gerombol

Metode	*k-means*		*fuzzy k-means*		*two step cluster*	
	Jumlah	Persen	Jumlah	Persen	Jumlah	Persen
Gerombol 1	35	32.4%	2	1.9%	4	3.7%
Gerombol 2	39	36.1%	54	50.0%	99	91.7%
Gerombol 3	7	6.5%	51	47.2%	4	3.7%
Gerombol 4	27	2.5%	1	0.9%	1	0.9%

Dari tabel di atas dapat dilihat bahwa pengelompokkan dengan 4 gerombol, untuk metode *k means* yang memiliki distribusi jumlah anggota yang tidak merata antara masing-masing gerombol, ini dikarenakan keragaman antar kelompoknya yang relatif besar, dan pada metode ini yang memiliki gerombol terbanyak adalah pada gerombol 2. Begitu juga dengan metode *two step cluster* yang memiliki distribusi jumlah anggota yang sangat tidak merata untuk masing-masing gerombolnya, ini disebabkan karena keragaman antar kelompoknya yang sangat besar. Sedangkan pada metode *fuzzy k-means*, jumlah masing-masing anggotanya juga tidak merata. Jumlah pembentukan banyaknya anggota pada metode-metode ini dipengaruhi oleh besarnya fungsi objektif dan besarnya kriteria penghentian yang dibentuk, serta kesesuaian metode dengan jumlah data yang digunakan.

Hasil perbandingan jumlah anggota yang identik dan besarnya persentasi *misclustering* antara metode *k-means*, *fuzzy k-means*, dan *two step cluster* dapat dilihat sebagai berikut:

Tabel 11. Persentasi *misclustering* 4 gerombol antara metode *k-means* dengan *fuzzy k-means*

Metode		*fuzzy k-means*			
		G1 = 51	G2 = 54	G3=2	G4=1
k-means	G1= 27	n = 15 55.6%	n = 12 44.4%	n = 0 0%	n = 0 0%
	G2= 39	n = 16 41.0%	n = 23 59.0%	n = 0 0 %	n = 0 0%
	G3= 35	n = 17 48.5%	n = 16 45.7%	n = 1 2.9%	n = 1 2.9%
	G4= 7	n = 3 42.9%	n = 3 42.9%	n = 1 14.2%	n =0 0%

Tabel 12. Persentasi *misclustering* 4 gerombol antara metode *k-means* dengan *two step cluster*

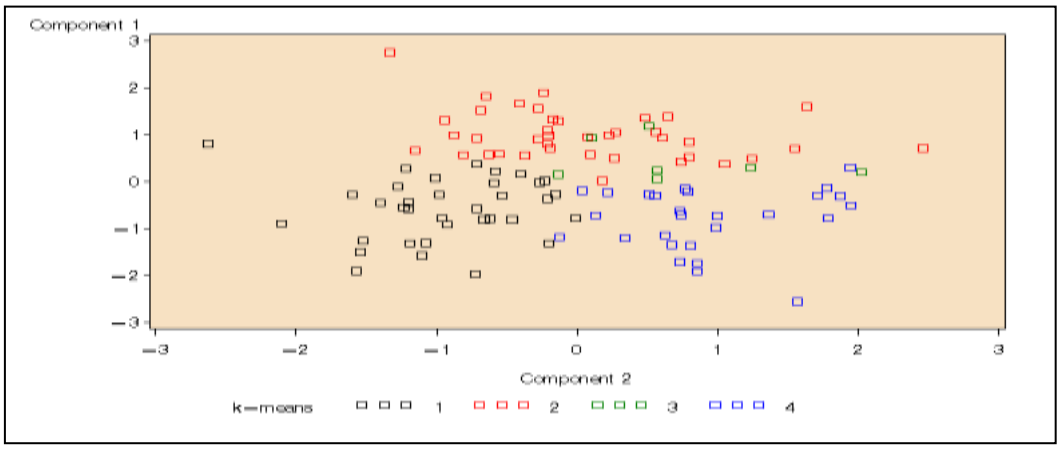
Metode		*two step cluster*			
		G1 = 99	G2 = 4	G3=4	G4=1
k-means	G1= 7	n = 7 100%	n = 0 0%	n = 0 0%	n = 0 0%
	G2= 27	n = 24 88.9%	n = 2 7.4%	n = 1 3.7%	n = 0 0%
	G3= 39	n = 35 89.8%	n = 0 0%	n = 3 7.7%	n = 1 2.5%
	G4= 35	n = 33 94.3%	n = 2 5.7%	n = 0 0%	n = 0 0%

Tabel 13. Persentasi *misclustering* 4 gerombol antara metode *fuzzy k-means* dengan *two step cluster*

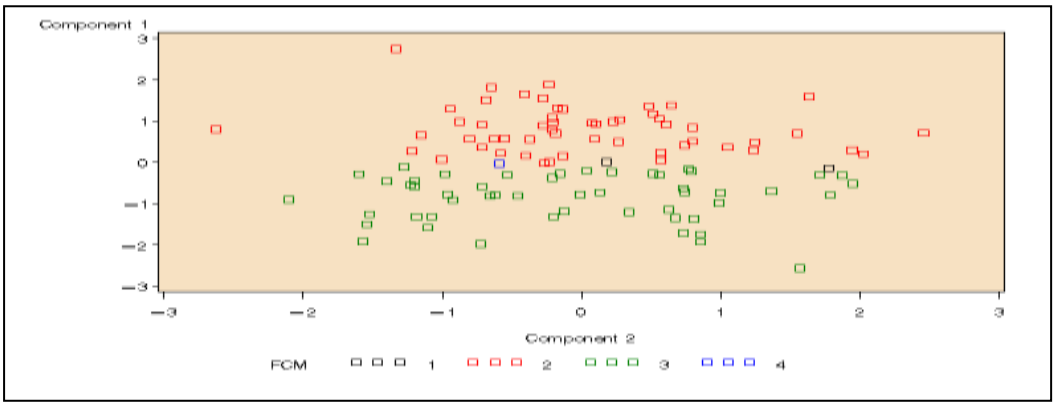
Metode		*two step cluster*			
		G1 = 99	G2 = 4	G3=4	G4=1
fuzzy k-means	G1= 2	n = 2 100%	n = 0 0%	n = 0 0%	n = 0 0%
	G2= 51	n = 47 92.2%	n = 4 7.8%	n = 0 7.4%	n = 0 0%
	G3= 1	n = 1 100%	n = 0 0%	n = 0 0%	n = 0 0%
	G4= 54	n = 49 90.7%	n = 0 0%	n = 0 0%	n = 1 1.9%

Berdasarkan tabel 14, 15, dan 16 di atas, untuk penggerombolan dengan 4 gerombol persentasi anggota identik terbesar dimiliki oleh metode *k-means* dan *fuzzy k-means* dan yang memiliki anggota yang identik terkecil adalah antara metode *fuzzy k-means* dengan metode *two step clustering*. Ini disebabkan oleh perbedaan keragaman yang besar antara kedua metode tersebut. Sedangkan pada persentasi salah penggerombolan (*misclustering*) untuk kondisi ini, pada tabel-tabel di atas terlihat bahwa untuk metode penggerombolan yang berbasis *k-means* terhadap *fuzzy k-means* memiliki persentasi salah penggerombolannya paling kecil yaitu sebesar 0%. Ini menunjukkan bahwa penggerombolan antara kedua metode ini memiliki hasil yang tidak jauh berbeda. Sedangkan bila dibandingkan dengan *two step cluster* terlihat bahwa terdapat nilai *misclustering* yang cukup besar, yaitu mencapai nilai 94.3%. Namun, nilai *misclustering* yang paling besar dimiliki oleh perbandingan metode antara *fuzzy k-means* dan *two step cluster* yaitu mencapai 100%. Ini terjadi karena pembentukan *misclustering* ini bergantung pada ketepatan metode dengan besarnya jumlah data yang digunakan.

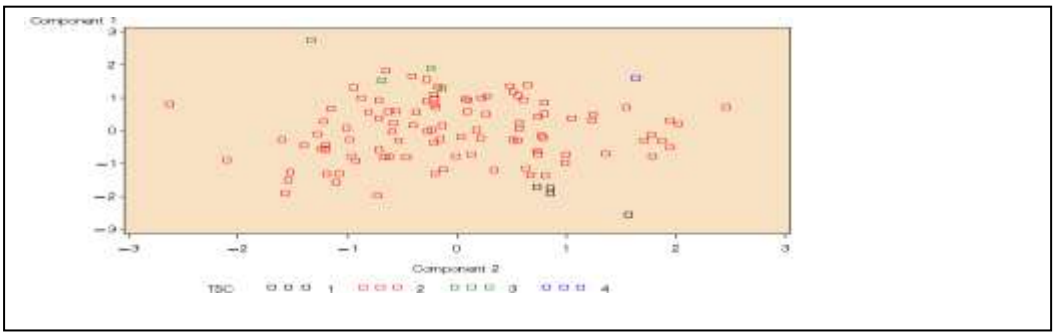
Plot komponen utama dari penggerombolan dengan 4 gerombol untuk masing-masing metode disajikan pada gambar berikut:



Gambar 10. Plot dua komponen utama 4 gerombol pada metode *k-means*



Gambar 11. Plot dua komponen utama 4 gerombol pada metode *fuzzy k-Means*



Gambar 12. Plot dua komponen utama 4 gerombol pada metode *two step clustering*

Untuk mengukur nilai penyebaran dari data hasil *clustering* digunakanlah nilai variansi masing masing gerombol. Berikut ini disajikan nilai variansi pada masing-masing metode penggerombolan:

Tabel 14. Variansi 2 Gerombol

	k-means	*fuzzy k-means*	*two step cluster*
Variance Within Cluster	13.263	15.730	15.356
Variance Between Cluster	310.673	524.234	459.652

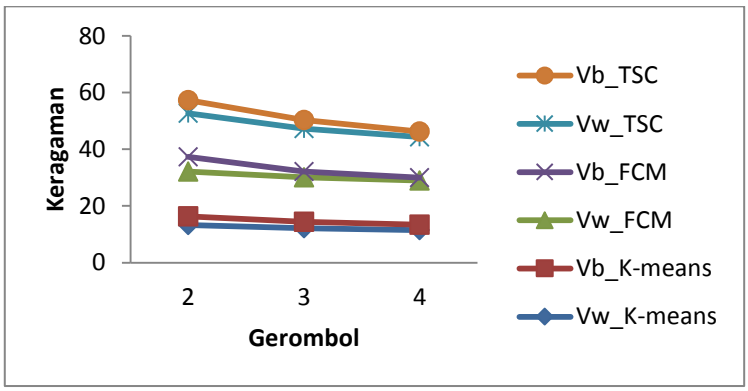
Tabel 15. Variansi 3 Gerombol

	k-means	*fuzzy k-means*	*two step cluster*
Variance Within Cluster	12.167	15.690	15.118
Variance Between Cluster	222.726	203.967	310.781

Tabel 16. Variansi 4 Gerombol

	k-means	*fuzzy k-means*	*two step cluster*
Variance Within Cluster	11.468	15.645	14.392
Variance Between Cluster	189.053	93.793	186.748

Dari tabel di atas dapat dilihat bahwa kehomogenan dalam *cluster* menurun seiring bertambahnya jumlah *cluster*,demikian juga halnya dengan keheterogenan antar *cluster*, untuk metode *k-means*, *fuzzy k-means*, dan *two step cluster* ternyata semakin bertambah jumlah *cluster* maka semakin menurun keheterogenan antar *cluster.* Kehomogenan semua algoritma ditunjukkan pada gambar berikut:



Gambar 13. Keragaman gerombol (*Variance cluster*)

Dari gambar 13 memperlihatkan bahwa performe ketiga metode tersebut hampir sama. Untuk masing-masing metode semakin bertambahnya jumlah gerombol maka keragaman dalam gerombolnya (*variance within cluster*) semakin menurun. Demikian halnya dengan keragaman antar kelompoknya (*variance between cluster*), semakin bertambahnya jumlah gerombol maka semakin menurun pula keragaman antar gerombolnya.

Untuk mengetahui *cluster* yang ideal, maka dapat dihat dari rasio perbandingan antara nilai keragaman dalam kelompok dengan nilai keragaman antar kelompok. Semakin kecil nilainya maka semakin bagus

pula *cluster* yang dihasilkan. Untuk lebih jelasnya disajikan pada tabel berikut:

Tabel 17. Rasio Variansi Gerombol

Gerombol	*k-means*	*fuzzy k-means*	*two step cluster*
2	0.042	0.030	0.033
3	0.054	0.076	0.048
4	0.060	0.166	0.077

Dari semua penggerombolan yang telah dilakukan dengan 2, 3 dan 4 gerombol, untuk semua metode yaitu metode *k-means*, *fuzzy k-means*, dan *two step cluster* maka hasil dari perbandingan keragaman dalam gerombol dengan keragaman antar gerombol menunjukkan bahwa, pada penggerombolan dengan 2 gerombol memiliki nilai yang jauh lebih kecil dibandingkan dengan 3 atau 4 gerombol. Ini berarti bahwa gerombol yang ideal adalah penggerombolan dengan 2 gerombol.

Kesimpulan

Berdasarkan uraian diatas dapat disimpulkan bahwa: (1) Jumlah gerombol ideal yang dihasilkan oleh masing-masing metode tersebut adalah 2 gerombol karena memiliki nilai rasio yang lebih kecil antara nilai rata-rata jumlah kuadrat dalam gerombol dengan antar gerombol; (2) Pada data Afifi, hasil dari masing-masing gerombol metode *k-means* dan *fuzzy k-means* lebih mirip pada penggerombolan 2 gerombol. Sedangkan metode *two step cluster* dari awal penggerombolan jumlah anggota gerombol yang agak jauh berbeda dengan kedua metode lainnya; (3) Jumlah anggota gerombol pada metode *two step cluster* agak jauh berbeda dibandingkan kedua metode lainnya diantaranya dipengaruhi oleh kesesuaian metode dengan jumlah data yang digunakan.

Daftar Pustaka

Agusta Y, 2007. K-Means-Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem dan Informatika Vol 3.* STIMIK. Bali.

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

Anonimous. 2001. *The SPSS TwoStep Cluster Component. A scalable component to segment your costumers more effectifely.* White paper-technical report, SPSS Inc Chicago.

Anonimous. 2004. *TwoStep Cluster Analysis*. Technical Report, SPSS Inc. Chicago.

Bacher, J., K. Wenzig and M. Vogler. 2004. SPSS TwoStep Cluster : A First Evaluation. Friedrich-Alexander-Universitat Erlangen-Nunberg.

Graham J Williams, 2008. *Data Mining Algorithms Cluster Analysis*. Adjunct Associate Professor, ANU.

Hong SL, 2006. *Experiment With K-Means, Fuzzy C-Means And Approaches To Choose K And C*. University of Central Florida. Orlando.

Kusdiati. 2006. Pengkajian Keakuratan *TwoStep Cluster* dalam menentukan Banyaknya Gerombol Populasi. Tesis. Departemen Statistika Institut Pertanian Bogor: IPB.

Kusumadewi, dkk. 2006. *Fuzzy Multi-Attribute Decision Making (FUZZY MADM)*. Yogyakarta. Graha Ilmu.

Santosa B, 2007. *Data Mining. Teknik Pemanfaatan Data Untuk Keperluan Bisnis.* Graha Ilmu. Yogyakarta.

Serban G, & Grigoreta SM. 2006. *A Comparison of Clustering Teqniques In Aspect Mining.* Studia Univ. Babes-Bolyai, Informatica, Volume L1.

Lathifaturrahmah
IAIN Antasari Banjarmasin
E-mail: lathifaturrahmah@iain-antasari.ac.id