KAJIAN METODE BERBASIS MODEL PADA ANALISIS KELOMPOK DENGAN PERANGKAT LUNAK MCLUST

Timbul Pardede (timbul@mail.ut.ac.id)

Jurusan Statistik FMIPA, Universitas Terbuka

ABSTRAK

Metode Ward dan metode K-rataan adalah metode kelompok yang teknik-teknik pengelompokannya hanya memperhatikan ukuran jarak antar objek-objek pengamatan tanpa mempertimbangkan aspek statistiknya. Metode kelompok berbasis model adalah metode kelompok yang didasarkan pada aspek statistik, yaitu kriteria kemungkinan maksimum. Metode kelompok berbasis model mempunyai sepuluh model dengan berbagai macam sifat geometris. Penyekatan data dilakukan dengan menggunakan algoritma Ekspektasi-Maksimum (EM), kemudian dengan pendekatan Bayesian Information Criterion (BIC) diperoleh model terbaik. Penelitian ini bertujuan untuk mengkaji efektivitas dari sepuluh metode berbasis model dan kemudian membandingkan hasil pengelompokannya dengan metode Ward dan metode K-rataan. Penelitian ini menggunakan data simulasi yang dibangkitkan melali program R versi 2.14.1 dan dianalisis dengan menggunakan program Mclust versi 4.0 dengan interface program R. Hasil penelitian menunjukkan bahwa metode kelompok berbasis model lebih efektif memisahkan kelompok-kelompok yang saling tumpang tindih dibandingkan dengan metode gerombol Ward dan K-rataan.

Kata kunci: Metode Ward, Metode K-mean, Metode berdasrkan model, Algoritma EM, BIC

ABSTRACT

Ward method and K-mean method are clustering method in which grouping only base on distance measure among observed objects, without considering statistical aspects. Model-based clustering is a method that use statistical aspects, as its theoretical basis i.e. probability maximum criterion. This model has ten models with a variety of geometrical characteristics. Data partition is conducted by utilizing EM (expectation-maximization) algorithm. Then by using Bayesian Information Criterion (BIC) the best model is obtained. This research aimed to assess the effectiveness of ten models from the model-based clustereng and then to compare result of grouping methods between model-based clustering with Ward clustering and K-mean clustering. This study used simulated data and applied data. Simulated data are generated with the R programs versions 2.14.1. Proses analysis was performed by using the Mclust programs vesions 4.0 with an interface the R programs versions 2.14.1. The results showed that model-based clustering was more effective in separating the condition of one separate group and two overlap groups than ward clustering and K-mean clustering.

Keywords: EM algorithm, BIC, K-mean clustering method, model-based clustering method, Ward clustering method

Analisis kelompok (*cluster analysis*) merupakan salah satu analisis statistik multivariat yang bertujuan untuk mengelompokan suatu objek amatan menjadi beberapa kelompok objek amatan berdasarkan karakteristik variabel-variabel yang dimiliki, sedemikian sehingga objek-objek yang terletak dalam kelompok yang sama cenderung mempunyai karakteristik yang lebih homogen dibandingkan dengan objek-objek pada kelompok yang berbeda. Pengelompokan objek amatan dilakukan berdasarkan ukuran kemiripan atau ketidakmiripan. Semakin tinggi kemiripan dua objek amatan maka semakin tinggi peluang untuk dikelompokan dalam suatu kelompok, sebaliknya semakin tinggi ketidakmiripannya maka semakin rendah peluang untuk dikelompokan dalam suatu kelompok.

Anderberg (1973) mengemukakan, terdapat beberapa metode kelompok yang dapat dibedakan berdasarkan proses algoritma yang dilakukan, diantaranya teknik yang berdasarkan ukuran jarak sebagai basis pengelompokannya. Metode kelompok berbasis ukuran jarak ini terdiri atas metode kelompok berhirarki dan metode kelompok tak berhirarki. Metode kelompok berhirarki, antara lain metode pautan tunggal (*single linkage*), metode pautan lengkap (*complete linkage*), metode pautan rataan (*average linkage*), metode terpusat (*centroid*), dan metode Ward (*Ward's method*). Adapun metode kelompok tak berhirarki, misalnya metode K-rataan. Metode kelompok ini memiliki teknik yang berbeda dalam proses pembentukan kelompok, namun teknik-teknik tersebut hanya memperhatikan ukuran jarak antar objek amatan. Metode-metode ini belum mempertimbangkan adanya aspek statistik.

Mc. Lachlan dan Basford (1988) memberikan suatu pendekatan dengan memperhatikan sebaran data yaitu analisis kelompok berbasis model (*model-based*). Metode kelompok berbasis model merupakan suatu algoritma kelompok dengan menggunakan analisis yang didasarkan pada aspek statistik di dalam menganalisis hasil kelompok. Fraley dan Raftery (1998) mengidentifikasi, enam model yang digunakan untuk mengelompokan objek amatan dengan berbagai sifat geometris yang diperoleh melalui komponen Gauss dengan parameter yang berbeda-beda. Pendekatan data dilakukan dengan menggunakan maksimum *likelihood* melalui algoritma Ekspektasi-Maksimum (EM), kemudian dengan pendekatan model Bayes berdasarkan *Bayesian Information Criterion* (BIC) diperoleh model terbaik. Pardede (2008), menggunakan enam model dengan metode berbasis model untuk membandingkan metode kelompok berbasis model dengan metode kelompok K-rataan. Kesimpulan yang diperoleh adalah metode berbasis model lebih baik dibandingkan metode K-rataan, akan tetapi pada kondisi kelompok yang saling tumpang tindih metode berbasis model dengan enam model belum mampu memisahkan objek-objek amatan. Fraley dan Raftery (1999) telah mengidentifikasi delapan model pada metode kelompok berbasis model yang digunakan untuk mengelompokan objek-objek amatan. Pada tahun 2010, Fraley dan Raftery telah mengidentifikasi sepuluh model untuk mengelompokan objek-objek amatan.

Berdasarkan paparan di atas, artikel ini mengkaji analisis kelompok berbasis model dengan sepuluh model yang telah diidentifikasi oleh Fraley dan Raftery (2010).Selanjutnya hasil analisis kelompok berbasis model ini dibandingkan dengan hasil analisis kelompok metode K-rataan dan metode Ward. Dari hasil analisis diperoleh efektivitas sepuluh model pada metode berbasis model, metode Ward dan metode K-rataan dalam mengelompokan objek-objek.

Metode Kelompok Ward

Pada metode kelompok berhirarki dengan penggabungan dianggap bahwa pada awalnya tiap-tiap objek amatan diperlakukan sebagai satu kelompok, sehingga jumlah kelompok yang ada sama dengan jumlah objek amatan. Kemudian menghitung jarak antar kelompok dengan kelompok

lainnya, dilanjutkan dengan menggabungkan dua kelompok terdekat menjadi satu kelompok baru. Langkah berikutnya jarak antara kelompok baru dengan kelompok lainnya dihitung kembali. Prosedur ini diulang terus hingga terbentuk suatu diagram pohon yang hanya terdiri atas satu kelompok yang beranggotakan semua objek amatan. Hasil kelompok metode berhirarki membentuk diagram pohon (*tree diagram*) atau *dendrogram* yang menggambarkan pengelompokan objek-objek amatan. Salah satu metode dari metode kelompok berhirarki dengan penggabungan adalah metode Ward. Metode Ward didasarkan pada kriteria jumlah kuadrat antara dua kelompok untuk seluruh variabel.

$$\delta_1 = \frac{n_k + n_i}{n_k + n_i + n_j},\ \delta_2 = \frac{n_k + n_j}{n_k + n_i + n_j}\ \ \delta_3 = \frac{n_k}{n_k + n_i + n_j}$$

dengan

nilai koefisien $\delta_1, \delta_2, \delta_3$ dan δ_4 sebagai faktor pembobot

$$d_{(i,j)k} = \delta_1 d_{ik} + \delta_2 d_{jk} + \delta_3\, d_{ij} + \delta_4 \left| d_{ik} - d_{jk} \right|$$

Metode Kelompok K-rataan

Salah satu metode tak berhirarki yang paling sering digunakan adalah metode kelompok K-rataan. Metode ini merupakan metode kelompok yang menyekat objek amatan ke dalam *k* kelompok. Metode ini pada umumnya digunakan pada data yang berukuran relatif besar.

Macqueen *dalam* Johnson dan Wichern (2007) menggambarkan algoritma kelompok untuk menyeleksi *n* unit data ke dalam *k* kelompok adalah berdasarkan kedekatan pusat (rataan) yang disusun dengan tahapan berikut:

1. Pilih *k* unit data pertama yang digunakan sebagai *k* pusat kelompok awal.
2. Gabungkan setiap *(n-k)* data yang merupakan sisa anggota ke pusat kelompok terdekat, kemudian dihitung masing-masing pusat (rataan) kelompok baru yang terbentuk dari hasil penggabungan.
3. Setelah semua data digabungkan pada tahap 2, pusat kelompok yang terbentuk dijadikan sebuah titik pusat kelompok, kemudian lakukan penggabungan kembali dari setiap unit data ke dalam titik pusat terdekat.
4. Lakukan ketiga tahapan di atas secara berulang hingga diperoleh suatu kelompok yang konvergen. Kelompok yang konvergen ditandai dengan adanya titik pusat yang tetap dan tidak ada lagi perubahan anggota di antara kelompok.

Metode Kelompok Berbasis Model

Pada analisis kelompok berbasis model, diasumsikan bahwa data dibangkitkan dari sebaran peluang campuran dengan setiap subpopulasi mewakili suatu kelompok yang berbeda (Fraley & Raftery, 1998). Misalnya $y = (y_1, y_2, ..., y_n)$ variabel acak ganda *p*, dengan *p* menyatakan dimensi data dan *n* menyatakan banyaknya objek amatan yang dianggap berasal dari campuran G subpopulasi $G_1, G_2, ..., G_g$ dengan fungsi kepekatan campurannya adalah:

$$\mathrm{f}_{mix}(y) = \sum_{k=1}^{G} \tau_k \, \mathrm{f}_k(y|\theta) \quad ; \quad y \in \Omega \qquad (1)$$

dengan $\mathrm{f}_k\left(y|\theta\right)$: fungsi kepekatan G_k, yaitu subpopulasi ke-*k* dengan vektor parameter θ yang tidak diketahui dan τ_k: proporsi data yang berasal dari subpopulasi ke-*i* dengan $\sum_{\mathrm{j=1}}^{\mathrm{G}}\tau_i = 1$ dan $\tau_i \geq 0$.

Dengan asumsi $y = \left(y_1, y_2, ..., y_n\right)$ bebas stokastik dan identik, dan fungsi $\mathrm{f}_k\left(y_i|\theta_k\right)$ merupakan fungsi kepekatan campuran objek amatan y_i dari kelompok ke-*k* maka fungsi kepekatan sebaran campuran (*mixture likelihood*) pada persamaan (1) adalah :

$$L_{mix}(\theta_1...\theta_G;\tau_1...\tau_G \mid y) = \prod_{i=1}^{n}\left[\sum_{k=1}^{G}\tau_k\, f_k\left(y_i|\theta_k\right)\right] \qquad (2)$$

dengan $f_k\left(y_i|\theta_k\right)$ adalah fungsi kepekatan variabel campuran normal ganda (*Gauss*) dengan parameter θ_k terdiri atas vektor rataan μ_k dan matriks kovariansi $\sum_k$, yang dinyatakan dalam bentuk: $f_k(y_i|\mu_k;\sum_k) = \dfrac{\exp\left\{-\frac{1}{2}\left(y_i-\mu_k\right)'\sum_k{}^{-1}\left(y_i-\mu_k\right)\right\}}{\left(2\pi\right)^{\frac{p}{2}}\left|\sum_k\right|^{\frac{1}{2}}}$.

Dengan demikian fungsi kepekatan sebaran campuran (*mixture likelihood*) ganda dengan parameter vektor rataan μ_k dan matriks kovariansi $\sum_k$ dapat ditulis dalam bentuk:

$$L_{mix}(\mu_1;\sum_1...\mu_k;\sum_k;\tau_1...\tau_G \mid y) = \prod_{i=1}^{n}\left[\sum_{k=1}^{G}\tau_k\,\frac{\exp\left\{-\frac{1}{2}\left(y_i-\mu_k\right)'\sum_k{}^{-1}\left(y_i-\mu_k\right)\right\}}{\left(2\pi\right)^{\frac{p}{2}}\left|\sum_k\right|^{\frac{1}{2}}}\right] \qquad (3)$$

Pada metode kelompok berbasis model, diasumsikan bahwa data dibangkitkan dengan fungsi kepekatan variabel campuran normal ganda yang dicirikan oleh kelompok-kelompok yang berpusat di sekitar μ_k. Karakteristik geometrik (bentuk, volume, dan orientasi) dihitung dari matriks kovariansi $\sum_k$ (Fraley & Raftery, 2002).

Branfield & Raftery (1993) mengembangkan metode kelompok berbasis model dengan memparameterisasikan setiap matriks kovariansi melalui suku-suku dekomposisi nilai ciri dalam bentuk:

$$\sum_k = \lambda_k D_k A_k D_k' \qquad (4)$$

dengan :

- D_k : matriks vektor ciri, yang menjelaskan orientasi dari komponen ke-*k*,
- A_k : matriks diagonal dengan masing-masing unsurnya proporsional terhadap nilai ciri dari $\sum_k$, yang menjelaskan bentuk,
- λ_k : akar ciri terbesar dari $\sum_k$, yang menjelaskan volume.

Pencirian sebaran geometrik (orientasi, volume, dan bentuk) diperoleh dari berbagai macam bentuk kelompok, atau terbatas pada kelompok yang sama dan matriks varians untuk semua komponen bisa

sama atau bervariasi. Tabel 1 menunjukkan matriks kovariansi $\sum_j$ untuk model campuran normal ganda dan interpretasi geometrik (Fraley & Raftery, 2010).

Tabel 1. Matriks Kovariasi $\sum_k$ dan Interpretasi Geometrik pada Model Campuran Normal Ganda

$\sum_j$	Volume	Bentuk Geometri	Orientasi	Tebaran	Simbol Mclust
$\lambda \mathrm{I}$	Sama	Sama	-	*Spherical*	EII
$\lambda_k I$	Berbeda	Sama	-	*Spherical*	VII
λA	Sama	Sama	Sumbu koordinat	Diagonal	EEI
$\lambda_k A$	Berbeda	Sama	Sumbu koordinat	Diagonal	VEI
λA_k	Sama	Berbeda	Sumbu koordinat	Diagonal	EVI
$\lambda_k A_k$	Berbeda	Berbeda	Sumbu koordinat	Diagonal	VVI
$\lambda DAD'$	Sama	Sama	Sama	*Ellipsoidal*	EEE
$\lambda D_k A D_k'$	Sama	Sama	Berbeda	*Ellipsoidal*	EEV
$\lambda_k D_k A D_k'$	Berbeda	Sama	Berbeda	*Ellipsoidal*	VEV
$\lambda_k D_k A_k D_k'$	Berbeda	Berbeda	Berbeda	*Ellipsoidal*	VVV

Sumber: (Fraley & Raftery, 2010)

Algoritma EM (*Expectation-Maximum*)

Algoritma EM merupakan metode perhitungan iterasi terhadap masalah pendugaan kemungkinan maksimum parameter pada data tidak lengkap (Dempster, Laird, and Rubin, 1977). Algoritma EM pada kelompok data lengkap diasumsikan menjadi $\boldsymbol{x}_i' = (\boldsymbol{y}_i', \boldsymbol{z}_i')$, dengan $\boldsymbol{y}_i'$ data teramati dan $\boldsymbol{z}_i'$ data yang tidak teramati (*missing*). Apabila $\boldsymbol{x}_i'$ adalah data yang berdistribusi bebas dan identik menurut distribusi peluang *f* dengan parameter θ maka fungsi *likelihood* data lengkap adalah $\mathrm{L_C}(x_i|\theta) = \prod_{i=1}^{n} f_j(x_i|\theta)$. Selanjutnya jika peluang variabel khusus tidak teramati dan tergantung pada pengamatan data $\boldsymbol{y}$ dan bukan $\boldsymbol{z}$ maka fungsi *likelihood* data lengkap menjadi:

$$\mathrm{L_O}(\boldsymbol{y}|\theta) = \int \mathrm{L_C}(\boldsymbol{x}|\theta) d\boldsymbol{z} \qquad (5)$$

Penduga maksimum *likelihood* (MLE) untuk paramater θ didasarkan pada proses pemaksimuman data pengamatan $\mathrm{L_O}(\boldsymbol{y}|\theta)$.

Pada EM untuk model campuran, data lengkap diasumsikan $\boldsymbol{x}_i' = (\boldsymbol{y}_i', \boldsymbol{z}_i')$ dengan $\boldsymbol{z}_i' = (z_{i1}, z_{i2}, ..., z_{ig})$ merupakan data yang tidak teramati, yaitu

$$\mathrm{z}_{ik} = \begin{cases} 1, & x_i \in G_k \\ 0, & \text{lainnya.} \end{cases} \quad ; i = 1, ..., n \;\; ; k = 1, ..., g \qquad (6)$$

dengan asumsi bahwa setiap $\boldsymbol{z}_i$ bebas dan identik menurut sebaran multinomial dari *G* kategori dengan peluang $\tau_1, \tau_2, ..., \tau_G$ dan fungsi kepekatan $\boldsymbol{y}_i$ terhadap $\boldsymbol{z}_i$ adalah $\prod_{k=1}^{G} f_k\left(y_i|\theta_k\right)^{z_{ik}}$, maka fungsi *log-likelihood* data lengkap (*complete-data log-likelihood*) adalah :

$$L(\theta_k, \tau_k, z_{ik}|x) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik} \log\left\{\tau_k f_k\left(x_i|\theta_k\right)\right\} \qquad (7)$$

Bila $f_k\left(x_i|\theta_k\right)$ merupakan model campuran sebaran normal ganda yaitu $f_k\left(x_i|\theta_k\right) = f_k\left(x_i|\mu_k;\Sigma_k\right)$, maka fungsi *log-likelihood* data lengkap pada model campuran normal ganda adalah:

$$L(\theta_k, \tau_k, z_{ik}|x) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik} \log\left\{\tau_k f_k\left(x_i|\mu_k;\Sigma_k\right)\right\} \qquad (8)$$

Dengan menggunakan algoritma EM , yaitu tahap E untuk pendugaan dan tahap M untuk pemaksimuman, maka iterasi tahap E model campuran normal ganda diperoleh

$$\hat{z}_{ik} = \frac{\hat{\tau}_k\, f_k\left(y_i\left|\hat{\mu}_k, \hat{\Sigma}_k\right.\right)}{\sum_{j=1}^{G} \hat{\tau}_j\, f_j\left(y_i\left|\hat{\mu}_j, \hat{\Sigma}_j\right.\right)} \qquad ; i = 1,...,n \; ; k = 1,...,G \qquad (9)$$

Sedangkan tahap M adalah untuk memaksimalkan persamaan (8) terhadap τ_k dan θ_k dengan z_{ik} tetap pada nilai yang dihitung pada tahap E.

Pemilihan Model Pengelompokan dengan Faktor Bayes

Pada analisis kelompok masalah yang paling sering menjadi pertanyaan adalah bagaimana menentukan metode kelompok yang digunakan dan berapa jumlah kelompok yang tepat. Seringkali para pengguna statistik melakukan coba-coba (*trial and error*) untuk mendapatkan hasil yang bermakna atau yang dapat diinterpretasikan sesuai dengan masalah kajiannya. Fraley & Raftery (1998) melakukan pendekatan model campuran melalui faktor Bayes dengan sistematika pemilihannya tidak hanya untuk parameterisasi model (metode kelompok yang digunakan), tetapi juga banyaknya kelompok. Pendekatan yang digunakan adalah dengan pendekatan BIC (*Bayesian Information Criterion*) dengan formulasi sebagai berikut:

$$2\log P\left(y|M_k\right) \approx 2\log P\left(y\left|\hat{\theta}_k, M_k\right.\right) - V_k \log(n) \equiv BIC_k$$

dengan

$P\left(y|M_k\right)$: integrasi *likelihood* untuk model M_k ,

$P\left(y\left|\hat{\theta}_k, M_k\right.\right)$: maksimum *likelihood* campuran untuk model M_k ,

V_k : banyaknya parameter bebas yang diduga pada model M_k ,

$\hat{\theta}_k$: dugaan kemungkinan maksimum untuk parameter θ pada model M_k.

Nilai BIC terbesar mengindikasikan bahwa model tersebut adalah model paling layak.

METODE

Sumber data yang digunakan adalah data simulasi dari himpunan campuran normal ganda yang dibangkitkan dengan menggunakan fungsi *mvnorm* pada perangkat lunak program R ver 2.14.1. Kriteria data simulasi yang dibangkitkan mengacu pada Pardede (2008), yakni data yang dibangkitkan dari sebaran campuran normal ganda yang terdiri atas tiga kelompok dengan tiga variabel dan jumlah amatan tiap kelompok adalah 50, 100, dan 150. Ketiga kelompok yang dibangkitkan dibuat dalam tiga macam kondisi, yaitu (1) ketiga kelompok saling terpisah, (2) satu kelompok terpisah dan dua kelompok tumpang tindih, dan (3) ketiga kelompok saling tumpang tindih. Untuk ketiga kondisi data, digunakan tiga jenis ukuran jarak antara dua pusat kelompok, yang disesuaikan dengan jauh dekatnya jarak antara vektor rataan kelompok. Selain itu, untuk melihat pengaruh tingkat korelasi antarvariabel terhadap hasil akhir kelompok, maka dicobakan juga tiga jenis tingkat korelasi. Pola-pola data simulasi secara lengkap disajikan pada Tabel 2.

Tabel 2. Pola Data Simulasi yang Akan Dibangkitkan

Jenis pengelompokan	Jarak antar dua pusat kelompok dan nilai variansi tiap variabel	Tingkat korelasi antar variabel		Banyak data tiap kelompok
Ketiga kelompok saling terpisah	Dekat, Sedang, Jauh $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$ (variansi kecil)	Rendah Sedang Tinggi	(0.2) (0.5) (0.8)	50 100 150
Satu terpisah, dua tumpang tindih	Dekat, Sedang, Jauh $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$ (variansi berbeda)	Rendah Sedang Tinggi	(0.2) (0.5) (0.8)	50 100 150
Ketiga kelompok saling tumpang tindih	Dekat, Sedang, Jauh $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$ (variansi besar)	Rendah Sedang Tinggi	(0.2) (0.5) (0.7)	50 100 150

Sumber: (Pardede, 2008)

Hasil data simulasi dilakukan tahapan analisis sebagai berikut:

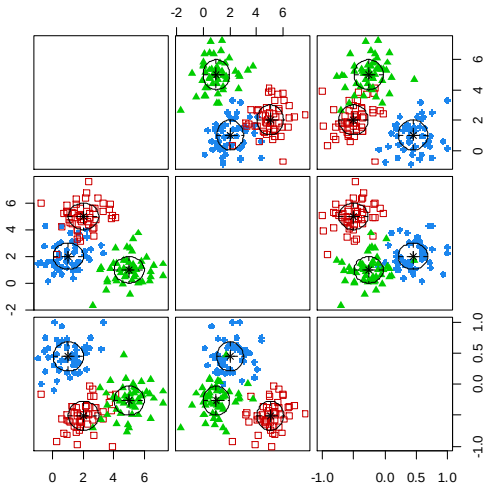
1. Dilakukan analisis kelompok dengan menggunakan perangkat lunak R ver 2.14.1 untuk metode Ward dan metode K-rataan. Untuk metode berbasis model, analisis kelompok dilakukan dengan menggunakan paket program Mclust ver 4.0 dengan *interface* perangkat lunak R ver 2.14.1 (*open source*).
2. Untuk metode berbasis model, hitung nilai BIC pada setiap model. Nilai BIC terbesar mengindikasikan bahwa model tersebut adalah model yang paling layak.
3. Hitung persentase salah pengelompokan dari setiap kelompok pada masing-masing metode, kemudian hasilnya dibandingkan dengan pengelompokan yang sebenarnya (ditentukan saat simulasi).
4. Persentase salah pengelompokan yang terkecil menunjukkan bahwa metode yang digunakan lebih baik.

HASIL DAN PEMBAHASAN

Data simulasi yang dibangkitkan terdiri atas 81 kasus data dengan setiap kasus data simulasi terdiri atas tiga kelompok. Semua kasus data dibedakan atas kondisi kelompok yakni jarak antarpusat kelompok dengan variansi setiap variabel sama atau berbeda pada setiap kelompok, tingkat korelasi, dan juga banyak data.

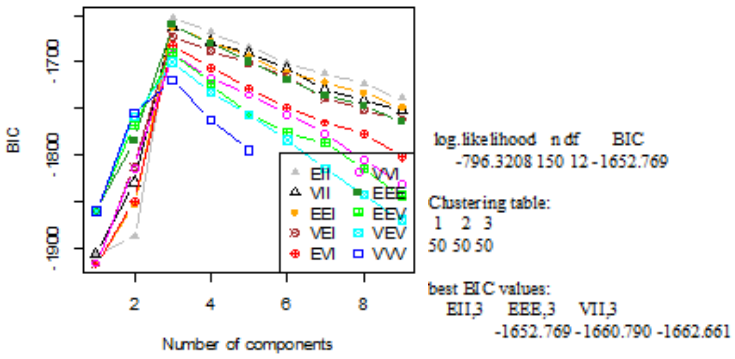
Kondisi Ketiga Kelompok Saling Terpisah

Pada kasus data simulasi dengan kondisi pengelompokan ketiga kelompok saling terpisah menunjukkan bahwa hasil pengelompokan yang diperoleh dengan metode Ward, metode K-rataan, dan metode berbasis model dapat mengelompokkan objek-objek amatan secara tepat dan sesuai dengan pengelompokan yang sebenarnya. Hal ini dintunjukan dengan nilai persentase salah pengelompokan pada metode Ward, metode K-rataan, dan metode berbasis model adalah 0%. Hal ini disebabkan oleh variansi dari masing-masing kelompok cenderung kecil sehingga setiap objek-objek amatan cenderung mengelompok di sekitar vektor rataan kelompok seperti terlihat pada Gambar 1.



Gambar 1. Matriks plot data simulasi untuk kondisi ketiga kelompok saling terpisah dengan jarak pusat kelompok dekat, variansi cenderung kecil, tingkat korelasi rendah, dan n=50

Hasil kelompok dengan metode berbasis model pada Gambar 2 menunjukkan bahwa dari 10 model yang dicobakan 3, model yang paling layak yang didasarkan pada nilai BIC paling besar, yaitu model EII dengan nilai BIC = -1652,769; model EEE dengan nilai BIC = -1660.790; dan model VII dengan nilai BIC = -1662.661. Nilai BIC yang paling besar dari tiga model yang paling layak terdapat pada model EII dengan nilai BIC = -1652,769 maka model terbaik terdapat pada model EII.



Gambar 2. Plot dan hasil kelompok metode berbasis model untuk kondisi ketiga kelompok saling terpisah dengan jarak pusat kelompok dekat dan variansi cenderung kecil, tingkat korelasi rendah, dan banyak data n=50.

Dari 27 kasus simulasi dengan kondisi ketiga kelompok saling terpisah, model terbaik terdapat pada model EEE yang tebaran datanya berbentuk *ellipsoidal*, kecuali pada kondisi banyak data *n*=50 dengan tingkat korelasi rendah (0,2) dan jarak antar pusat kelompok dekat, sedang, jauh menghasilkan model terbaik EII yang berbentuk *Spherical*, dan pada tingkat korelasi sedang (0,5) menghasilkan model terbaik EEV yang berbentuk *ellipsoidal.* Nilai BIC dan model terbaik pada metode berbasis model disajikan pada Tabel 3. Persentase salah pengelompokan kelompok tidak terpengaruh terhadap tingkat korelasi antarvariabel dan banyak objek amatan pada tiap kelompok.

Tabel 3 . Nilai BIC dan Model Terbaik pada Metode Kelompok Berbasis Model

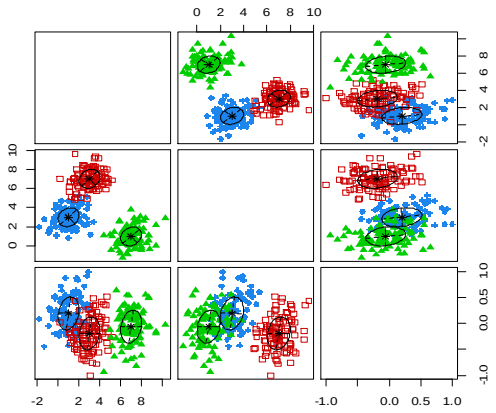
Banyak data	Jarak antar pusat kelompok	Tingkat korelasi	Ketiga kelompok saling terpisah $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=1$		satu kelompok terpisah, dua kelompok tindih $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=25$		Keiga kelompok saling tumpang tindih $\sigma_1^2=25,\sigma_2^2=25,\sigma_3^2=25$	
			Nilai BIC	Model	Nilai BIC	Model	Nilai BIC	Model
50	Dekat (d=5.099)	Rendah	-1652,769	EII	-2116,695	EEI	-2786,649	VII
		Sedang	-1578,400	EEE	-2051,198	EEE	-2740,561	EII
		Tinggi	-1343,213	EEE	-1826,012	EEE	-2639,369	EEE
	Sedang (d=7.483)	Rendah	-1657,665	EII	-2137,900	EEE	-2838,331	VII
		Sedang	-1578,746	EEE	-2061,239	EEE	-2826,904	EII
		Tinggi	-1343,213	EEE	-1826,045	EEE	-2726,018	EEE
	Jauh (d= 9.899)	Rendah	-1657,665	EII	-2148,996	EEE	-2936,891	EII
		Sedang	-1578,746	EEV	-2061,576	EEE	-2925,486	EII
		Tinggi	-1343,213	EEE	-1826,045	EEE	-2764,566	EEE
100	Dekat (d=5.099)	Rendah	-3264,336	EEE	-4.205,786	EEE	-5595,077	EII
		Sedang	-3093,097	EEE	-4018,418	EEE	-5489,043	EII
		Tinggi	-2622,166	EEE	-3587,805	EEE	-5211,399	EEE
	Sedang (d=7.483)	Rendah	-3268,231	EEE	-4216,587	EEE	-5709,513	EII
		Sedang	-3093,232	EEE	-4058,344	EEE	-5637,339	EII
		Tinggi	-2622,166	EEE	-3.587,829	EEE	-5403,803	EEE
	Jauh (d= 9.899)	Rendah	-3268,231	EEE	-4233,654	EEE	-5833,120	VII
		Sedang	-3093,232	EEE	-4058,876	EEE	-5799,674	EEV
		Tinggi	-2622,166	EEE	-3587,829	EEE	-5486,362	EEE

Tabel 3 . Lanjutan

Banyak data	Jarak antar pusat kelompok	Tingkat korelasi	Ketiga kelompok saling terpisah $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=1$		satu kelompok terpisah, dua kelompok tindih $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=25$		Keiga kelompok saling tumpang tindih $\sigma_1^2=25,\sigma_2^2=25,\sigma_3^2=25$	
			Nilai BIC	Model	Nilai BIC	Model	Nilai BIC	Model
150	Dekat (d=5.099)	Rendah	-4858,828	EEE	-6221,692	EEE	-8349,897	VII
		Sedang	-4601,239	EEE	-6009,734	EEE	-8197,145	EII
		Tinggi	-3896,198	EEE	-6300,827	EEE	-7812,342	EEE
	Sedang (d=7.483)	Rendah	-4865,295	EEE	-5698,814	EEE	-8537,436	EII
		Sedang	-4602,796	EEE	-6047,875	EEE	-8432,451	EEE
		Tinggi	-3896,198	EEE	-5344,670	EEE	-8.060,247	EEE
	Jauh (d= 9.899)	Rendah	-4865,295	EEE	-6312,636	EEE	-8726,841	EII
		Sedang	-4602,796	EEE	-6051,276	EEE	-8631,423	EEE
		Tinggi	-3896,198	EEE	-5344,692	EEE	-8186,772	EEE

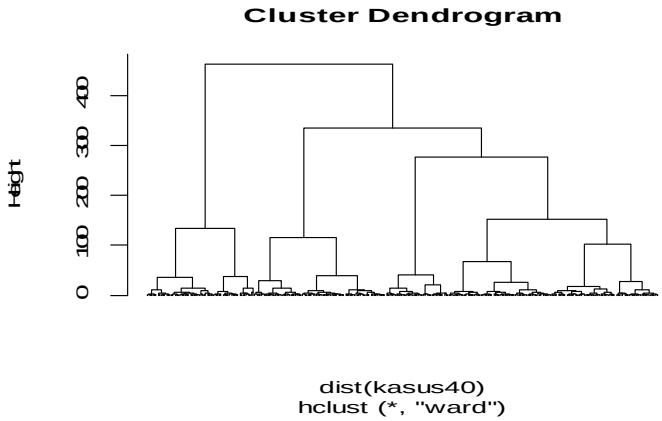
Kondisi Satu Kelompok Terpisah dan Dua Kelompok Tumpang Tindih

Untuk kasus data simulasi dengan satu kelompok terpisah dan dua kelompok tumpang tindih disajikan pada Gambar 3 dengan kondisi jarak antarpusat kelompok sedang (d=7.483), variansi ketiga kelompok adalah $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=25$, tingkat korelasi antarvariabel rendah (0.2), dan banyak amatan tiap kelompok adalah *n*=100.



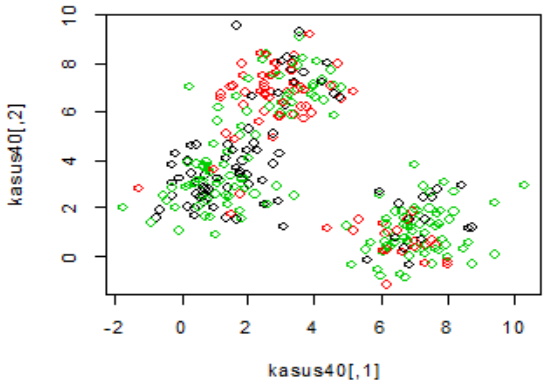
Gambar 3. Matriks plot data simulasi untuk kondisi satu kelompok terpisah dan dua kelompok tumpang tindih dengan jarak sedang, variansi $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=25$, tingkat korelasi sedang, dan *n*=100.

Hasil pengelompokan dengan metode Ward diperoleh hasil bahwa dari 100 objek amatan pada kelompok 1 terdapat 2 objek amatan masuk ke dalam kelompok 2, dari 100 objek amatan pada kelompok 2 terdapat 62 objek amatan masuk ke dalam kelompok 1, dan dari 100 objek amatan pada kelompok 3 terdapat 23 objek amatan masuk ke kelompok 2 (Gambar 4). Persentase salah pengelompokannya adalah 29,00%.



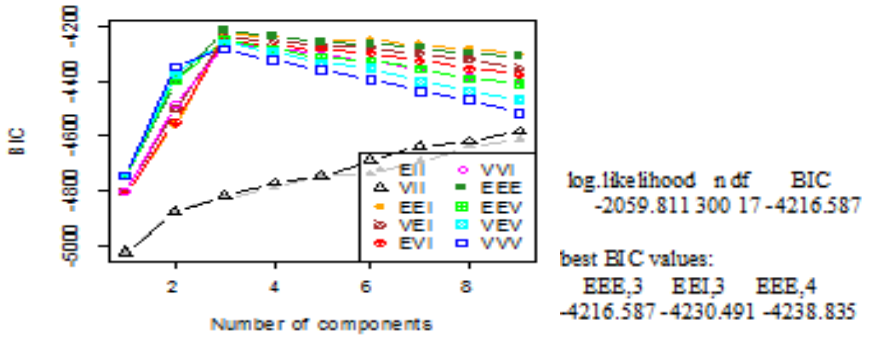
Gambar 4. Dendogram hasil pengelompokan metode Ward dengan kondisi satu kelompok terpisah dan dua kelompok tumpang tindih dengan jarak sedang, variansi $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$, tingkat korelasi sedang, dan *n*=100.

Pada metode K-rataan, hasil pengelompokan yang diperoleh adalah pada kelompok 1 terdapat 20 objek amatan masuk ke dalam kelompok 2 dan 23 objek amatan masuk ke dalam kelompok 3 dan hanya 57 objek amatan yang dengan tepat masuk ke dalam kelompok 1. Untuk kelompok 2, terdapat 50 objek amatan masuk ke dalam kelompok 1 dan 15 objek amatan masuk ke dalam kelompok 3, dan hanya 35 objek amatan dengan tepat masuk ke dalam kelompok 3. Hal ini menunjukkan bahwa lebih dari 50% objek amatan tidak terkelompok pada tempatnya (Gambar 5).



Gambar 5. Plot hasil pengelompokan metode K rataan dengan kondisi satu kelompok terpisah dan dua kelompok tumpang tindih dengan jarak sedang, variansi $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$, tingkat korelasi sedang, dan *n*=100.

Pada metode berbasis model, dari 10 model yang dicobakan terdapat tiga model yang paling layak, yakni model EEE (3 kelompok) dengan nilai BIC = -4216.587; model EEI (3 kelompok) dengan nilai BIC = -4230.491; dan model EEE (4 kelompok) dengan nilai BIC = -4238.835. Model terbaik dari tiga model yang paling layak terdapat nilai BIC yang paling besar yaitu pada model EEE (Gambar 6).



Gambar 6. Plot dan hasil pengelompokan metode berbasis model dengan kondisi satu kelompok terpisah dan dua kelompok tumpang tindih dengan jarak pusat kelompok sedang, variansi $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$, tingkat korelasi sedang dan *n*=100.

Untuk metode berbasis model, dari 27 kasus simulasi pada kondisi satu kelompok terpisah dan dua kelompok tumpang tindih menghasilkan model yang sama untuk setiap kasus yaitu model EEE yang tebaran datanya berbentuk *ellipsoidal*. Kecuali pada kondisi jarak antarpusat kelompok dekat, banyak data n=50 dan tingkat korelasi sedang menghasil model EEV (*ellipsoidal*) sebagai model terbaik. Nilai BIC dan Model terbaik pada metode berbasis model disajikan pada Tabel 3.

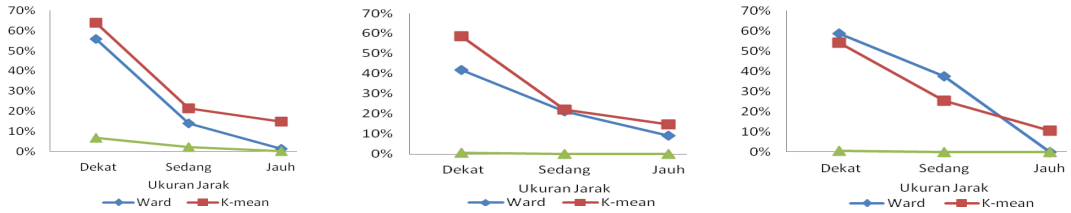
Untuk 27 kasus dengan kondisi satu kelompok terpisah dan dua kelompok saling tumpang tindih diperoleh hasil bahwa persentase salah pengelompokan pada metode berbasis model jauh lebih kecil dibandingkan dengan metode Ward dan metode K-rataan. Hal ini menunjukkan bahwa metode berbasis model lebih cenderung dapat memisahkan ketiga kelompok dibandingkan dua metode kelompok lainnya. Persentase salah pengelompokan pada kondisi satu kelompok terpisah dan dua kelompok tumpang tindih disajikan pada Tabel 4.

Tabel 4. Persentase Salah Pengelompokan pada Kondisi Satu Kelompok Terpisah dan Dua Kelompok Saling Tumpang Tindih.

Jarak antar pusat kelompok	Metode	*n*=50			*n*=100			*n*=150		
		Tingkat Korelasi			Tingkat Korelasi			Tingkat Korelasi		
		Rendah	Sedang	Tinggi	Rendah	Sedang	Tinggi	Rendah	Sedang	Tinggi
Dekat	Ward	56,00%	42,00%	58,67%	55,67%	53,00%	56,33%	52,00%	51,78%	42,00%
	K-rataan	64,00%	58,67%	54,00%	56,67%	57,67%	55,67%	60,67%	60,00%	56,89%
	Model	6,67%	0,67%	0,67%	3,33%	2,00%	0,00%	4,44%	2,67%	1,11%
Sedang	Ward	14,00%	21,33%	37,33%	29,00%	16,00%	13,00%	27,78%	30,44%	21,11%
	K-rataan	21,33%	22,00%	25,33%	55,67%	29,33%	7,67%	24,44%	50,22%	45,11%
	Model	2,00%	0,00%	0,00%	1,33%	0,00%	0,00%	0,89%	0,22%	0,00%
Jauh	Ward	1,33%	9,33%	10,67%	50,67%	0,00%	0,00%	10,44%	0,00%	0,00%
	K-rataan	14,67%	14,67%	0,00%	55,67%	11,67%	10,67%	12,22%	11,33%	11,78%
	Model	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%

Ditinjau dari jarak antarpusat kelompok, terjadi penurunan persentase salah pengelompokan dengan semakin jauh jarak antar pusat kelompok untuk ketiga metode kelompok, yang disajikan pada Gambar 7. Penurunan salah persentase pengelompokan ini disebabkan oleh ukuran jarak

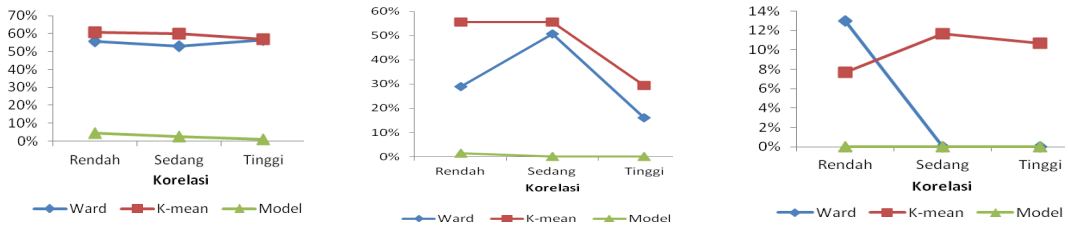
antarvektor rataan kelompok yang relatif makin jauh untuk semua kondisi, sehingga objek-objek amatan akan semakin mengelompok di sekitar vektor rataannya.



(a) (b) (c)

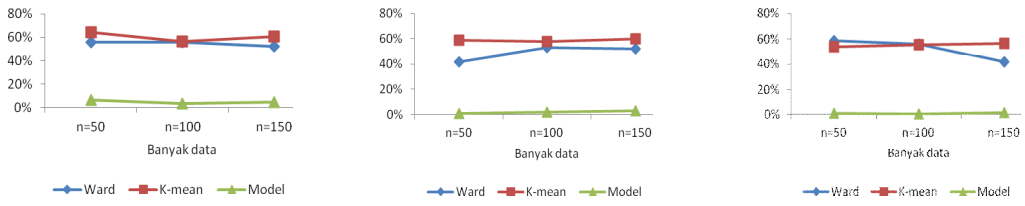
Gambar 7. Plot persentase salah pengelompokan pada ukuran jarak dengan tingkat korelasi (a)rendah, (b)sedang, dan (c)tinggi dengan n=50.

Untuk tingkat korelasi rendah, sedang, dan tinggi menunjukkan bahwa pada metode berbasis model terjadi penurunan persentase salah pengelompokan dari tingkat korelasi rendah ke tingkat korelasi tinggi, walaupun penurunan ini hampir tidak ada perbedaan yang berarti. Hal ini menunjukkan bahwa tingkat korelasi yang berbeda tidak berpengaruh secara signifikan pada kondisi kelompok pada kondisi satu kelompok terpisah dan dua kelompok tumpang tindih (Gambar 8).



(a) (b) (c)

Gambar 8. Persentase salah pengelompokan yang didasarkan pada tingkat korelasi dengan ukuran jarak (a) dekat, (b) sedang, dan (c) jauh dengan banyak data *n*=100.



(a) (b) (c)

Gambar 9. Persentase salah pengelompokan didasarkan pada banyaknya data dengan tingkat korelasi (a) rendah, (b) sedang, dan (c) tinggi dengan jarak antar pusat kelompok dekat.

Ditinjau dari banyak objek amatan, banyak amatan tiap kelompok n= 50 mempunyai pola persentase salah pengelompokan yang tidak jauh berbeda dengan objek amatan tiap kelompok sebesar 100 dan 150. Hal ini berarti bahwa banyak amatan tiap kelompok tidak berpengaruh terhadap hasil kelompok (Gambar 9).

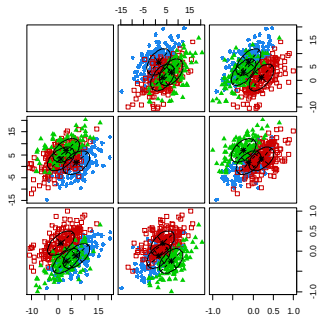
Dari hasil pengelompokan ketiga metode kelompok yang dibandingkan dengan kondisi satu kelompok terpisah dan dua kelompok tumpang tindih menunjukkan bahwa metode berbasis model lebih efektif dalam memisahkan kelompok-kelompok kelompok dibandingkan metode Ward dan metode K-rataan.

Ketiga Kelompok Saling Tumpang Tindih

Pada Gambar 10 disajikan data dengan kondisi jarak antarpusat kelompok sedang (d=7.483) dengan variansi ketiga kelompok adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, tingkat korelasi antar variabel sedang (0.5), dan *n*=150. Secara visual matriks plot pada Gambar 10 menggambarkan kondisi ketiga kelompok saling tumpang tindih.

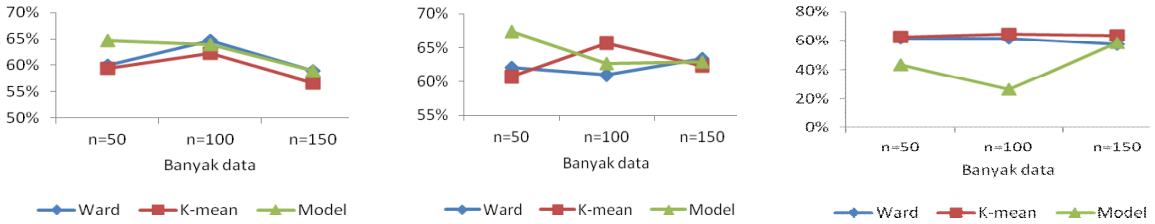
Secara keseluruhan dari 27 kasus simulasi pada kondisi ketiga kelompok saling tumpang tindih, ketiga metode kelompok tidak mampu memisahkan kelompok yang saling tumpang tindih. Metode berbasis model menghasilkan model yang bervariasi pada setiap kasus simulasi, yakni model dengan tebaran datanya berbentuk *spherical* (VII, EII) dan model yang tebaran datanya berbentuk *ellipsoidal* (EEE, EEV). Tingkat korelasi rendah dan sedang menghasilkan model VII dan EII, sedangkan tingkat korelasi tinggi menghasilkan model EEE. Nilai BIC dan Model terbaik pada metode berbasis model disajikan pada Tabel 3.

Pada tingkat korelasi tinggi dengan jarak antar pusat kelompok relatif sedang dan jauh, persentase salah pengelompokan lebih rendah bila dibandingkan dengan metode Ward dan metode K-rataan. Namun pada tingkat korelasi rendah dan tingkat korelasi sedang, metode berbasis model tidak mampu memisahkan kelompok yang saling tumpang tindih, bahkan sebagian besar persentase salah pengelompokannya lebih besar dibandingkan dengan metode Ward dan metode K-rataan. Hal ini disebabkan oleh objek amatannya mengelompok pada satu kelompok, sehingga secara geometris dari 10 model pada metode berbasis model tidak mampu memisahkan kelompok yang saling tumpang tindih. Bahkan metode berbasis model ini menganjurkan bahwa akan lebih efektif jika pengelompokannya dibentuk dalam satu atau dua atau empat kelompok.



Gambar 10. Matriks plot data untuk kondisi ketiga kelompok saling tumpang tindih, jarak sedang, variansi $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, tingkat korelasi sedang, dan *n*=150.

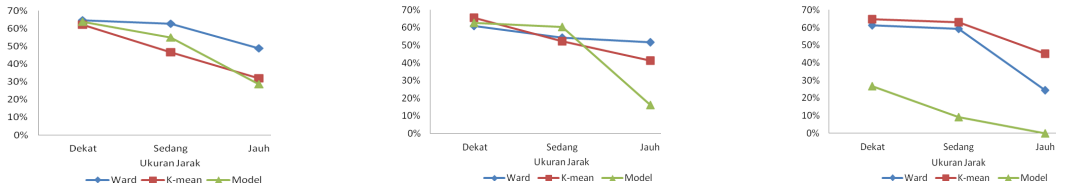
Pada kondisi ketiga kelompok saling tumpang tindih, perbedaan banyak objek amatan tiap kelompok tidak berpengaruh terhadap persentase salah pengelompokannya, baik pada tingkat korelasi maupun pada jarak antar pusat kelompok (Gambar 11).



(a) (b) (c)

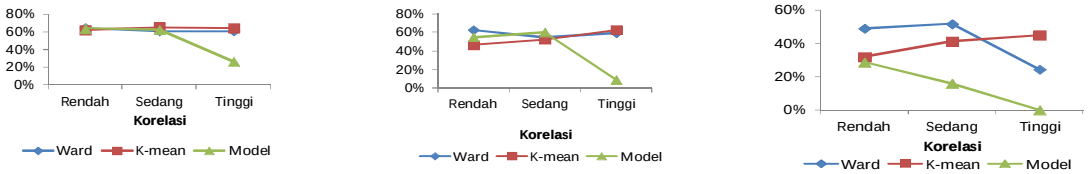
Gambar 11. Plot persentase salah pengelompokan didasarkan pada banyaknya data dengan tingkat korelasi (a) rendah, (b) sedang, dan (c) tinggi dan jarak dekat.

Ditinjau dari jarak antar pusat kelompok, terjadi penurunan persentase salah pengelompokan dengan semakin jauh jarak antar pusat kelompok untuk ketiga metode kelompok baik baik pada tingkat korelasi maupun pada banyak objek amatan tiap kelompok. Hal ini dapat dilihat berdasarkan persentase salah pengelompokan yang dihasilkan, yang disajikan pada Gambar 12.



(a) (b) (c)

Gambar 12. Persentase salah pengelompokan didasarkan pada ukuran jarak dengan tingkat korelasi (a) rendah, (b) sedang, dan (c) tinggi dengan *n*=100.



(a) (b) (c)

Gambar 13. Plot persentase salah pengelompokan yang didasarkan pada tingkat korelasi dengan jarak (a) dekat, (b) sedang, dan (c) jauh dan *n*=100.

Untuk tingkat korelasi rendah, sedang, dan tinggi menunjukkan bahwa pada metode berbasis model terjadi penurunan persentase salah pengelompokan dari tingkat korelasi rendah ke tingkat korelasi tinggi. Hal ini menunjukkan bahwa tingkat korelasi yang berbeda berpengaruh secara signifikan pada kondisi kelompok pada kondisi ketiga kelompok saling tumpang tindih (Gambar 13).

Dari hasil pengelompokan ketiga metode kelompok yang dibandingkan dengan kondisi ketiga kelompok saling tumpang tindih menunjukkan bahwa metode berbasis model lebih efektif memisahkan kelompok yang saling tumpang tindih apabila tingkat korelasi tinggi dan jarak antarpusat kelompok relatif sedang dan jauh.

Tabel 5. Persentase salah pengelompokan pada kondisi ketiga kelompok saling tumpang tindih

Jarak antar pusat kelompok	Metode	n=50			n=100			n=150		
		Tingkat Korelasi			Tingkat Korelasi			Tingkat Korelasi		
		Rendah	Sedang	Tinggi	Rendah	Sedang	Tinggi	Rendah	Sedang	Tinggi
Dekat	Ward	60,00%	62,00%	61,33%	64,67%	61,00%	61,33%	58,89%	63,33%	57,56%
	K-rataan	59,33%	60,67%	62,67%	62,33%	65,67%	64,67%	56,67%	62,22%	63,78%
	Model	64,67%	67,33%	43,33%	64,00%	62,67%	26,67%	58,89%	62,89%	58,44%
Sedang	Ward	51,33%	46,00%	48,67%	62,67%	54,33%	59,33%	42,22%	57,33%	60,00%
	K-rataan	45,33%	52,67%	65,33%	46,67%	52,33%	63,00%	47,33%	51,33%	64,67%
	Model	54,00%	64,67%	12,00%	55,00%	60,33%	9,00%	63,11%	23,78%	8,00%
Jauh	Ward	40,00%	41,33%	42,00%	49,00%	51,67%	24,33%	30,00%	49,56%	31,56%
	K-rataan	48,67%	41,33%	44,67%	32,00%	41,33%	45,00%	31,56%	45,33%	45,78%
	Model	24,00%	22,67%	3,33%	28,67%	16,00%	0,00%	28,89%	14,44%	1,78%

Sebaliknya, apabila tingkat korelasi tinggi dengan jarak antarpusat kelompok relatif dekat dan juga pada tingkat korelasi rendah dan sedang dengan jarak antar pusat kelompok dekat, sedang dan jauh, ketiga metode yang dibandingkan tidak efektif dalam memisahkan kelompok yang tumpang tindih. Persentase salah pengelompokan pada kondisi ketiga kelompok saling tumpang tindih secara lengkap disajikan pada Tabel 5.

SIMPULAN

Berdasarkan penelitian ini, dihasilkan beberapa kesimpulan. sebagai berikut :

Ukuran data pada tiap kelompok tidak berpengaruh terhadap hasil persentase salah pengelompokan yang dihasilkan. Pada ukuran jarak, semakin jauh jarak antarpusat kelompok dengan variansi yang tetap maka persentase salah pengelompokan yang dihasilkan semakin kecil.Pada metode berbasis model, semakin besar tingkat korelasi antarvariabel maka persentase salah pengelompokan yang dihasilkan semakin kecil, sedangkan pada metode Ward dan K-rataan, persentase salah pengelompokan yang dihasilkan sangat bervariasi.

Untuk kondisi ketiga kelompok saling terpisah, ketiga metode yang dibandingkan memberikan hasil pengelompokan yang sama dan sesuai dengan hasil pengelompokan sebenarnya.

Untuk kondisi satu kelompok terpisah dan dua kelompok tumpang tindih, metode berbasis model memberikan hasil yang lebih baik dibandingkan dengan metode Ward dan metode K-rataan.

Untuk kondisi ketiga kelompok saling tumpang tindih dengan tingkat korelasi tinggi dan jarak antarpusat kelompok sedang dan jauh, hasil pengelompokan berbasis model lebih baik dibandingkan dengan metode Ward dan metode K-rataan. Sedangkan untuk tingkat korelasi tinggi dengan jarak antarpusat kelompok relatif dekat, maupun tingkat korelasi rendah dan sedang dengan

jarak antar pusat kelompok dekat, sedang dan jauh, ketiga *metode* kelompok tidak cukup efektif memisahkan ketiga kelompok yang saling tumpang tindih.

REFERENSI

Anderberg, M.R. (1973). *Cluster analysis for applications*, New York: Academic Press

Branfield, J. D. & Raftery, A. E. (1993) Model-based gaussian and non-gaussian *cluster*ing. *Biometrics,* 49, 803-821.

Dempster, A. P., Laird, N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm*, J. R. Statistics Society* B, 39, 1-38.

Fraley, C. & Raftery A.E. (1998). How many c*luster*? Which c*luster*ing method? Answer via model-based c*luster* analysis. *The Computer Journal,* 41, 578-588.

Fraley, C. & Raftery, A. E. (1999). MCLUST:Software for model-based clustering analysis. *Journal of Classifications,*16, 297-306.

Fraley, C. & Raftery, A. E. (2002). MCLUST: Software for rvlodel-Based c*luster*ing, density estimation and discriminant analysis. *Technical Report* 415, University of Washington, Department of Statistics.

Fraley C, & Raftery A. E. (2010). Mclust version 3 for R: Normal mixture modeling and model-based c*luster*ing." *Technical Report* 504. University of Washington, Department of Statistics.

Johnson, R. A. & Wichern, D. W. (2007). *Applied multivariate statistical analysis*, (6th Ed). New Jersey: Prentice-Hall.

Mc Lachlan, G.J. & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering.* New York: Marcel Dekker.

Pardede,T. (2008). Perbandingan Metode Berbasis Model (*Model-Based*) dengan Metode Metode K-rataan dalam Analsis Gugus. *Jurnal Sigma, Sains dan Teknologi,* 11(2), 157-166 .