LAPORAN PENELITIAN BIDANG ILMU
KELOMPOK TINGKAT LANJUT

KAJIAN METODE BERBASIS MODEL PADA ANALISIS *CLUSTER* DENGAN PERANGKAT LUNAK MCLUST

Oleh:
Drs. Timbul Pardede, M.Si
Drs. Budi Prasetyo, M.Si

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TERBUKA
2012

LEMBAR PENGESAHAN

1.	a. Judul Penelitian	:	Kajian Metode Berbasis Model pada Analisis *Cluster* dengan Perangkat Lunak Mclust
	b. Bidang Penelitian	:	Keilmuan
	c. Klasifikasi Penelitian	:	Lanjut
2.	Ketua Peneliti		
	a. Nama Lengkap	:	Drs. Timbul Pardede, MSi
	b. NIP	:	19650508 199103 1 004
	c. Golongan/ Pangkat	:	III/c; Penata
	d. Jabatan Akademik Fakultas dan Unit Kerja	:	Lektor, FMIPA
	e. Program Studi	:	Statistika
3.	Anggota Peneliti		
	a. Jumlah Anggota	:	1 orang
	b. Nama Anggota/Unit	:	Drs. Budi Prasetyo, M.Si.
4.	a. Periode Penelitian	:	Maret – November 2012
	b. Lama Penelitian	:	9 bulan
5.	Biaya Penelitian	:	Rp. 29.700.000 (Dua puluh sembilan juta tujuh ratus ribu rupiah)
6.	Sumber Biaya	:	Universitas Terbuka

Jakarta, Februari 2012

Mengetahui,
Dekan Fakultas MIPA-UT

Peneliti,

Dr. Nuraini Soleiman, M.Ed
NIP. 19540730 198601 2 001

Drs. Timbul Pardede, MSi.
NIP. 19650508 199103 1 004

Mengetahui,
Ketua LPPM

Menyetujui,
Kepala Pusat Keilmuan

Dra. Dewi Artati Patmo Putri, M.A, Ph.D
NIP. 19610724 198701 2 001

Dra. Endang Nugraheni, M.Ed., M.Si
NIP. 19570422 198503 2 001

DAFTAR ISI

Halaman

BAB I
PENDAHULUAN

1.1. Latar Belakang

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

Anderberg (1973) mengemukakan, terdapat beberapa metode *cluster* yang dapat dikelompokan berdasarkan proses algoritma yang dilakukan, diantaranya teknik yang berdasarkan ukuran jarak sebagai basis pengelompokannya. Metode *cluster* berbasis ukuran jarak ini terdiri dari metode *cluster* berhirarki dan metode *cluster* tak berhirarki. Metode *cluster* 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 *cluster* tak berhirarki, misalnya metode K-rataan. Metode *cluster* ini memiliki teknik-teknik yang berbeda-beda dalam proses pembentukan kelompok, namun teknik-teknik tersebut hanya memperhatikan ukuran jarak antar objek-objek pengamatan. Metode-metode ini belum mempertimbangkan aspek statistiknya, seperti sebaran datanya.

Dengan memperhatikan sebaran data yang digunakan dalam analisis *cluster*, Mc Lachlan & Basford (1988) memberikan suatu pendekatan lain yaitu analisis *cluster* berbasis model (*model-based*). Metode *cluster* berbasis model merupakan suatu metode yang berbeda dengan metode *cluster* yang didasarkan pada ukuran jarak. Metode ini merupakan suatu algoritma *cluster* dengan menggunakan analisis yang didasarkan pada aspek statistik di dalam memutuskan hasil *cluster*. Fraley & Raftery (1998) mengidentifikasi, terdapat enam model yang digunakan untuk mengelompokan objek pengamatan dengan berbagai macam 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 *cluster* berbasis model dengan metode *cluster* K-rataan. Pendugaan parameter dilakukan dengan metode maksimum *likelihood*. Kesimpulan yang diperoleh adalah metode berbasis model lebih baik dibandingkan metode K-rataan, akan tetapi dalam keadaan bentuk *cluster* tertentu (objek-objek pengamatan saling tumpang tindih) metode berbasis model dengan enam model belum mampu memisahkan objek-objek yang saling tumpang tindih.

Dengan perkembangan teknologi dan semakin banyaknya *software* komputer yang mendukung dalam melakukan analisis *cluster* baik dalam bentuk angka maupun dalam bentuk visual maka semakin bertambah pula model-model pada metode berbasis model. Fraley & Raftery (1998), mengidentifikasi enam model pada metode *cluster* berbasis model yang digunakan untuk mengelompokan objek pengamatan. Satu tahun berikutnya, tahun 1999 Fraley & Raftery telah mengidentifikasi delapan model pada metode *cluster* berbasis model yang digunakan untuk mengelompokan objek pengamatan. Perangkat lunak yang digunakan untuk menganalisis metode berbasis model adalah Mclust dengan *interface* perangkat lunak S-Plus. Bahkan pada tahun 2010, Fraley & Raftery telah mengidentifikasi sepuluh model untuk mengelompokan objek pengamatan dengan menggunakan perangkat lunak Mclust ver 3.4.11 dengan *interface* perangkat lunak R ver 2.14.1.

Berdasarkan paparan diatas, peneliti ingin melakukan pengkajian analisis *cluster* berbasis model dengan sepuluh model yang telah diidentifikasi oleh Fraley & Raftery dengan menggunakan data bangkitan maupun data sekunder sebagai contoh penerapan. Hasil analisis *cluster* berbasis model ini selanjutnya dibandingkan dengan metode yang didasarkan pada jarak antar objek-objek pengamatan, yaitu metode K-rataan dan metode Ward. Dari hasil analisis diharapkan akan diperoleh efektivitas kesepuluh model untuk mengelompokan objek-objek pengamatan.

1.2. Perumusan Masalah

Berdasarkan uraian latar belakang di atas dapat dirumuskan masalah penelitian sebagai berikut:

1. Seberapa jauh efektivitas analisis *cluster* berbasis model dengan sepuluh model ditinjau dari berbagai jenis data simulasi yang dibangkitan berdasarkan jumlah objek pengamatan, kondisi jarak antar pusat *cluster*, dan kondisi tingkat korelasi antarvariabel?
2. Seberapa jauh efektivitas analisis *cluster* berbasis model dengan sepuluh model bila dibandingkan dengan metode K-rataan dan metode Ward pada data simulasi?
3. Seberapa jauh efektivitas analisis *cluster* berbasis model dengan sepuluh model bila dibandingkan dengan metode K-rataan dan metode Ward pada data sekunder sebagai contoh terapan?

1.3. Tujuan Penelitian

Secara umum penelitian ini bertujuan untuk:

1. Mengkaji efektivitas analisis *cluster* berbasis model dengan sepuluh model ditinjau dari berbagai jenis data simulasi yang dibangkitan berdasarkan jumlah objek pengamatan, kondisi jarak antar pusat *cluster*, dan kondisi tingkat korelasi antarvariabel.
2. Membandingkan metode *cluster* berbasis model dengan metode *cluster* berbasis jarak seperti metode *cluster* K-rataan dan metode Ward pada data simulasi.
3. Mengkaji efektivitas analisis *cluster* berbasis model dengan sepuluh model bila dibandingkan dengan metode K-rataan dan metode Ward pada data sekunder sebagai contoh terapan.

1.4. Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut :

1. Bagi pengembangan ilmu pengetahuan, diharapkan dari hasil penelitian ini dapat menerapkan teori-teori, mengembangkan wawasan dan dinamika keilmuan dalam analisis *cluster* khususnya metode *cluster* berbasis model.
2. Bagi pihak-pihak yang ingin melakukan kajian lebih dalam mengenai analisis *cluster* berbasis model, diharapkan penelitian ini dapat menjadi referensi dan landasan bagi penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA

2.1. Analisis *Cluster*

Analisis *cluster* merupakan salah satu analisis statistik multivariat yang bertujuan untuk mengelompokan objek pengamatan kedalam kelompok-kelompok objek pengamatan berdasarkan karakteristik dari Variabel-variabel yang dimiliki. sedemikian sehingga objek-objek yang terletak dalam kelompok yang sama cenderung mepunyai karakteristik relatif lebih homogen berdasarkan kemiripan atau ketidakmiripan karakteristik-karakteristik yang dimiliki (Hair *et al*., 1998).

Ukuran ketidakmiripan antarobjek pengamatan yang digunakan dalam analisis *cluster* adalah jarak antarobjek. Jarak antar dua objek harus didefinisikan sedemikian rupa sehingga semakin pendek jarak antar dua objek, semakin kecil ketakmiripannya, yang berarti semakin besar peluang untuk dikelompokan dalam suatu *cluster*. Sebaliknya semakin besar jarak antar dua objek, semakin besar pula nilai ukuran ketidakmiripannya, yang berarti semakin kecil peluang untuk dikelompokan dalam suatu *cluster*. Nilai ukuran ketidakmiripan yang sering digunakan pada analisis *cluster* adalah jarak Euclid dan jarak Mahalanobis. Jarak Mahalanobis digunakan bila semua variabel saling berkorelasi atau tidak saling ortogonal, sebaliknya jarak Euclid digunakan bila antarvariabel saling bebas atau saling ortogonal (Johnson & Wichern, 2007).

Jarak Euclid antara objek ke-*i* dan objek ke-*j* dengan *p* variabel didefinisikan sebagai berikut :

$$d_{ij} = \left\{ \sum_{k=1}^{p} \left(X_{ik} - X_{jk} \right)^2 \right\}^{1/2}$$

dan jarak Mahalanobis didefinisikan sebagai berikut :

$$d_{ij} = \left\{ \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right)' S^{-1} \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right) \right\}^{1/2}$$

dengan *S* adalah matriks kovariansi.

Menurut Anderberg (1973), analisis *cluster* terdiri dari beberapa metode *cluster*, antara lain metode *cluster* berhirarki dan metode *cluster* tak berhirarki. Metode *cluster* berhirarki digunakan apabila banyak *cluster* yang akan dibentuk belum diketahui sebelumnya dan jumlah objek amatan relatif kecil. Metode *cluster*

berhirarki ini dapat dibedakan menjadi dua metode yaitu metode penggabungan (*agglomerative*) dan metode pembagian (*divisive*). Metode *cluster* tak berhirarki digunakan apabila banyak *cluster* yang akan dibentuk secara apriori sudah diketahui terlebih dahulu dan jumlah objek amatan relatif besar. Salah satu metode *cluster* tak berhirarki adalah Metode K-rataan.

2.2. Metode *Cluster* Berhirarki dengan Penggabungan

Pada metode *cluster* berhirarki dengan penggabungan dianggap bahwa pada awalnya tiap-tiap objek pengamatan diperlakukan sebagai satu *cluster*, sehingga jumlah *cluster* yang ada sama dengan jumlah objek pengamatan. Tahap selanjutnya dengan menghitung jarak antar *cluster* dengan *cluster* lainnya, dilanjutkan dengan menggabungkan berdasarkan jarak antar dua *cluster* terdekat menjadi satu *cluster* baru. Langkah berikutnya jarak antara *cluster* baru dengan *cluster* lainnya dihitung kembali, yang biasanya disebut dengan perbaikan matriks jarak. Prosedur ini diulang terus hingga terbentuk suatu diagram pohon yang hanya terdiri dari satu *cluster* yang beranggotakan semua objek pengamatan. Hasil *cluster* metode berhirarki membentuk diagram pohon (*tree diagram*) atau *dendrogram* yang menggambarkan pengelompokan objek berdasarkan jarak.

Dalam analisis *cluster* berhirarki dengan penggabungan ini dikenal beberapa metode yang digunakan untuk memperbaiki jarak antar *cluster* (Anderberg, 1973), yaitu :

1. Metode Pautan Tunggal (*Single Linkage*)
   Metode ini menggabungkan *cluster* berdasarkan jarak terpendek (minimum) antar*cluster*
2. Metode Pautan Lengkap (*Complete Linkage*)
   Metode ini menggabungkan *cluster* berdasarkan jarak terpanjang (maksimum) antar*cluster*.
3. Metode Pautan Rataan (*Average Linkage*)
   Metode pautan rataan menggabungkan *cluster* dengan cara menghitung jarak antara rata-rata pasangan seluruh anggota *cluster*.
4. Metode Terpusat (*Centroid Method*)
   Metode ini menghitung jarak antara dua *cluster* sebagai jarak antara rataan dari semua objek amatan dalam satu *cluster* dengan *cluster* lain. Pengelompokan dimulai dari pasangan observasi dengan jarak paling mendekati jarak rata-rata.

5. Metode Ward (*Ward's Methods*)

Metode Ward didasarkan pada kriteria jumlah kuadrat antara dua *cluster* untuk seluruh variabel. Metode ini cenderung digunakan untuk mengkombinasikan *cluster-cluster* dengan jumlah kecil.

Secara umum ukuran jarak yang digunakan untuk analisis *cluster* berhirarki dengan penggabungan ini dapat dituliskan sebagai berikut:

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

dengan nilai koefisien $\delta_1, \delta_2, \delta_3$ dan δ_4 sebagai faktor pembobot untuk masing-masing metode dapat dilihat pada Tabel 1.

Tabel 1. Ukuran jarak yang digunakan pada analisis *cluster* berhirarki dengan penggabungan.

Metode	δ_1	δ_2	δ_3	δ_4
Single Linkage	$1/2$	$1/2$	0	$-1/2$
Complete Linkage	$1/2$	$1/2$	0	$1/2$
Average Linkage (unweighted)	$1/2$	$1/2$	0	0
Average Linkage (weighted)	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Centroid Method	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$\frac{n_i n_j}{\left(n_i + n_j\right)^2}$	0
Ward's Methods	$\frac{n_k + n_i}{n_k + n_i + n_j}$	$\frac{n_k + n_j}{n_k + n_i + n_j}$	$\frac{n_k}{n_k + n_i + n_j}$	0

Sumber: (Härdle & Simar, 2007)

2.3. Metode *Cluster* Tak Berhirarki

Metode *cluster* tak berhirarki digunakan bila banyaknya *cluster* yang akan dibentuk sudah diketahui sebelumnya. Diawali dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (misalnya dua atau lebih *cluster*). Setelah jumlah *cluster* ditentukan, maka proses *cluster* dilakukan dengan tanpa mengikuti proses hirarki. Salah satu metode tak berhirarki yang paling sering digunakan adalah metode *cluster* K-rataan. Metode ini merupakan metode *cluster* yang menyekat objek pengamatan ke dalam *k cluster*. Metode ini pada umumnya diaplikasikan pada gugus data yang berukuran relatif besar.

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

1 Mengambil *k* unit data pertama yang digunakan sebagai *k* pusat *cluster* awal.

2 Menggabungkan setiap *(n-k)* data yang merupakan sisa anggota ke pusat *cluster* terdekat, kemudian dihitung masing-masing pusat (rataan) *cluster* baru yang terbentuk dari hasil penggabungan.

3 Langkah selanjutnya, setelah semua data digabungkan pada tahap 2, pusat *cluster* yang terbentuk dijadikan sebuah titik pusat (rataan) *cluster*, berikutnya dilakukan penggabungan kembali dari setiap unit data ke dalam titik pusat terdekat.

Suatu *cluster* yang konvergen diperoleh dengan memperbaiki secara berulang titik pusat *cluster* yang terbentuk pada tahap ke-3 melalui penggabungan semua *n* data ke titik pusat terdekat. *Cluster* yang konvergen ditandai dengan adanya titik pusat yang tetap dan tidak ada lagi perubahan anggota diantara *cluster*

2.4. Metode *Cluster* Berbasis Model

2.4.1. Model Campuran

Pada analisis *cluster* model campuran, diasumsikan bahwa data dibangkitkan dari sebaran peluang campuran dengan setiap subpopulasi mewakili suatu *cluster* 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 pengamatan 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

$\mathrm{f}_k(y|\theta)$: fungsi kepekatan G_k, yaitu subpopulasi ke-*k* dengan vektor parameter θ yang tidak diketahui

τ_k : merupakan proporsi data yang berasal dari subpopulasi ke-*i* dengan $\sum_{j=1}^{G}\tau_i=1$ dan $\tau_i\geq 0$

Dengan asumsi $y=(y_1, y_2, ..., y_n)$ bebas stokastik dan identik, dan fungsi kepekatan $f_k(y_i|\theta_k)$ merupakan fungsi kepekatan campuran objek pengamatan y_i dari *cluster* ke-*k* maka fungsi kepekatan sebaran campuran (*mixture likelihood*) pada persamaan (1) adalah :

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

Dalam penelitian ini difokuskan pada kasus dimana $f_k(y_i|\theta_k)$ adalah fungsi kepekatan variabel ganda campuran normal (*Gauss*) dengan parameter θ_k terdiri dari vektor rataan μ_k dan matriks kovariansi $\sum_k$, yang dinyatakan dalam bentuk :

$$f_k(y_i|\mu_k; \sum_k) = \frac{\exp\left\{-\frac{1}{2}(y_i - \mu_k)' \sum_k^{-1} (y_i - \mu_k)\right\}}{(2\pi)^{\frac{p}{2}} |\sum_k|^{\frac{1}{2}}}$$

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

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

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

Branfield & Raftery (1993) mengembangkan metode *cluster* 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, bentuk) mungkin akan diperoleh dari berbagai macam bentuk *cluster*, atau terbatas pada *cluster* yang sama dan matriks varians untuk semua komponen bisa sama atau bervariasi. Sebagai ilustrasi, model $\sum_k = \lambda I$ menunjukkan bahwa semua *cluster* berbentuk *spherical* dan memiliki volume sama. Model $\sum_k = \sum$ menunjukkan semua *cluster* mempunya ciri geometrik yang sama tetapi tidak harus *spherical* dan $\sum_k$ tidak terbatas. Model $\sum_j = \lambda DAD'$ mempunyai ciri geometrik sama dan semua *cluster* berbentuk *ellipsoidal*. Model $\sum_k = \lambda_k D_k A_k D_k'$ mempunyai model tanpa batasan dimana setiap *cluster* mempunyai ciri geometrik yang berbeda. Tabel-1 menunjukkan matriks kovariansi $\sum_j$ untuk model campuran normal ganda dan interpretasi geometrik (Fraley & Raftery, 2010).

Tabel 2. Matriks kovariasi $\sum_k$ dan interpretasi geometrik pada model campuran normal ganda.

$\sum_j$	**Volume**	**Bentuk Geometri**	**Orientasi**	**Tebaran**	**Simbol Mclust**
λI	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)

2.4.2. Algoritma EM (*Expectation-Maximum*) untuk model campuran

Algoritma EM merupakan metode perhitungan iterasi terhadap masalah pendugaan kemungkinan maksimum parameter pada data tidak lengkap (Dempster, Laird, and Rubin, 1977). Algoritma EM pada *cluster*, data lengkap diasumsikan menjadi $\boldsymbol{x}_i' = (\boldsymbol{y}_i', \boldsymbol{z}_i')$, dengan $\boldsymbol{y}_i'$ merupakan 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:

$$\mathrm{L_C}(x_i|\theta)=\prod_{i=1}^{n} f_j\left(x_i|\theta\right)$$

Selanjutnya jika peluang variabel khusus tidak teramati dan tergantung pada pengamatan data $\boldsymbol{y}$ dan bukan $\boldsymbol{z}$ maka fungsi *likelihood* data lengkap menjadi:

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

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 $\boldsymbol{x}_i'=\left(\boldsymbol{y}_i',\boldsymbol{z}_i'\right)$ dengan $\boldsymbol{z}_i'=\left(z_{i1},z_{i2},....,z_{ig}\right)$ merupakan data yang tidak teramati, yaitu

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

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 :

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

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:

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

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

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

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.

Fraley dan Raftery (1998) membuat algoritma EM pada model campuran Gauss sebagai berikut:

Mulai

Tahapan E

Hitung $$\hat{z}_{ik} = \frac{\hat{\tau}_k \, f_k\left(y_i \middle| \hat{\mu}_k, \hat{\Sigma}_k\right)}{\sum_{j=1}^{G} \hat{\tau}_j \, f_j\left(y_i \middle| \hat{\mu}_j, \hat{\Sigma}_j\right)}$$

atau $$\hat{z}_{ik} = \frac{\hat{\tau}_k \left|\hat{\Sigma}_j\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(y_i - \hat{\mu}_k\right)' \hat{\Sigma}_k^{-1}\left(y_i - \hat{\mu}_k\right)\right\}}{\sum_{j=1}^{G} \hat{\tau}_j \left|\hat{\Sigma}_j\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(x_i - \hat{\mu}_j\right)' \hat{\Sigma}_j^{-1}\left(x_i - \hat{\mu}_j\right)\right\}}$$

dengan f_k diperoleh dari persamaan (3)

Tahapan M

Maksimumkan $\hat{z}_{ik}$ dari persamaan (8)

$$n_k \longleftarrow \sum_{i=1}^{n} z_{ik}$$

$$\hat{\tau}_k \longleftarrow \frac{n_k}{n}$$

$$\hat{\mu}_k \longleftarrow \frac{\sum_{i=1}^{n} \hat{z}_{ik} \, y_i}{n_k}$$

$\hat{\Sigma}_k$: sesuai dengan model pada Tabel 1.

Ulang Sampai kriteria konvergen dipenuhi.

2.4.3. Pemilihan Model *Cluster* dengan Faktor Bayes

Pada analisis *cluster* masalah yang paling sering muncul adalah bagaimana menentukan metode *cluster* yang digunakan dan berapa jumlah *cluster* yang ada. Seringkali para pengguna statistik melakukan coba-coba (*trial and error*) untuk mendapatkan hasil yang bermakna atau yang dapat diinterpretasikan sesuai dengan

masalah kajiannya, sehingga hampir semua metode digunakan dan kemudian hasilnya dibandingkan. Solusi untuk menangani kedua masalah ini, Fraley & Raftery (1998) melakukan pendekatan model campuran melalui faktor Bayes. Salah satu keuntungan pendekatan model campuran dengan menggunakan pendekatan faktor Bayes adalah dapat membandingkan antarmodel. Sistematika pemilihan tidak hanya untuk parameterisasi model (metode *cluster* yang digunakan), tetapi juga banyaknya *cluster*. Pendekatan yang digunakan adalah dengan pendekatan BIC (*Bayesian Information Criterion*) dengan formulasi sebagai berikut:

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

dimana

$P\left(y|M_k\right)$: 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 yang paling layak. Fraley & Raftery (1998) menyimpulkan suatu konvensi standar untuk kalibrasi perbedaan nilai BIC antar*cluster*, yakni bila perbedaan nilai BIC kurang dari 2 maka tingkat pemisahan *cluster* cukup lemah, perbedaan nilai BIC antara 2 sampai dengan 6 maka tingkat pemisahan *cluster* cukup, perbedaan nilai BIC antara 6 sampai dengan 10 maka tingkat pemisahan *cluster* cukup kuat, dan bila perbedaan nilai BIC lebih dari 10 maka tingkat pemisahan *cluster* sangat kuat.

2.4.4. Strategi Metode *Cluster* Berbasis Model

Fraley & Raftery (1998) membuat strategi metode *cluster* berbasis model dengan cara mengkombinasikan *cluster* berhirarki penggabungan, algoritma EM, dan faktor Bayes dengan langkah-langkah sebagai berikut :

*. Tentukan banyak *cluster* maksimum (*M*), dan himpunan model campuran ganda normal.

*. Lakukan pengelompokan dengan berhirarki penggabungan untuk setiap model campuran normal ganda. Hasil pengelompokan ini

ditransformasi ke dalam variabel indikator, kemudian digunakan sebagai nilai awal untuk algoritma EM

*. Lakukan algoritma EM untuk setiap model dan masing-masing banyak *cluster* 2, 3, ….,*M*, yang diawali dengan klasifikasi *cluster* berhirarki.

*. Hitung nilai BIC untuk kasus satu *cluster* pada setiap model dan untuk model *likelihood* campuran dengan parameter optimal dari algoritma EM untuk 2, 3, …., m *cluster*.

*. Plotkan nilai BIC untuk setiap model.

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

III. METODE PENELITIAN

3.1. Tempat dan Waktu Penelitian

Penelitian dilaksanakan di UT Pusat selama 9 bulan, mulai Maret 2012 sampai dengan November 2012.

3.2. Sumber data

Sumber data yang digunakan dalam penelitian ini adalah data himpunan campuran normal ganda hasil simulasi yang dibangkitkan dengan menggunakan fungsi *mvnorm pada* perangkat lunak program R ver 2.14.1 dan data sekunder yang diperoleh dari paket Mclust ver 3.4.11. Kriteria data simulasi yang dibangkitkan mengacu pada Pardede (2008), yakni terdiri dari tiga macam jumlah amatan yang dibangkitkan dari sebaran normal ganda (Gaussian), yaitu 50, 100, dan 150 jumlah amatan dengan masing-masing terdiri dari tiga variabel dan tiga *cluster*. Pemilihan jumlah *cluster* dan jumlah variabel ini dikaitkan dengan penggunaan di lapangan, yaitu mengacu pada contoh kasus data *Iris*. Contoh penerapan data *Iris* ini sering digunakan sebagai contoh penerapan dalam analisis *cluster*.

Ketiga *cluster* yang akan dibangkitkan dibuat dalam 3 macam kondisi, yaitu (1) ketiga *cluster* saling terpisah, (2) satu *cluster* terpisah dan dua *cluster* tumpang tindih, dan (3) ketiga *cluster* saling tumpang tindih. Untuk membangkitkan ketiga kondisi tersebut, maka digunakan 3 jenis ukuran jarak antara dua nilai tengah (pusat) *cluster*, yang disesuaikan dengan jauh dekatnya jarak antara vektor rataan *cluster*. Hal ini didasarkan pada pemikiran bahwa semakin dekat jarak antara kedua pusat *cluster*, semakin banyak pengamatan yang tumpang tindih. Sebaliknya semakin jauh jarak antara kedua pusat *cluster*, semakin sedikit pengamatan yang tumpang tindih (Pardede, 2008). Di samping itu, untuk melihat pengaruh tingkat korelasi antarvariabel terhadap hasil akhir pengelompokan, maka dicobakan juga 3 jenis tingkat korelasi, yaitu korelasi rendah (0,25), korelasi sedang (0,5). dan korelasi tinggi (0,8). Pemilihan tingkat korelasi ini didasarkan pada kesimetrisan tingkat korelasi yang bernilai pada $|r| \leq 1$.

Dengan jumlah amatan yang beragam ini, diharapkan dapat diketahui efektivitas analisis *cluster* berbasis model pada jumlah amatan yang berbeda-beda. Pola-pola data simulasi yang akan dibangkitkan secara lengkap dapat dilihat pada Tabel 3.

Tabel 3. Pola data simulasi yang akan dibangkitkan

Jenis pengelompokan	Jarak antar dua pusat *cluster* dan nilai variansi tiap variabel	Tingkat korelasi antar variabel	Banyak data tiap *cluster*
Ketiga *cluster* saling terpisah	Dekat, Sedang, Jauh $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$ (variansi kecil)	Rendah (0.25) Sedang (0.5) Tinggi (0.75)	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 (0.25) Sedang (0.5) Tinggi (0.75)	50 100 150
Ketiga *cluster* saling tumpang tindih	Dekat, Sedang, Jauh $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$ (variansi besar)	Rendah (0.25) Sedang (0.5) Tinggi (0.75)	50 100 150

Sumber: (Pardede, 2008)

Untuk mendukung hasil penelitian ini, digunakan data terapan yakni data *Iris*. Data *Iris* merupakan data sekunder yang diperoleh dari paket perangkat lunak R dan Mclust. Penggunaan data sekunder ini untuk melihat efektivitas analisis *cluster* pada salah satu kondisi data simulasi.

Dengan demikian, jumlah kasus simulasi yang akan dianalisis sebanyak 81 kasus dan contoh penerapan ada sebanyak satu data sekunder. Setiap data simulasi dan data sekunder dilakukan analisis dengan menggunakan metode Ward, metode K-rataan, dan metode berbasis model. Metode Ward dipilih, karena metode ini umumnya digunakan pada kumpulan data yang relatif kecil dan merupakan salah satu metode *cluster* berhirarki. Sedangkan metode K-rataan dipilih, karena metode ini umumnya diaplikasikan pada kumpulan data yang relatif besar dan merupakan salah satu metode *cluster* tak berhirarki.

3.3. Bangkitan Data Simulasi

Data yang dibangkitkan merupakan data himpunan campuran normal ganda berdimensi 3. Data simulasi ini memerlukan dua faktor, yaitu vektor-vektor rataan setiap *cluster* (μ_1, μ_2 dan μ_3) dan matriks kovariansi ($\sum$) setiap *cluster*. Vektor rataan setiap *cluster* menggambarkan ukuran pemusatan setiap *cluster* dan matriks kovariansi menggambarkan ukuran sebaran data disekitar vektor rataanya.

Tahapan yang dilakukan untuk membangkitan data himpunan sebaran campuran normal pada setiap kasus adalah sebagai berikut :

1. Menentukan banyak *cluster* (*G*=3 *cluster*), banyak variabel (*p*=3) dan banyak data tiap *cluster* (*n*=50, 100, 150) dengan sebaran setiap data bangkitan adalah $G_k \sim MNV_3(\mu_k, \sum_k)$
2. Menentukan sebaran data untuk masing-masing *cluster* berdasarkan parameter vektor rataan $(\mu_1 \quad \mu_2 \quad \mu_3)$ dan matriks kovariansi $(\sum_1 \quad \sum_2 \quad \sum_3)$. Untuk membangkitkan matriks kovariansi tersebut dilakukan dengan cara:
   a. Menentukan matriks $S_k^{1/2}$ yang merupakan matriks diagonal berdimensi 3x3 dengan masing-masing elemen diagonalnya adalah standar deviasi masing-masing variabel.
   b. Menentukan matriks korelasi antar variabel, yaitu $R_k; k = 1,2,3$
   c. Menghitung matriks kovariansi $\sum_k = S_k^{1/2} R_k S_k^{1/2}$
3. Membangkitkan data variabel acak multivariat normal berdimensi tiga untuk *cluster* ke-*k* sebanyak n_k , yaitu $G_k \sim MNV_3(\mu_k, \sum_k); k = 1,2,3$
4. Menggabungkan ketiga jenis *cluster* menjadi sebuah kasus simulasi.
5. Ulangi tahap 2-6 untuk 80 kasus simulasi lainnya.

3.4. Prosedur Analisis Data

Sebelum melakukan analisis *cluster* terlebih dahulu dilakukan analisis plot skor komponen utama pada setiap kasus, dengan tujuan untuk melihat secara visual sebaran data dan banyaknya *cluster* yang terbentuk dari hasil simulasi. Selanjutnya, dari hasil data bangkitan dilakukan tahapan sebagai berikut:

1. Lakukan analisis *cluster* dengan menggunakan perangkat lunak R ver 2.14.1 untuk metode Ward.
2. Lakukan analisis *cluster* dengan menggunakan perangkat lunak R ver 2.14.1 untuk metode K-rataan.
3. Lakukan analisis *cluster* dengan menggunakan paket program Mclust ver 3.14.11 dengan *interface* R ver 2.14.1 pada metode *cluster* berbasis model, dengan tahapan sebagai berikut:
    a. Tentukan banyak *cluster* maksimum (*M*), dan himpunan model campuran ganda normal.
    b. Lakukan pengelompokan dengan berhirarki penggabungan untuk setiap model campuran normal ganda.
    c. Hasil pengelompokan ini ditransformasi ke dalam variabel indikator, yang kemudian digunakan sebagai nilai awal untuk algoritma EM
    d. Lakukan algoritma EM untuk setiap model dan masing-masing banyak *cluster* 2, 3, ….,*M*.
    e. Hitung nilai BIC untuk kasus satu *cluster* pada setiap model dan untuk model *likelihood* campuran dengan parameter optimal dari algoritma EM untuk 2, 3, …., m *cluster*.
    f. Plotkan nilai BIC untuk setiap model.
    g. Nilai BIC terbesar mengindikasikan bahwa model tersebut adalah model yang paling layak.
4. Lakukan kajian tentang hasil pengelompokan masing-masing metode dengan pengelompokan yang sebenarnya (ditentukan saat simulasi).
5. Hitung rataan persentase salah pengelompokan dari setiap *cluster* pada masing-masing metode, kemudian hasilnya dibandingkan.
6. Rataan persentase salah pengelompokan yang terkecil menunjukkan bahwa metode yang digunakan lebih baik.
7. Lakukan langkah 1-4 untuk 81 pola data simulasi dan untuk data sekunder (data pohon dan data *Iris*).

IV. HASIL DAN PEMBAHASAN

4.1. Data Simulasi

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

Sebagai dasar untuk melihat kondisi pengelompokan dari hasil simulasi, maka kondisi pengelompokan yang dibentuk sebaiknya terdiri dari tiga kondisi *cluster*, yaitu 1) kondisi ketiga *cluster* saling terpisah dengan banyak objek pengamatan tiap *cluster* sebesar 50, 100, dan 150 amatan, 2) kondisi satu *cluster* terpisah dan dua *cluster* saling tumpang tindih dengan banyak objek pengamatan tiap *cluster* sebesar 50, 100, dan 150 amatan, 3) kondisi ketika *cluster* saling tumpang tindih dengan banyak objek pengamatan tiap *cluster* sebesar 50, 100, dan 150 amatan. Untuk melihat kondisi pengelompokan dari hasil bangkitan data, maka secara visual data hasil simulasi disajikan plot skor dua komponen utama yang secara lengkap dapat dilihat pada Lampiran 1, Lampiran 2, dan Lampiran 3.

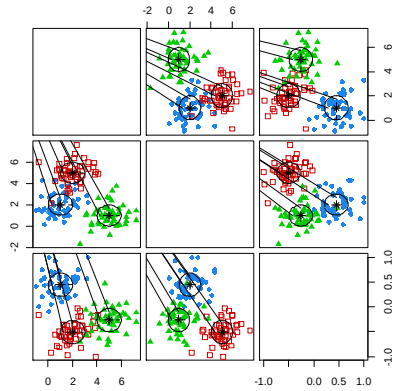
Setiap kasus data simulasi ini akan digunakan sebagai data awal untuk menganalisis efektivitas analisis *cluster* dengan 1)metode berbasis model dengan sepuluh model yang dicobakan, 2)metode K-rataan, dan 3)metode Ward. Pada metode berbasis model dipilih model terbaik yang didasarkan atas nilai BIC terbesar. Selanjutnya hasil analisis *cluster* dari masing-masing metode dibandingkan berdasarkan rataan persentase salah pengelompokannya. Metode terbaik didasarkan pada rataan persentase salah pengelompokan yang terkecil. Semakin kecil rataan persentase kesalahan pengelompokan yang dihasilkan maka metode tersebut semakin efektif dalam mengelompokan objek-objek pengamatan.

4.2. Kondisi Ketiga *Cluster* Saling Terpisah

Pada kondisi pengelompokan dengan ketiga *cluster* saling terpisah terdapat 27 kasus pola simulasi data yang dibedakan atas jarak antar pusat *cluster* (dekat, sedang, dan jauh) dengan variansi dari ketiga *cluster* adalah $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$; tingkat korelasi antar variabel adalah rendah (0.2), sedang (0.5), dan tinggi (0.8); dan banyak objek pengamatan pada tiap *cluster* adalah n=50, n=100, dan n=150.

Sebagai ilustrasi pertama diambil kasus data simulasi dengan kondisi pengelompokan ketiga *cluster* saling terpisah, jarak antar pusat *cluster* dekat

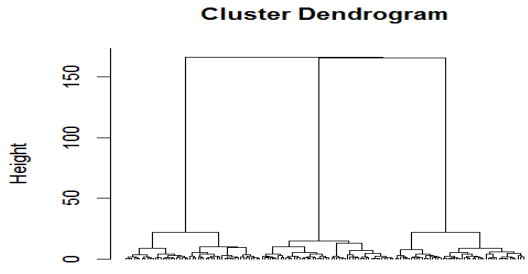
(d=5.099) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$, tingkat korelasi antarvariabel adalah rendah (0.2), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=50.



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

Secara visual matriks plot (Gambar 1) dari kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* dekat, variansi cenderung kecil, tingkat korelasi rendah, dan banyak data sebesar *n*=50 menunjukkan bahwa ketiga *cluster* saling terpisah.

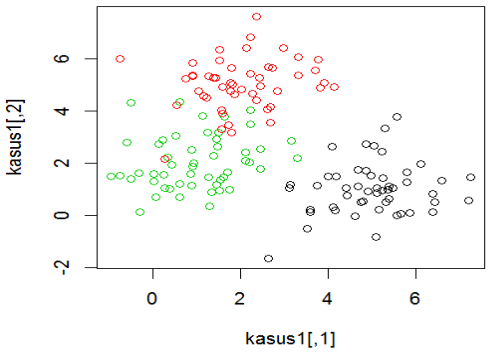
Hasil pengelompokan yang diperoleh dengan metode Ward menunjukkan bahwa metode Ward dapat mengelompokan objek-objek pengamatan secara tepat dengan pengelompokan yang sebenarnya yang ditentukan pada saat simulasi (Gambar 2). Hal ini mengindikasikan bahwa rataan persentasi salah pengelompokannya adalah 0%.




```
[1]   1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[51]  2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[101] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Gambar 2. Dendogram dan hasil pengelompokan metode Ward dengan kondisi data simulasi untuk kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* dekat, variansi cenderung kecil, tingkat korelasi rendah, dan banyak data *n*=50.

Pada metode K-rataan, analisis dilakukan sampai 100 kali iterasi hingga diperoleh *cluster* yang konvergen. Hasil pengelompokan menunjukkan bahwa metode K-rataan juga sesuai dan tepat untuk mengelompokan objek-objek pengamatan dengan pengelompokan yang sebenarnya, seperti terlihat pada Gambar 3. Hal ini juga menggambarkan bahwa rataan persentasi salah pengelompokan dengan metode K-mean adalah 0%.



Cluster means:

	[,1]	[,2]	[,3]
1	5.000000	1.000000	2.0000000
2	2.017460	5.024688	0.9480177
3	1.002833	2.035104	4.9715124

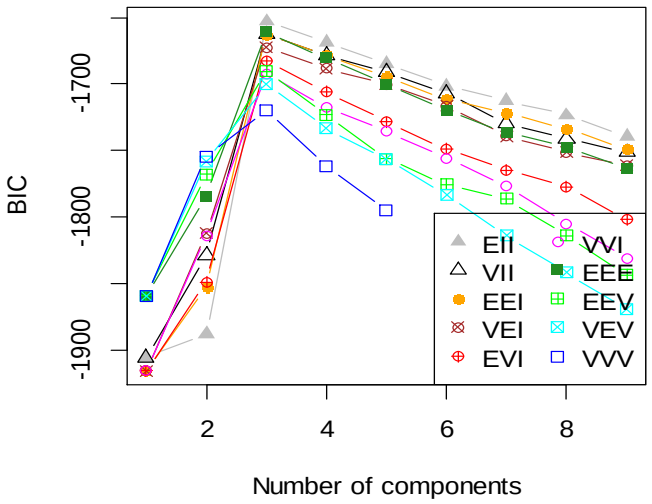
Clustering vector:

```
 [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[51] 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[101] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Gambar 3. Plot dan hasil pengelompokan metode K-mean dengan kondisi data simulasi untuk kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* dekat dan variansi cenderung kecil, tingkat korelasi rendah, dan banyak data sebesar n=50.

Pada metode berbasis model, 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 nilai BIC yang paling besar yaitu pada model EII (Gambar

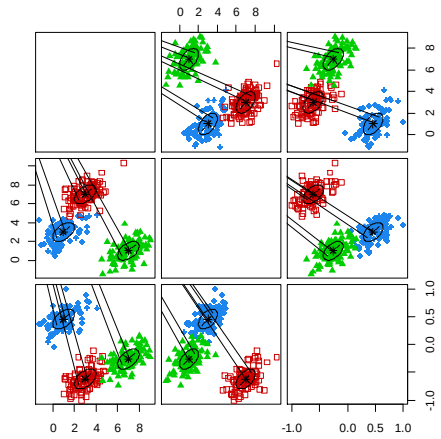


4).

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EII (spherical, equal volume) model with 3 components:

 log.likelihood   n df      BIC
    -796.3208 150 12 -1652.769

Clustering table:
 1  2  3
50 50 50

best BIC values:
   EII,3    EEE,3    VII,3
-1652.769 -1660.790 -1662.661
```

Gambar 4. Plot dan hasil *cluster* metode berbasis model dengan model terbaik adalah EII untuk kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* dekat dan variansi cenderung kecil, tingkat korelasi rendah, dan banyak data sebesar n=50.

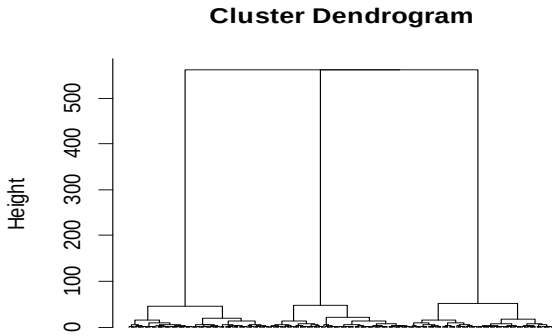
Sebagai ilustrasi kedua diambil kasus data simulasi dengan kondisi pengelompokan ketiga *cluster* saling terpisah, jarak antarpusat *cluster* sedang (d=7.483) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$, tingkat korelasi antarvariabel adalah sedang (0.5), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=100.



Gambar 5. Matriks plot data simulasi untuk kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* dekat dan variansi cenderung kecil dan tingkat korelasi sedang dan banyak data sebesar *n*=100

Secara visual matriks plot (Gambar 5) dari kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* dekat, variansi cenderung kecil, tingkat korelasi rendah, dan banyak data sebesar *n*=50 menunjukkan bahwa ketiga *cluster* saling terpisah.

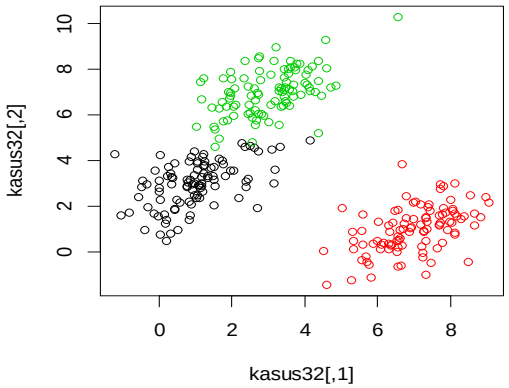
Berdasarkan pengelompokan dengan metode Ward diperoleh hasil bahwa metode Ward sesuai dan tepat mengelompokan objek-objek pengamatan dengan pengelompokan yang sebenarnya, seperti terlihat pada dendogram di Gambar 6. Hal ini menunjukkan rataan persentasi salah pengelompokannya adalah 0%.




```
[1]   1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [51] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[101] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[151] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[201] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[251] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

Gambar 6. Dendogram dan hasil pengelompokan metode Ward dengan kondisi data simulasi untuk kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* sedang dan variansi cenderung kecil, tingkat korelasi sedang, dan banyak data sebesar *n*=100.

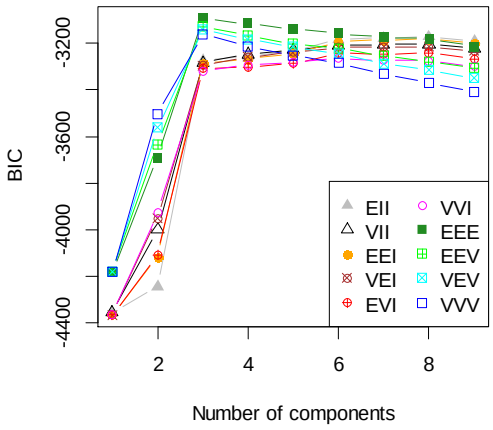
Pada metode K-rataan, analisis dilakukan sampai 100 kali iterasi hingga diperoleh *cluster* yang konvergen. Berdasarkan pengelompokan diperoleh hasil bahwa metode K-rataan juga sesuai dan tepat dapat pengelompokan objek-objek pengamatan dengan pengelompokan yang sebenarnya (ditentukan saat simulasi), seperti terlihat pada Gambar 7. Hal ini juga menunjukkan rataan persentasi salah pengelompokannya adalah 0%.



K-means clustering with 3 clusters of sizes 100, 100, 100
Cluster means:
[,1] [,2] [,3]
1 1 3 7
2 7 1 3
3 3 7 1
Clustering vector:
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[51] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[101] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[151] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[201] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[251] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Gambar 10. Plot dan hasil pengelompokan metode K-mean dengan kondisi data simulasi untuk kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* sedang dan variansi cenderung kecil, tingkat korelasi rendah, dan banyak data sebesar n=100.

Pada metode berbasis model, dari 10 model yang dicobakan terdapat tiga model yang paling layak yang didasarkan pada nilai BIC paling besar, yakni model EEE (3 *cluster*) dengan nilai BIC = -3093.232; model EEE (4 *cluster*) dengan nilai BIC = -3116.187; dan model EEV dengan nilai BIC = -3127.454. Nilai BIC yang paling besar dari tiga model yang paling layak terdapat pada model EEE dengan nilai BIC = -3093.232 maka model terbaik terdapat pada nilai BIC yang paling besar yaitu pada model EII (Gambar 8).




```
-------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
-------------------------------------------------
Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 3 components
 log.likelihood   n df      BIC
    -1498.134 300 17 -3093.232
Clustering table:
  1   2   3
100 100 100
best BIC values:
   EEE,3    EEE,4    EEV,3
-3093.232 -3116.187 -3127.454
```

Gambar 8 Plot dan hasil *cluster* metode berbasis model dengan model terbaik adalah EEE untuk kondisi ketiga *cluster* saling terpisah dengan jarak pusat *cluster* dan variansi cenderung kecil dan tingkat korelasi rendah dan banyak data sebesar n=100.

Untuk 27 kasus yang kondisi ketiga *cluster* saling terpisah diperoleh hasil rataan persentase salah pengelompokan yang sama besar baik pada metode Ward, metode K-rataan, dan metode berbasis model, yaitu sebesar 0%. Hal ini menunjukkan bahwa ketiga metode dapat mengelompokan objek-objek pengamatan secara sempurna dan sesuai dengan pengelompokan yang sebenarnya (ditentukan saat simulasi). Hal ini disebabkan oleh variansi dari masing-masing *cluster* cenderung kecil ($\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1$) sehingga setiap objek-objek pengamatan cenderung mengelompok di sekitar vektor rataan *cluster*.

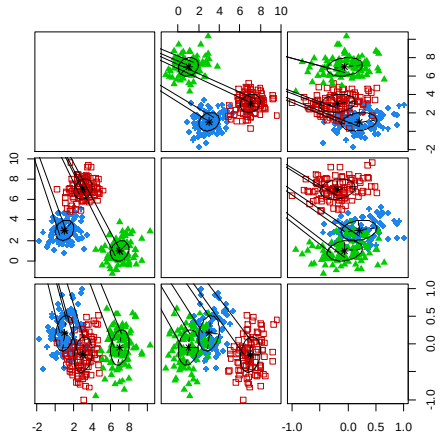
Untuk metode berbasis model, dari 27 kasus simulasi pada kondisi ketiga *cluster* saling terpisah, model terbaik terdapat pada model EEE yang tebaran datanya berberntuk *ellipsoidal*, kecuali pada kondisi banyak data n=50 dengan jarak antar pusat *cluster* dekat, sendang, jauh pada tingkat korelasi rendah (0,2) menghasilkan model terbaik EII yang berbentuk *Spherical*, dan pada tingkat korelasi sedang (0,5) menghasilkan model terbaik EEV yang tebaran datanya berbentuk *ellipsoidal* (Lampiran 4).

Rataan persentase salah pengelompokan tidak terpengaruh terhadap jarak antar pusat *cluster* (jarak dekat, jarak sedang, dan jarak jauh). Demikian juga tingkat korelasi antarvariabel (rendah (0.2), sedang (0.5), dan tinggi (0.8)); dan banyak objek pengamatan pada tiap *cluster* (*n*=50, *n*=100, dan *n*=150) tidak berpengaruh pada rataan persentase salah pengelompokan antar *cluster*.

4.3. Satu *Cluster* Terpisah dan Dua *Cluster* Tumpang Tindih

Pada kondisi pengelompokan dengan satu *cluster* terpisah dan dua *cluster* tumpang tindih terdapat 27 kasus simulasi data yang dibedakan atas jarak antarpusat *cluster* (dekat, sedang, dan jauh) dengan variansi dari ketiga *cluster* adalah $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$; tingkat korelasi antarvariabel adalah rendah (0.2), sedang (0.5), dan tinggi (0.8); dan banyak objek pengamatan pada tiap *cluster* adalah *n*=50, *n*=100, dan *n*=150.

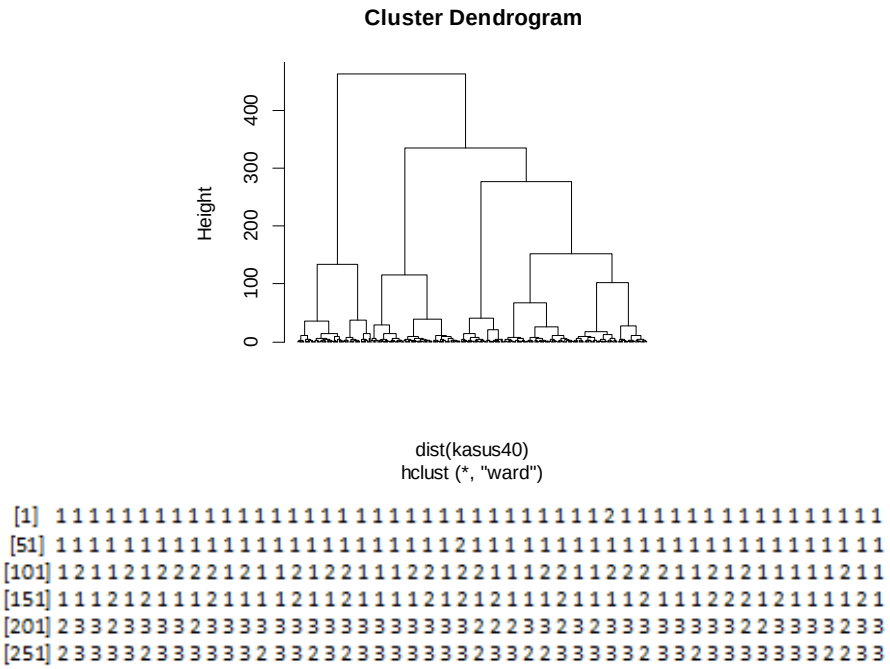
Sebagai ilustrasi diambil kasus data simulasi dengan satu *cluster* terpisah dan dua *cluster* tumpang tindih, jarak antarpusat *cluster* sedang (d=7.483) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$, tingkat korelasi antarvariabel rendah (0.2), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=100.



Gambar 9. Matriks plot data simulasi untuk kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih dengan jarak pusat *cluster* sedang dengan variansi $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$, tingkat korelasi sedang, dan banyak data *n*=100.

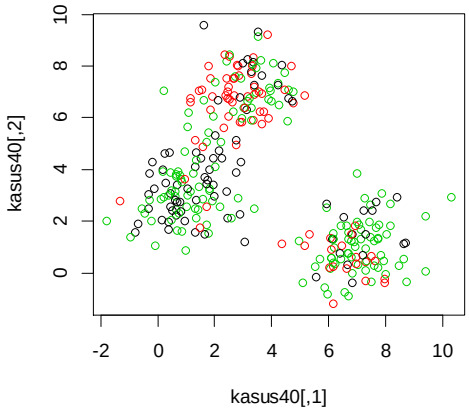
Hasil pengelompokan dengan metode Ward diperoleh hasil bahwa dari 100 objek amatan pada *cluster* 1 terdapat 2 objek amatan masuk ke dalam *cluster* 2, dari 100 objek amatan pada *cluster* 2 terdapat 62 objek amatan masuk ke dalam *cluster* 1,

dan dari 100 objek amatan pada *cluster* 3 terdapat 23 objek amatan masuk ke *cluster* 2 (Gambar 10). Rataan persentasi salah pengelompokannya adalah 29,00%.



Gambar 10. Dendogram dan hasil pengelompokan metode Ward dengan kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih dengan jarak pusat *cluster* sedang dengan variansi $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$ dan tingkat korelasi sedang dan banyak data sebesar n=100.

Pada metode K-rataan, hasil pengelompokan yang diperoleh adalah pada *cluster* 1 terdapat 20 masuk ke dalam *cluster* 2 dan 23 masuk pada *cluster* 3 dan hanya 57 objek amatan yang dengan tepat masuk ke dalam *cluster* 1. Untuk *cluster* 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 *cluster* 3. Hal ini menunjukkan bahwa rataan lebih dari 50% objek amatan tidak terkelompok pada tempatnya (Gambar 11).



Cluster means:

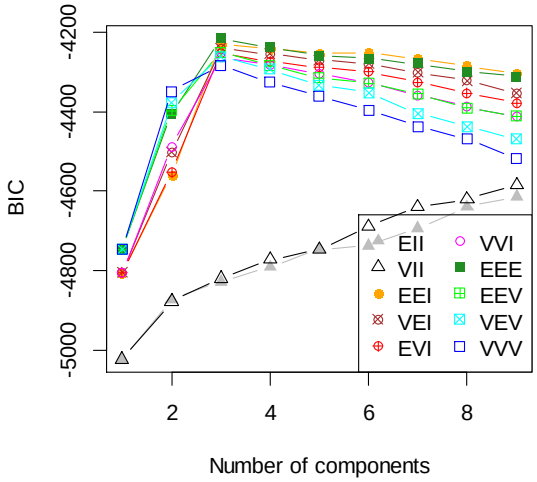
	[,1]	[,2]	[,3]
1	2.766106	3.686675	10.677701
2	3.747886	4.936593	-3.348148
3	4.107311	3.029604	3.384300

Clustering vector:

```
 [1] 3 1 1 3 1 1 1 1 3 3 3 3 1 1 1 3 3 2 3 3 1 1 1 3 1 3 3 3 1 3 1 1 3 2 3 3 1 3 1 1 3 1 1 3 3 3 1 3 1 1
[51] 3 1 1 1 1 1 1 3 3 1 1 1 1 1 1 1 1 3 1 3 3 3 3 1 3 2 3 3 3 3 3 1 3 3 3 3 3 3 1 3 1 1 3 1 1 1 1 3 3 1 2
[101] 3 2 2 3 2 3 2 2 2 2 3 2 3 3 2 2 2 2 3 3 3 2 2 3 2 2 2 3 1 2 2 1 3 2 2 2 2 3 2 2 1 2 3 1 3 3 2 3 1
[151] 1 2 3 2 1 2 3 1 3 2 1 3 3 2 2 3 3 2 3 2 3 3 2 3 2 1 2 2 1 2 2 1 2 1 3 2 3 3 3 2 2 2 3 2 3 2 1 1 2 1
[201] 2 3 3 2 3 3 3 3 2 3 3 3 3 3 3 3 3 3 1 3 1 3 3 3 2 2 2 3 1 3 3 2 3 1 3 3 3 3 3 1 3 2 3 1 3 3 2 3 1
[251] 2 3 3 1 3 2 3 3 1 3 3 3 2 1 3 3 3 2 1 3 3 2 1 3 1 3 3 3 2 1 3 2 2 3 3 3 3 3 2 3 3 2 3 3 3 3 1 1 2 2 1 3
```

Gambar 11. Plot dan hasil pengelompokan metode K mean dengan kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih dengan jarak pusat *cluster* sedang dengan variansi $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$ dan tingkat korelasi sedang dan banyak data sebesar n=100.

Pada metode berbasis model, dari 10 model yang dicobakan terdapat tiga model yang paling layak, yakni model EEE (3 *cluster*) dengan nilai BIC = -4216.587; model EEI (3 *cluster*) dengan nilai BIC = -4230.491; dan model EEE (4 *cluster*) dengan nilai BIC = -4238.835. Model terbaik dari tiga model yang paling layak terdapat nilai BIC yang paling besar yaitu pada model EEE (Gambar 12).




```
Classification:
 [1] 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1
 [51] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[101] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2
[151] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[201] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[251] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 3 components:

 log.likelihood   n df      BIC
      -2059.811 300 17 -4216.587

best BIC values:
   EEE,3    EEI,3    EEE,4
-4216.587 -4230.491 -4238.835
```

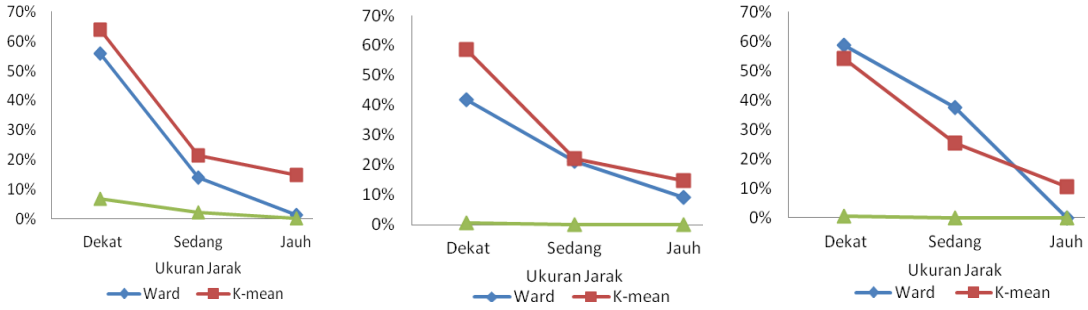
Gambar 12. Plot dan hasil pengelompokan metode berbasis model dengan kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih dengan jarak pusat *cluster* sedang dengan variansi $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 25$ dan tingkat korelasi sedang dan banyak data sebesar n=100.

Untuk metode berbasis model, dari 27 kasus simulasi pada kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih menghasilkan model yang sama untuk setiap kasus yaitu model EEE yang tebaran datanya berbentuk *ellipsoidal*. Kecuali pada kondisi jarak antarpusat *cluster* dekat, banyak data n=50 dan tingkat korelasi sedang menghasil model EEV (*ellipsoidal*) sebagai model terbaik. Model dan nilai BIC pada setiap kasus disajikan pada Lampiran 4.

Untuk 27 kasus yang kondisi satu *cluster* terpisah dan dua *cluster* saling tumpang tindih diperoleh hasil bahwa rataan persentase salah pengelompokan pada metode berbasis model jauh lebih kecil dibandingkan dengan metode Ward dan metode K-mean. Hal ini menunjukkan bahwa metode pengelompokan berbasis model lebih cenderung dapat memisahkan ketiga *cluster* dibandingkan dua metode

cluster lainnya. Hasil pengelompokan untuk 27 kasus pada kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih secara lengkap disajikan pada Lampiran 5.

Ditinjau dari jarak antarpusat *cluster*, terjadi penurunan persentasi salah pengelompokan dengan semakin jauh jarak antar pusat *cluster* untuk ketiga metode *cluster*. Hal ini dapat dilihat berdasarkan persentasi salah pengelompokan yang dihasilkan, yang disajikan pada Gambar 13.

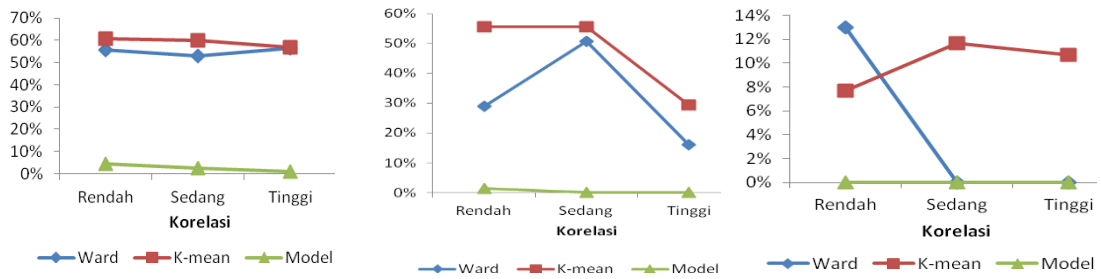


(a) (b) (c)

Gambar 13. Persentasi salah pengelompokan didasarkan pada ukuran jarak dengan tingkat korelasi (a) rendah, (b) sedang, dan (c) tinggi dengan banyak data n=50.

Penurunan salah persentasi ini disebabkan oleh ukuran jarak antarvektor rataan *cluster* yang relatif makin jauh untuk semua kondisi, sehingga objek-objek pengamatan akan semakin mengelompok di sekitar vektor rataannya.

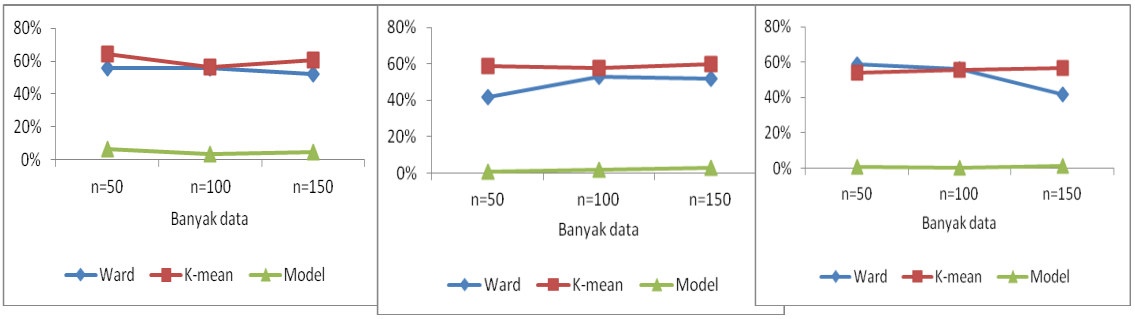
Untuk tingkat korelasi rendah, sedang, dan jauh menunjukkan bahwa pada metode berbasis model terjadi penurunan persentase salah pengelompokan dari tingkat korelasi rendah ke tingkat korelasi tinggi, walaupun penurunan ini hampir tidak ada perbedaan yang berarti. Hal ini menunjukkan bahwa tingkat korelasi yang berbeda tidak berpengaruh secara signifikan pada kondisi *cluster* pada kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih (Gambar 14).



(a) (b) (c)

Gambar 14. Persentasi salah pengelompokan yang didasarkan pada tingkat korelasi dengan ukuran jarak (a) dekat, (b) sedang, dan (c) jauh dengan banyak data n=100.

Ditinjau dari banyak objek pengamatan, banyak amatan tiap *cluster* sebesar 50 mempunyai pola persentase salah pengelompokan yang tidak jauh berbeda dengan objek pengamatan tiap *cluster* sebesar 100 dan 150. Hal ini berarti bahwa banyak amatan tiap *cluster* yang dicobakan tidak terlalu berpengaruh terhadap hasil *cluster* (Gambar 15).



(a) (b) (c)

Gambar 15. Persentasi salah pengelompokan didasarkan pada banyaknya data dengan tingkat korelasi (a) rendah, (b) sedang, dan (c) tinggi dengan jarak antar pusat *cluster* dekat.

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

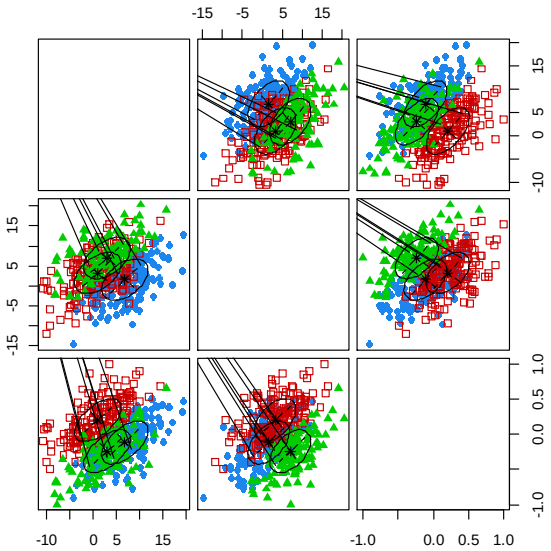
4.4. Ketiga *Cluster* Saling Tumpang Tindih

Untuk kondisi pengelompokan dengan ketiga *cluster* saling tumpang tindih terdapat 27 kasus simulasi data yang dibedakan atas jarak antar pusat *cluster* (dekat, sedang, dan jauh) dengan variansi dari ketiga *cluster* adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$; tingkat korelasi antar variabel adalah rendah (0.2), sedang (0.5), dan tinggi (0.8); dan banyak objek pengamatan pada tiap *cluster* adalah *n*=50, *n*=100, dan *n*=150.

Sebagai ilustrasi pertama diambil kasus data simulasi dengan ketiga *cluster* saling tumpang tindih, jarak antarpusat *cluster* sedang (d=7.483) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, tingkat korelasi antar variabel sedang (0.5), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=150.

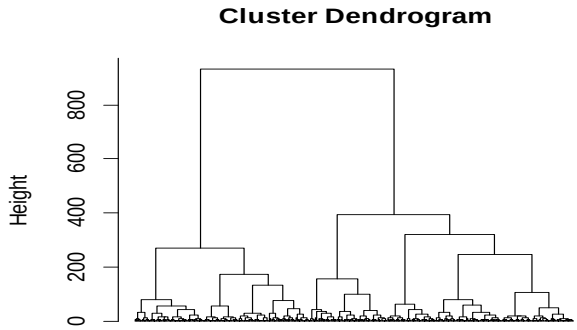
Secara visual matriks plot dari kondisi ketiga *cluster* saling terpisah dengan jarak antarpusat *cluster* sedang dengan variansi cenderung besar, tingkat korelasi

sedang (0,5), dan banyak data sebesar *n*=150 menunjukkan bahwa ketiga *cluster* saling tumpang tindih (Gambar 16).



Gambar 16. Matriks plot data simulasi untuk kondisi ketiga *cluster* saling tumpang tindih, jarak antar pusat *cluster* sedang (d=7.483) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, tingkat korelasi antar variabel sedang (0.5), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=150.

Pengelompokan dengan metode Ward diperoleh hasil bahwa dari 150 objek amatan pada *cluster* 1 terdapat 47 objek amatan masuk ke dalam *cluster* 2 dan 25 objek amatan masuk ke dalam *cluster* 3, dari 150 objek amatan pada *cluster* 2 terdapat 49 objek amatan masuk ke dalam *cluster* 1 dan 21 objek amatan masuk ke dalam *cluster* 3, dan dari 150 objek amatan pada *cluster* 3 terdapat 52 objek amatan masuk ke dalam *cluster* 1 dan 64 objek amatan masuk ke dalam *cluster* 2 (Gambar 17). Rataan persentasi salah pengelompokannya adalah 57,33%. Hal ini memnunjukkan bahwa metode Ward tidak mampu memisahkan *cluster* yang saling tupang tindih.

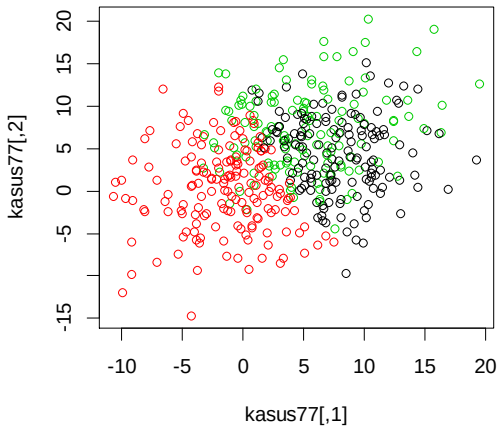


dist(kasus77)
hclust (*, "ward")

```
[1] 1 2 1 1 3 3 2 3 3 3 3 3 2 3 2 1 1 3 3 3 2 3 3 2 3 2 3 3 1 3 3 3 3 3 3 3 3 3 3 1 3 1 3 3 2 1 1 3 1
 [51] 1 1 1 1 1 2 2 1 2 3 3 1 1 2 3 1 3 1 3 3 1 2 3 3 1 2 2 2 3 3 1 3 2 1 2 2 2 3 3 1 3 3 1 2 3 2 3 2 3 3
[101] 3 1 3 3 1 3 3 3 3 1 3 1 1 3 1 3 1 3 3 3 3 3 1 1 1 3 1 3 3 3 3 3 1 1 2 3 1 1 1 3 2 3 1 3 1 3 1 3 3 1 3 3
[151] 1 1 1 2 1 2 1 3 3 1 1 1 1 1 1 1 2 1 1 2 2 3 1 3 3 1 1 2 1 1 1 3 1 1 3 1 3 1 3 3 3 1 1 3 3 1 1 1 1 3 1
[201] 3 3 2 3 1 3 3 2 1 3 2 1 1 3 2 1 3 1 3 2 1 3 3 2 1 1 1 1 3 3 1 1 2 1 2 3 3 3 1 3 3 2 3 1 1 3 1 1 3 2
[251] 1 3 1 1 3 1 3 1 1 1 2 3 3 1 2 1 1 3 1 3 1 3 1 1 3 2 1 3 1 1 3 1 3 2 1 1 1 3 2 1 1 1 1 1 1 1 1 3 3 1
[301] 3 1 1 1 3 3 1 1 3 3 2 1 1 1 3 1 2 1 3 2 1 2 1 1 1 1 1 1 1 2 3 2 3 1 3 2 2 1 1 1 3 2 1 1 3 1 2 1 3 2
[351] 3 3 1 3 1 2 1 2 1 1 1 1 3 3 3 1 1 2 1 2 2 3 2 1 3 2 3 3 1 1 3 1 1 2 3 1 3 1 3 3 1 1 2 2 1 3 2 3 2 3
[401] 2 2 2 2 1 2 2 1 3 1 1 3 1 3 3 3 3 1 1 3 3 1 1 2 3 1 2 3 2 1 3 3 1 3 2 1 3 1 3 3 3 3 1 2 1 3 3 1 3 1 3 1 3
```

Gambar 17. Dendogram dan hasil pengelompokan metode Ward dengan kondisi ketiga *cluster* saling tumpang tindih, jarak antar pusat *cluster* sedang (d=7.483) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, tingkat korelasi antar variabel sedang (0.5), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=150.

Hasil pengelompokan dengan metode K-rataan diperoleh hasil bahwa dari 150 objek amatan pada *cluster* 1 terdapat 66 objek amatan masuk ke dalam *cluster* 2 dan 8 objek amatan masuk ke dalam *cluster* 3, dari 150 objek amatan pada *cluster* 2 terdapat 32 objek amatan masuk ke dalam *cluster* 1 dan 58 objek amatan masuk ke dalam *cluster* 3, dan dari 150 objek amatan pada *cluster* 3 terdapat 28 objek amatan masuk ke dalam *cluster* 1 dan 39 objek amatan masuk ke dalam *cluster* 2 (Gambar 18). Rataan persentasi salah pengelompokannya adalah 51,33%. Hal ini memnunjukkan bahwa metode K-mean juga tidak mampu memisahkan *cluster* yang saling tupang tindih.



Cluster means:

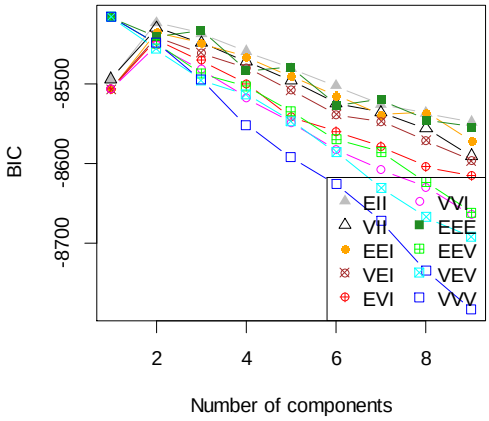
	[,1]	[,2]	[,3]
1	7.891147	4.4279495	1.9496817
2	-1.400521	0.1792769	0.1141793
3	5.186067	7.0636385	9.8577783

Clustering vector:

```
 [1] 2 2 2 1 3 3 2 3 3 3 3 3 2 3 2 2 2 3 3 1 2 3 3 2 3 2 3 3 2 3 3 3 3 3 3 3 3 3 1 3 2 3 2 3 3 2 2 2 3 1
 [51] 2 2 1 2 2 2 2 2 2 3 3 2 2 2 3 2 3 3 3 3 3 2 2 3 3 2 3 2 2 2 2 3 2 3 2 3 2 2 2 2 2 3 3 2 3 3 1 2 3 2 3 2 3 3
[101] 3 2 3 3 2 3 3 3 3 2 3 2 1 3 2 3 2 3 3 3 3 1 2 3 3 2 3 3 3 2 2 2 2 3 2 2 2 3 2 2 2 3 2 3 2 3 3 3 2 3 3
[151] 1 2 2 2 1 2 1 3 3 1 2 2 2 2 2 1 2 3 1 2 2 3 3 1 1 1 2 2 1 1 2 1 2 1 1 1 2 1 2 3 3 3 1 3 3 2 1 1 2 3 2
[201] 3 1 2 1 1 3 3 2 2 3 2 2 2 1 2 1 3 2 3 2 3 1 3 2 1 1 1 2 1 3 2 2 2 1 2 3 1 3 2 3 1 2 1 1 2 3 2 1 3 2
[251] 1 3 2 1 1 1 1 1 2 2 2 3 1 2 2 1 2 1 2 1 1 3 1 3 2 1 3 2 1 1 1 1 1 1 1 2 2 2 3 1 2 1 2 2 2 2 2 1 2 3 1 2
[301] 1 1 1 1 1 3 1 1 3 3 2 1 1 1 1 3 1 1 3 1 1 2 1 2 1 2 1 1 1 2 3 2 1 1 1 2 2 1 1 2 3 1 1 2 1 1 2 1 1 1 1
[351] 3 1 1 3 2 2 1 2 1 1 1 1 1 3 3 1 1 1 1 2 2 3 2 1 1 2 3 3 2 1 1 2 1 2 1 1 1 1 3 3 2 1 2 2 1 1 2 1 2 3
[401] 2 2 2 2 1 2 2 1 1 1 3 1 3 3 1 3 1 1 3 3 2 2 2 3 2 2 1 2 1 1 1 1 3 1 1 1 1 3 1 1 1 2 1 3 1 1 3 1 1
```

Gambar 18. Plot dan hasil pengelompokan metode K mean dengan kondisi ketiga *cluster* saling tumpang tindih, jarak antar pusat *cluster* sedang (d=7.483) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, tingkat korelasi antar variabel sedang (0.5), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=150.

Pada metode berbasis model, terdapat tiga model yang paling layak yakni model EEE (3 *cluster*) dengan nilai BIC = -8432.451; model EII (3 *cluster*) dengan nilai BIC = -8435.782; dan model VII (3 *cluster*) dengan nilai BIC = -8447.913. Model terbaik dari tiga model yang paling layak terdapat nilai BIC yang paling besar yaitu pada model EII (Gambar 19).



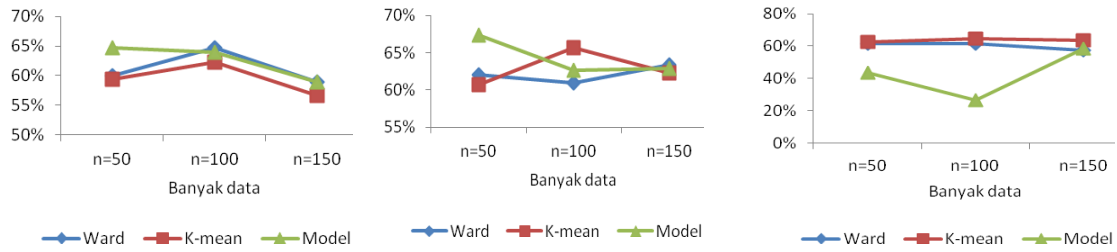

```
 [1] 3 2 1 1 2 2 2 2 2 2 1 2 1 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 1
 [51] 2 1 3 1 2 3 2 3 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 1 1 2 2 2 2 2 2 2
[101] 2 2 2 2 1 1 2 2 2 2 2 2 3 2 1 2 2 2 2 2 2 3 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2
[151] 1 3 2 1 3 3 3 3 3 3 3 3 2 3 3 3 2 3 2 3 3 3 3 3 3 1 3 3 2 1 3 1 3 3 1 3 2 2 2 3 2 2 3 3 3 3 2
[201] 3 3 1 3 3 3 2 2 3 3 3 3 3 1 3 2 2 3 3 3 3 2 3 3 3 3 3 3 1 3 3 1 3 2 3 3 3 2 3 1 3 3 3 3 2 3 2 3
[251] 3 2 2 3 3 3 1 3 3 3 3 3 1 3 3 2 3 3 2 3 1 3 3 3 1 3 1 1 1 1 2 1 3 3 3 3 3 1 3 3 3 2 2 1 3 3 3 3 2
[301] 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 1 1 1 1 1 3 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 3 1 1 1 1 1
[351] 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 3 1 2 1 1 1 1 2 1 2 1 2 1 1 1 1 2 2 2 1 1 1 1 1 3 1 1 1
[401] 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 2 1 1 1 2 1 1 1 2 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 2 1 1 1 1 2 1 1

classification table:
  1   2   3
160 173 117

best BIC values:
   EEE,3    EII,3    VII,3
-8432.451 -8435.782 -8447.913
```

Gambar 19. Plot dan hasil pengelompokan metode berbasis model dengan kondisi ketiga *cluster* saling tumpang tindih, jarak antar pusat *cluster* sedang (d=7.483) dengan variansi ketiga *cluster* adalah $\sigma_1^2 = 25, \sigma_2^2 = 25, \sigma_3^2 = 25$, tingkat korelasi antar variabel sedang (0.5), dan banyak objek pengamatan pada tiap *cluster* adalah *n*=150.

Secara keseluruhan dari 27 kasus simulasi pada kondisi ketiga *cluster* saling tumpang tindih, ketiga metode *cluster* tidak mampu memisahkan memisahkan *cluster* yang saling tumpang tindih. Metode *cluster* 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). Pada tingkat korelasi rendah dan sedang cenderung menghasilkan model VII dan EII, sedangkan tingkat korelasi tinggi menghasilkan model EEE. Model dan nilai BIC pada setiap kasus disajikan pada Lampiran 4. Pada tingkat korelasi tinggi dengan jarak antar pusat *cluster* relatif sedang dan jauh, rataan persentasi salah pengelompokan *cluster* lebih baik bila dibandingkan dengan motode Ward dan metode K-rataan. Namun pada tingkat korelasi rendah dan tingkat korelasi sedang, metode berbasis model tidak mampu memisahkan *cluster* yang saling tumpang tindih, bahkan sebagian besar persentase salah pengelompokannya lebih besar dibandingkan dengan metode Ward dan metode K-rataan. Hal ini mungkin

disebabkan oleh objek-objek pengamatannya mengelompok pada satu *cluster*, sehingga secara geometris dari 10 model metode pengelompokan berbasis model tidak mampu memisahkan *cluster* yang saling tumpang tindih. Bahkan metode pengelompokan berbasis model ini menganjurkan bahwa akan lebih efektif jika pengelompokannya dibagi dalam satu atau dua atau empat *cluster*.

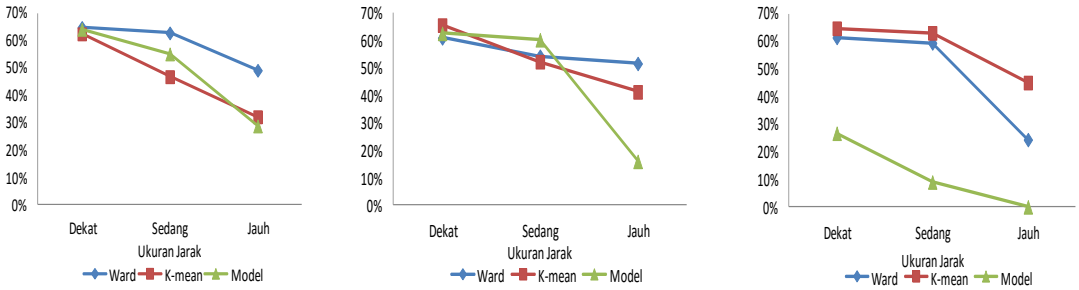
Pada kondisi ketiga *cluster* saling tumpang tindih ini, perbedaan banyak objek-objek pengamatan tiap *cluster* tidak terlalu berpengaruh terhadap persentase salah pengelompokannya, baik pada tingkat korelasi maupun pada jarak antar pusat *cluster* (Gambar 20).



(a) (b) (c)

Gambar 20. Persentasi salah pengelompokan didasarkan pada banyaknya data dengan tingkat korelasi (a) rendah, (b) sedang, dan (c) tinggi dengan jarak antar pusat *cluster* dekat.

Ditinjau dari jarak antar pusat *cluster*, terjadi penurunan persentasi salah pengelompokan dengan semakin jauh jarak antar pusat *cluster* untuk ketiga metode *cluster* baik baik pada tingkat korelasi maupun pada banyak objek amatan tiap *cluster*. Hal ini dapat dilihat berdasarkan persentasi salah pengelompokan yang dihasilkan, yang disajikan pada Gambar 21.

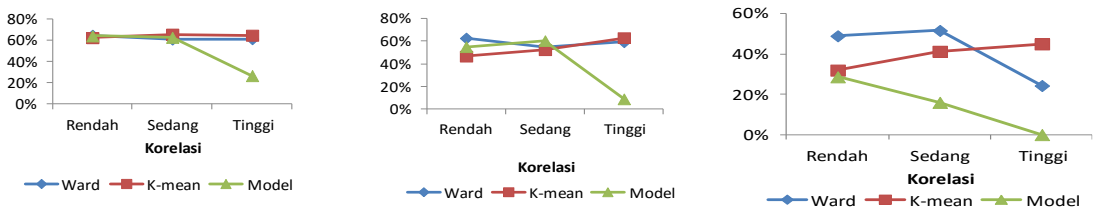


(a) (b) (c)

Gambar 21. Persentasi salah pengelompokan didasarkan pada ukuran jarak dengan tingkat korelasi (a) rendah, (b) sedang, dan (c) tinggi dengan banyak data n=100.

Untuk tingkat korelasi rendah, sedang, dan jauh 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 *cluster* pada kondisi ketiga *cluster* saling tumpang tindih (Gambar 22).



(a) (b) (c)

Gambar 22. Persentasi salah pengelompokan yang didasarkan pada tingkat korelasi dengan ukuran jarak (a) dekat, (b) sedang, dan (c) jauh dengan banyak data n=100.

Dari hasil pengelompokan ketiga metode *cluster* yang dibandingkan dengan kondisi ketiga *cluster* saling tumpang tindih menunjukkan bahwa metode pengelompokan berbasis model lebih efektif memisahkan *cluster* yang saling tumpang tindih apabila tingkat korelasi tinggi dan jarak antarpusat *cluster* relatif sedang dan jauh. Sebaliknya, apabila tingkat korelasi tinggi dengan jarak antarpusat *cluster* relatif dekat dan juga pada tingkat korelasi rendah dan sedang dengan jarak antar pusat *cluster* dekat, sedang dan jauh, ketiga metode yang dibandingkan tidak efektif dalam memisahkan *cluster* yang tumpang tindih.

4.5. Data *Iris*

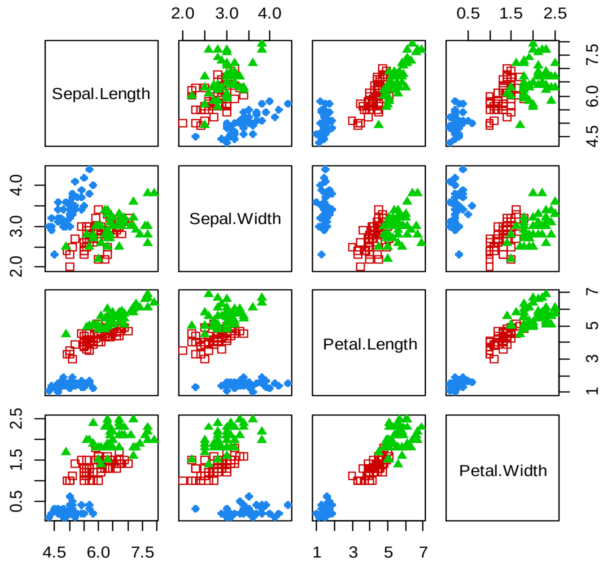
Data *Iris* merupakan contoh klasik yang sering digunakan dalam buku-buku teks statistik untuk mengilustrasikan masalah analisis *cluster*. Data *Iris* ini adalah sejenis bunga yang terdiri dari 4 variabel yaitu, panjang petal, lebar petal, panjang sepal, dan lebar sepal. Masing-masing variabel terdiri dari 150 pengamatan, setiap ukuran variabel terbagi atas tiga spesies yaitu *Iris setosa* , *Iris versicolor*, dan *Iris virginica* yang masing-masing terdiri dari 50 pengamatan (lihat Lampiran 7).

Sebelum menerapkan analisis *cluster* terhadap data *Iris*, terlebih dahulu diberikan gambaran umum tentang statistik deskriptif dan matriks plot keempat variabel yang diamati yang disajikan pada Tabel 3 dan Gambar 23.

Tabel 3. Statistik deskriptif data *Iris*

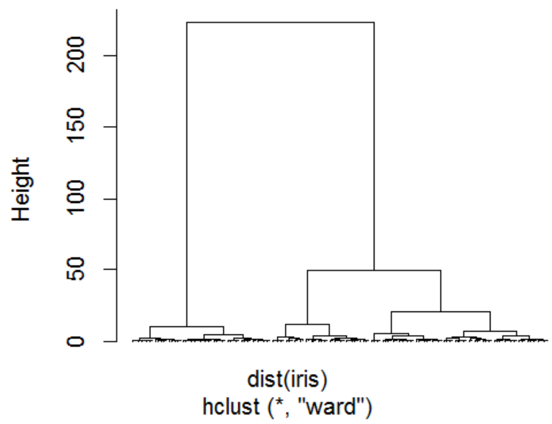
Kelompok spesies	Variabel	Rataan	Standar Deviasi
Iris setosa	Panjang sepal	5.006	0.353
	Lebar sepal	3.428	0.379
	Panjang petal	1.462	0.174
	Lebar petal	0.246	0.105
Iris versicolor	Panjang sepal	5.936	0.516
	Lebar sepal	2.770	0.314
	Panjang petal	4.260	0.470
	Lebar petal	1.326	0.198
Iris virginica	Panjang sepal	6.577	0.636
	Lebar sepal	2.974	0.323
	Panjang petal	5.552	0.552
	Lebar petal	2.026	0.275

Berdasarkan statistik deskriptif data *Iris* pada Tabel 5 tampak bahwa rataan dan standar deviasi variabel panjang petal untuk spesies *Iris setosa* jauh lebih kecil dibandingkan dengan spesies *Iris versicolor* dan *Iris virginica*, demikian juga untuk variabel lebar petal dan panjang sepal, walaupun perbedaannya tidak sebesar panjang petal. Variabel panjang sepal untuk spesies *Iris setosa*, rataan dan standar deviasinya sedikit lebih besar daripada spesies *Iris versicolor* dan *Iris virginica*. Matriks plot data *Iris* pada Gambar 23 menunjukkan bahwa *Iris setosa* terpisah dari spesies *Iris versicolor* dan *Iris virginica*. Gambaran umum dari statistik deskriptif dan matriks plot data *Iris* ini dapat mewakili kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih.



Gambar 23. Matriks plot data *Iris*

Hasil pengelompokan dengan metode Ward diperoleh hasil bahwa untuk spesies *Iris setosa* dan spesies *Iris versicolor* secara sempurna dapat dipisahkan dari spesies lainnya sedangkan untuk spesies *Iris virginica*, 36 objek amatan sesuai dan tepat masuk ke kelompok spesies *Iris virginica* dan 14 objek amatan masuk ke dalam kelompok spesies *Iris versicolor.*

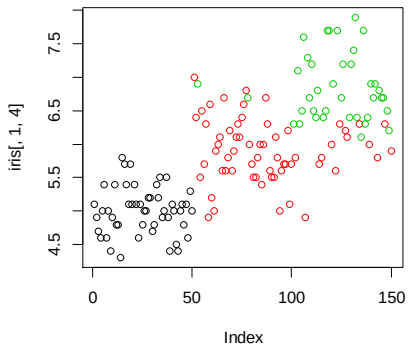



```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [51] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[101] 3 2 3 3 3 3 2 3 3 3 3 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3 3 3 2 3 3 2
```

Gambar 24. Dendogram dan hasil pengelompokan metode Ward pada data.

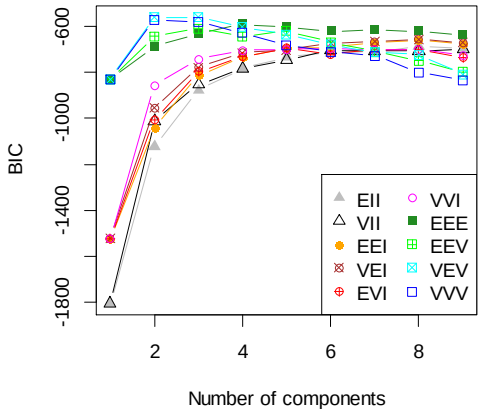
Hasil pengelompokan dengan metode K-rataan diperoleh hasil bahwa untuk spesies *Iris setosa* secara sempurna dapat dipisahkan dari spesies lainnya. Untuk spesies *Iris versicolor*, 49 objek amatan sesuai dengan tepat masuk pada kelompok

spesies *Iris versicolor* dan 1 objek amatan masuk ke dalam kelompok spesies *Iris virginica*. Sementara spesies *Iris virginica*, 36 objek amatan sesuai dan tepat masuk ke kelompok spesies *Iris virginica* dan 14 objek amatan masuk ke dalam kelompok spesies *Iris versicolor*.




```
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [51] 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[101] 3 2 3 3 3 3 2 3 3 3 3 3 2 2 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3 3 3 2 3 3 2
```

Gambar 25. Plot dan hasil pengelompokan metode K-rataan pada data *Iris*




```
Classification:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [51] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 2 3 2 2 2 2 3 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[101] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

classification table:
 1  2  3
50 45 55

best BIC values:
   VEV,3     VVV,3     EEV,3
-562.5514 -580.8399 -610.0853
```

Gambar 26. Plot dan hasil pengelompokan metode berbasis model pada data *Iris*

Pada metode berbasis model, terdapat tiga model paling layak yakni model VEV (3 *cluster*) dengan nilai BIC = -562,5541, model VVV (3 *cluster*) dengan nilai

BIC -580,8399, dan model EEV (3 *cluster*) dengan nilai BIC -610,0853. Model terbaik dari tiga model yang paling layak terdapat pada nilai BIC yang paling besar yaitu pada model VEV yang tebaran datanya berbentuk e*llipsoidal*. Hasil pengelompokan dengan metode berbasis model diperoleh hasil bahwa ketiga spesies, yakni spesies *Iris setosa* dan spesies *Iris versicolor*, sementara pada spesies *Iris virginica* terdapat 5 objek amatan masuk ke dalam kelompok *Iris versicolor*.

Untuk spesies *Iris versicolor*, metode Ward dan metode berbasis model secara sempurna dapat dipisahkan dengan spesies lainnya, sedangkan metode K-rataan dua spesies *Iris versicolor* masuk ke dalam spesies *Iris virginica.* Untuk spesies *Iris virginica* hanya metode berbasis model yang dapat memisahkan antar spesies, sementara metode Ward dan metode K-rataan terdapat 14 objek amatan masuk pada kelompok spesies *Iris versicolor.* Salah pengelompokan terkecil terjadi pada metode pengelompokan berbasis model sebesar 3.33% (5 amatan), sementara persentase salah pengelompokan metode Ward sebesar 9,33% (14 amatan) dan metode K-rataan sama besar, yaitu 10.67% (16 amatan). Hasil pengelompokan untuk masing-masing metode *cluster* untuk data *Iris* disajikan pada Tabel 4.

Salah pengelompokan yang terjadi pada data *Iris* ini hanya melibatkan spesies *Iris versicolor* dan spesies *Iris virginica*, sementara untuk spesies *Iris setosa* tidak terpengaruh untuk ketiga metode. Hal ini disebabkan oleh cukup dekatnya jarak antar pusat *cluster* spesies *Iris versicolor* dengan spesies *Iris virginica* (d= 1,62), sementara jarak antar pusat *cluster* spesies *Iris setosa* dengan spesies *Iris versicolor* (d=3.21) dan jarak antar pusat *cluster* spesies *Iris setosa* dengan spesies *Iris virginica* (d=4.75) cukup jauh, sehingga menyebabkan spesies IS memang benar-benar terpisah dari dua spesies lainnya. Hal ini juga didukung data visual matriks plot data *Iris* pada Gambar 23.

Tabel 6. Hasil pengelompokan data *Iris* menjadi 3 gorombol dan persentase salah pengelompokannya.

Metode *cluster*	***Iris setosa* (50,0,0)**	***Iris versicolor* (0,50,0)**	***Iris virginica* (0,0,50)**	**Salah pengelompokan**
Ward	(50,0,0)	(0,50,0)	(0,14,36)	14 (9.33%)
k-rataan	(50,0,0)	(0,48,2)	(0,14,36)	16 (10.67%)
Berbasis model	(50,0,0)	(0,45,5)	(0,0,50)	5 (3.33%)

Ket. (50,0,0) : 50 masuk kelompok IS, 0 masuk kelompok IC dan 0 masuk kelompok IV

V. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan penelitian ini, dihasilkan beberapa kesimpulan sebagai berikut :

1. Semakin jauh jarak antarpusat *cluster* dengan variansi yang tetap maka persentase salah pengelompokan yang dihasilkan semakin kecil.
2. Besar kecilnya ukuran data pada tiap *cluster* tidak berpengaruh terhadap hasil persentase salah pengelompokan yang dihasilkan.
3. Pada metode berbasis model, semakin besar tingkat korelasi antarvariabel maka persentase salah pengelompokan yang dihasilkan semakin kecil.
4. Untuk kondisi ketiga *cluster* saling terpisah, ketiga metode yang dibandingkan memberikan hasil pengelompokan yang sama dan sesuai dengan hasil pengelompokan sebenarnya.
5. Untuk kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih, metode berbasis model memberikan hasil yang lebih baik dibandingkan dengan metode Ward dan metode K-rataan.
6. Untuk kondisi ketiga *cluster* saling tumpang tindih dengan tingkat korelasi tinggi dan jarak antarpusat *cluster* sedang dan jauh, hasil pengelompokan berbasis model lebih baik dibandingkan dengan metode Ward dan metode K-rataan. Sedangkan rendah dan sedang dengan jarak antar pusat *cluster* dekat, sedang dan jauh, ketiga metode pengelompokan tidak cukup efektif memisahkan ketiga *cluster* yang saling tumpang tindih.

5.2. Saran

Kesimpulan ini berlaku untuk variabel yang memiliki sebaran campuran normal dan tanpa ada data pencilan. Diperlukan penelitian lebih lanjut dengan memperhatikan sebaran campuran tidak normal, data yang mengandung pencilan.

DAFTAR PUSTAKA

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

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

Dempster, A. P., Laird, N. M. and Rubin D. B. (1977). Maximum Likelihood from Incomplete Data Via The EM Algorithm, *J. R. Statistics Society* B, Vol 39, hal 1-38.

Fraley, C. & Raftery A.E. (1998). How Many *Cluster*? Which *Cluster*ing Method? Answer via Model-Based *Cluster* Analysis. *The Computer Journal* 41; 578-588

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

Fraley, C. and Raftery, A. E. (2002). MCLUST: Software for rvlodel-Based *Cluster*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 *Cluster*ing." *Technical Report* 504, University of Washington, Department of Statistics.

Hair, J.E., Jr., R.E. Anderson, R.L. Tatham, and W.C. Black. (1998). *Multivariate Data Analysis*, Prentice-Hall, Inc., 5th ed.

Härdle W. and Simar L. (2007). *Applied Multivariate Statistical Analysis*, 2 th Edition, Springer-Verlag: Berlin Heidelberg

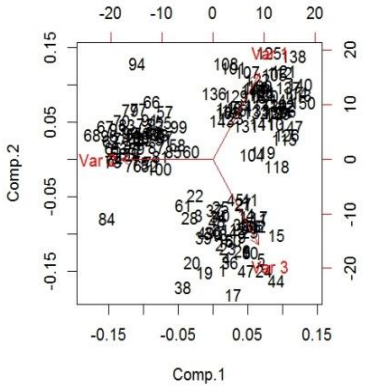
Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6 th Edition, New Jersey: Prentice-Hall.

Mc Lachlan, G.J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

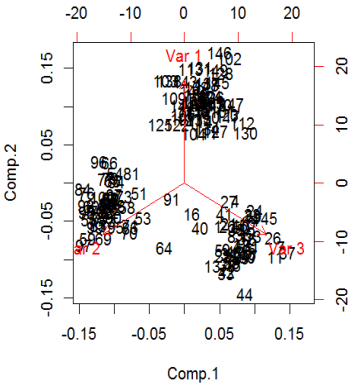
Pardede,T. (2008). Perbandingan Metode Berbasis Model (*Model-Based*) dengan Metode Metode K-mean dalam Analsis Gugus. *Jurnal Sigma, Sains dan Teknologi* Vol 11, No. 2; 157-166

LAMPIRAN

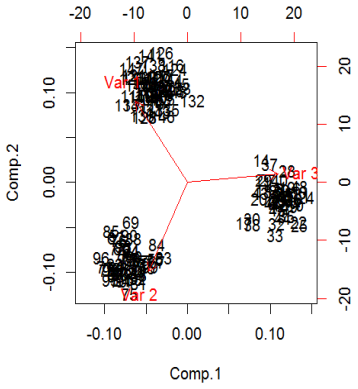
Lampiran 1. Pola simulasi data untuk kondisi ketiga *cluster* saling terpisah dengan banyak objek pengamatan untuk tiap *cluster* sebesar n=50



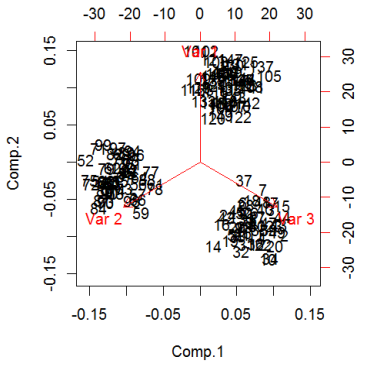
a. Jarak dekat, tingkat korelasi rendah dan n=50



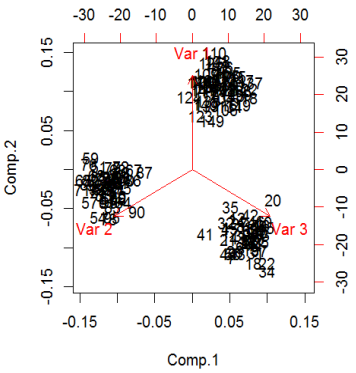
b. ,Jarak sedang, tingkat korelasi rendah dan n=50



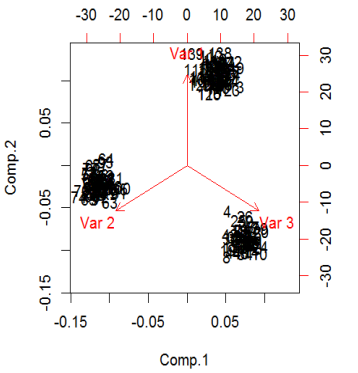
c. Jarak jauh, tingkat korelasi rendah dan n=50



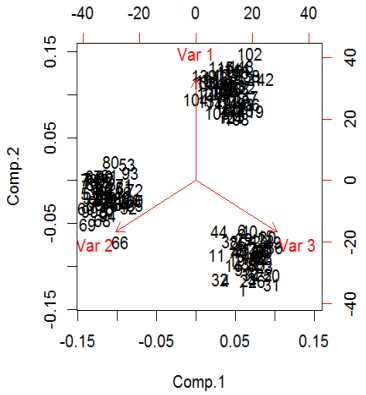
d. Jarak dekat, tingkat korelasi sedang dan n=50



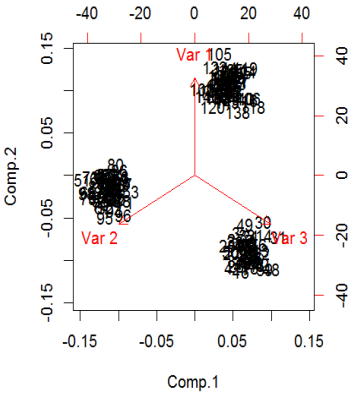
e. Jarak sedang, tingkat korelasi sedang dan n=50



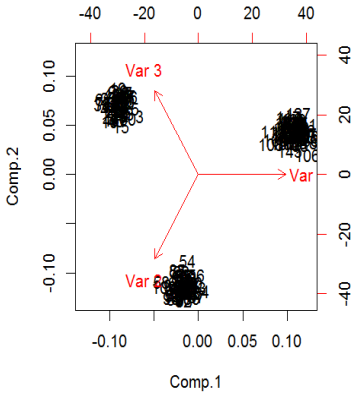
f. Jarak jauh, tingkat korelasi sedang dan n=50



g. Jarak dekat, tingkat korelasi tinggi dan n=50

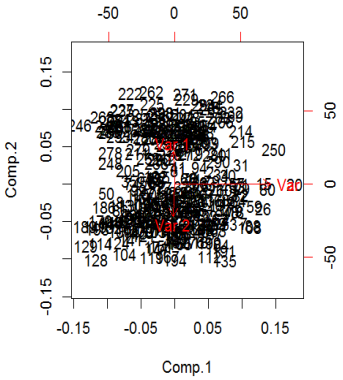


h. Jarak sedang, tingkat korelasi tinggi dan n=50

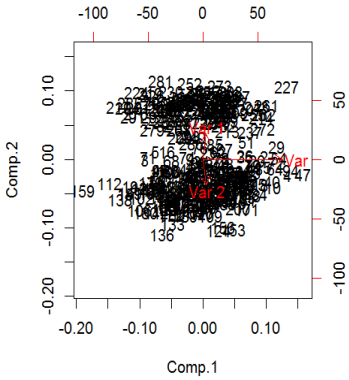


i. Jarak jauh, tingkat korelasi tinggi dan n=50

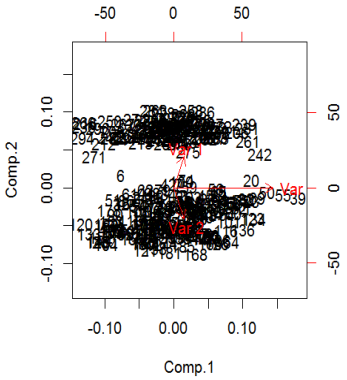
Lampiran 2. Pola simulasi data untuk kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih dengan banyak objek pengamatan tiap *cluster* sebesar n=100



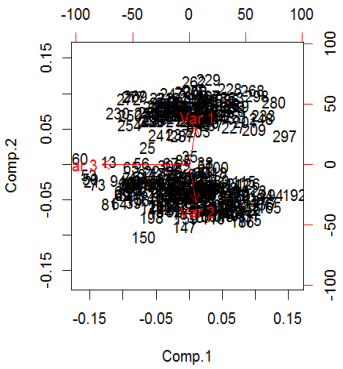
a. Jarak dekat, tingkat korelasi rendah dan n=100



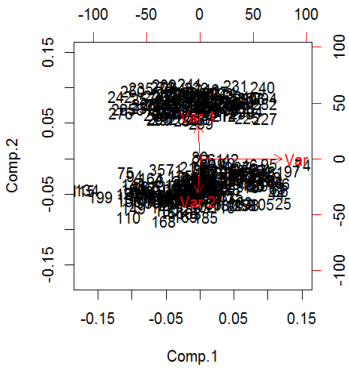
b. Jarak sedang, tingkat korelasi rendah dan n=100



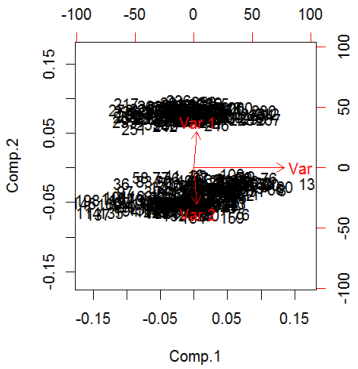
c. Jarak jauh, tingkat korelasi rendah dan n=100



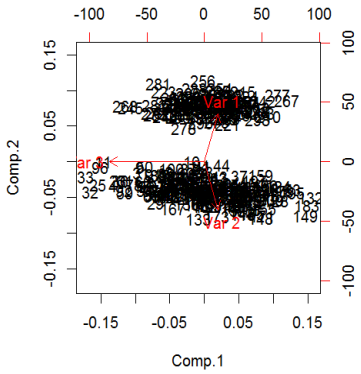
d. Jarak dekat, tingkat korelasi sedang dan n=100



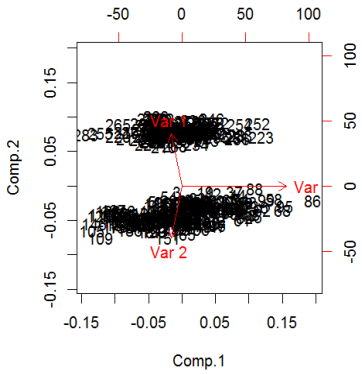
e. Jarak sedang, tingkat korelasi sedang dan n=100



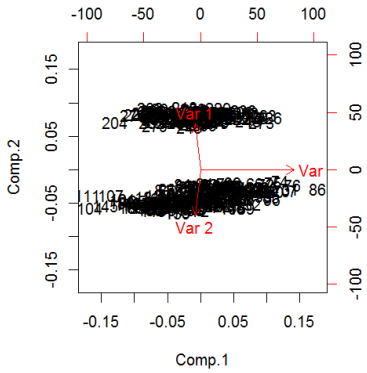
f. Jarak jauh, tingkat korelasi sedang dan n=100



g. Jarak dekat, tingkat korelasi tinggi dan n=100

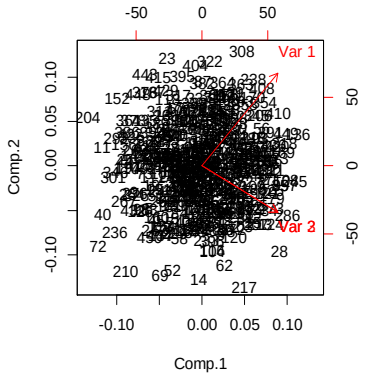


h. Jarak sedang, tingkat korelasi tinggi dan n=100

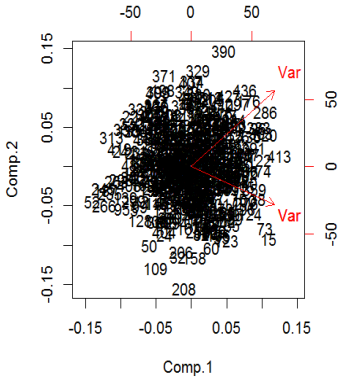


i. Jarak jauh, tingkat korelasi tinggi dan n=100

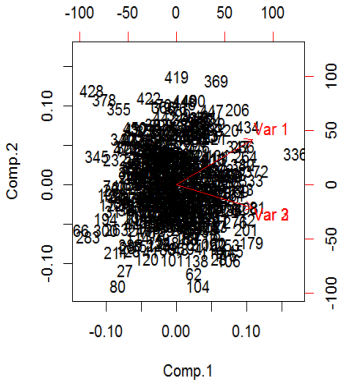
Lampiran 3. Pola simulasi data untuk kondisi ketiga *cluster* saling tumpang tindih dengan banyak objek pengamatan untuk tiap *cluster* sebesar n=150



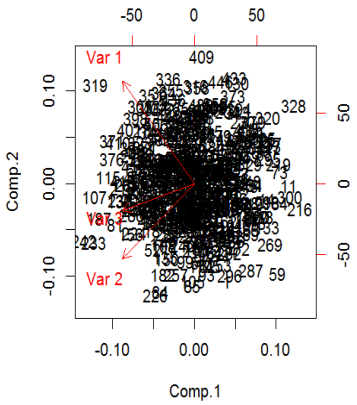
a. Jarak dekat, tingkat korelasi rendah dan n=150



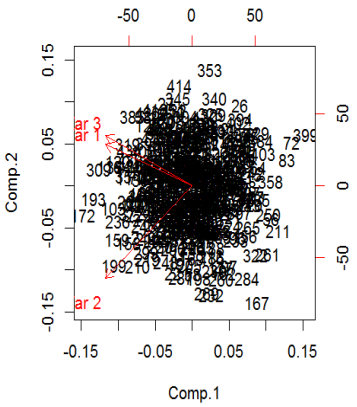
b. Jarak sedang, tingkat korelasi rendah dan n=150



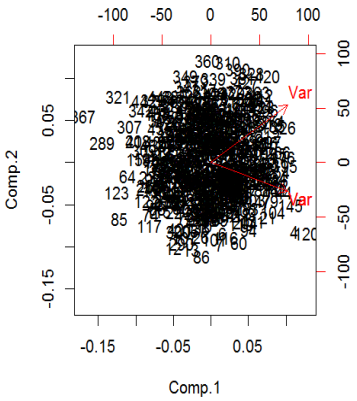
c. Jarak jauh, tingkat korelasi rendah dan n=150



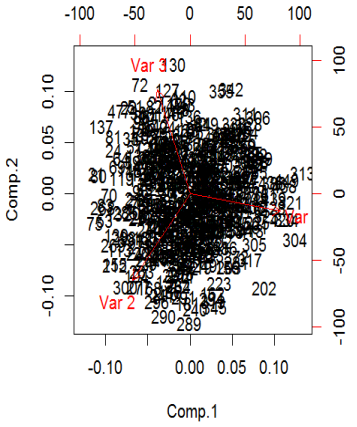
d. Jarak dekat, tingkat korelasi sedang dan n=150



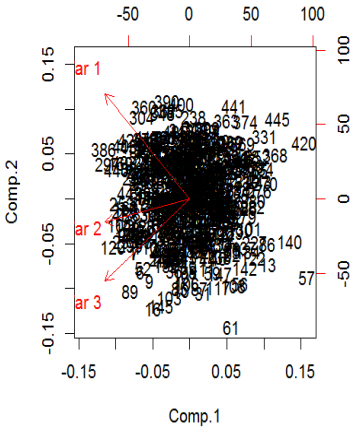
e. Jarak sedang, tingkat korelasi sedang dan n=150



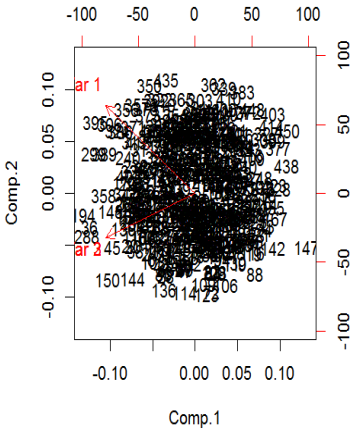
f. Jarak jauh, tingkat korelasi sedang dan n=150



g. Jarak dekat, tingkat korelasi tinggi dan n=150



h. Jarak sedang, tingkat korelasi tinggi dan n=150



i. Jarak jauh, tingkat korelasi tinggi dan n=150

Lampiran 4. Nilai BIC dan model terbaik pada metode *cluster* berbasis model

Banyak data	Jarak antar pusat cluster	Tingkat korelasi	Ketiga *cluster* saling terpisah $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=1$		satu *cluster* terpisah dan dua *cluster* tindih $\sigma_1^2=1,\sigma_2^2=1,\sigma_3^2=25$		Keiga *cluster* 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 (0,2)	-1652,769	EII	-2116,695	EEI	-2786,649	VII
		Sedang (0,5)	-1578,400	EEE	-2051,198	EEE	-2740,561	EII
		Tinggi (0,8)	-1343,213	EEE	-1826,012	EEE	-2639,369	EEE
	Sedang (d=7.483)	Rendah (0,2)	-1657,665	EII	-2137,900	EEE	-2838,331	VII
		Sedang (0,5)	-1578,746	EEE	-2061,239	EEE	-2826,904	EII
		Tinggi (0,8)	-1343,213	EEE	-1826,045	EEE	-2726,018	EEE
	Jauh (d= 9.899)	Rendah (0,2)	-1657,665	EII	-2148,996	EEE	-2936,891	EII
		Sedang (0,5)	-1578,746	EEV	-2061,576	EEE	-2925,486	EII
		Tinggi (0,8)	-1343,213	EEE	-1826,045	EEE	-2764,566	EEE
100	Dekat (d=5.099)	Rendah (0,2)	-3264,336	EEE	-4.205,786	EEE	-5595,077	EII
		Sedang (0,5)	-3093,097	EEE	-4018,418	EEE	-5489,043	EII
		Tinggi (0,8)	-2622,166	EEE	-3587,805	EEE	-5211,399	EEE
	Sedang (d=7.483)	Rendah (0,2)	-3268,231	EEE	-4216,587	EEE	-5709,513	EII
		Sedang (0,5)	-3093,232	EEE	-4058,344	EEE	-5637,339	EII
		Tinggi (0,8)	-2622,166	EEE	-3.587,829	EEE	-5403,803	EEE
	Jauh (d= 9.899)	Rendah (0,2)	-3268,231	EEE	-4233,654	EEE	-5833,120	VII
		Sedang (0,5)	-3093,232	EEE	-4058,876	EEE	-5799,674	EEV
		Tinggi (0,8)	-2622,166	EEE	-3587,829	EEE	-5486,362	EEE
150	Dekat (d=5.099)	Rendah (0,2)	-4858,828	EEE	-6221,692	EEE	-8349,897	VII
		Sedang (0,5)	-4601,239	EEE	-6009,734	EEE	-8197,145	EII
		Tinggi (0,8)	-3896,198	EEE	-6300,827	EEE	-7812,342	EEE
	Sedang (d=7.483)	Rendah (0,2)	-4865,295	EEE	-5698,814	EEE	-8537,436	EII
		Sedang (0,5)	-4602,796	EEE	-6047,875	EEE	-8432,451	EEE
		Tinggi (0,8)	-3896,198	EEE	-5344,670	EEE	-8.060,247	EEE
	Jauh (d= 9.899)	Rendah (0,2)	-4865,295	EEE	-6312,636	EEE	-8726,841	EII
		Sedang (0,5)	-4602,796	EEE	-6051,276	EEE	-8631,423	EEE
		Tinggi (0,8)	-3896,198	EEE	-5344,692	EEE	-8186,772	EEE

Lampiran 5. Hasil pengelompokan pada kondisi satu *cluster* terpisah dan dua *cluster* tumpang tindih

a. Banyak data tiap *cluster* n=50

Jarak antar pusat *cluster*	Tingkat korelasi	Metode	*Cluster* 1			*Cluster* 2			*Cluster* 3			Persentasi salah pengelompokan
			50	**0**	**0**	**0**	**50**	**0**	**0**	**0**	**50**	
		Ward	32	6	12	24	20	6	31	5	14	56,00%
	Rendah	K-mean	24	20	6	22	11	17	24	7	19	64,00%
		Model	50	0	0	0	48	2	3	5	42	6,67%
		Ward	31	19	0	24	26	0	0	20	30	42,00%
Dekat	Sedang	K-mean	21	14	15	20	20	10	23	6	30	58,67%
		Model	50	0	0	0	49	1	0	0	50	0,67%
		Ward	17	20	13	15	31	4	10	26	14	58,67%
	Tinggi	K-mean	31	4	15	22	20	8	20	12	18	54,00%
		Model	50	0	0	0	50	0	1	0	49	0,67%
		Ward	41	9	0	10	40	0	2	0	48	14,00%
	Rendah	K-mean	39	11	0	10	40	0	3	8	39	21,33%
		Model	48	0	2	1	49	0	0	0	50	2,00%
		Ward	49	1	0	14	36	0	9	8	33	21,33%
Sedang	Sedang	K-mean	37	0	13	8	41	1	11	0	39	22,00%
		Model	50	0	0	0	50	0	0	0	50	0,00%
		Ward	36	13	1	32	18	0	0	10	40	37,33%
	Tinggi	K-mean	37	11	2	6	39	5	5	9	36	25,33%
		Model	50	0	0	0	50	0	0	0	50	0,00%
		Ward	49	1	0	1	49	0	0	0	50	1,33%
	Rendah	K-mean	42	0	8	1	46	3	10	0	40	14,67%
		Model	50	0	0	0	50	0	0	0	50	0,00%
		Ward	36	14	0	0	50	0	0	0	50	9,33%
Jauh	Sedang	K-mean	44	0	6	0	44	6	0	10	40	14,67%
		Model	50	0	0	0	50	0	0	0	50	0,00%
		Ward	50	0	0	0	50	0	0	0	50	0,00%
	Tinggi	K-mean	50	0	0	0	43	7	0	9	41	10,67%
		Model	50	0	0	0	50	0	0	0	50	0,00%

b. Banyak data tiap *cluster* n=100

Jarak antar pusat *cluster*	Tingkat korelasi	Metode	*Cluster* 1			*Cluster* 2			*Cluster* 3			Persentasi salah pengelompokan
			100	**0**	**0**	**0**	**100**	**0**	**0**	**0**	**100**	
		Ward	41	25	34	20	74	6	36	46	18	55,67%
	Rendah	K-mean	35	42	23	27	52	21	10	47	43	56,67%
		Model	95	5	0	3	96	1	1	0	99	3,33%
		Ward	67	9	24	48	36	16	59	3	38	53,00%
Dekat	Sedang	K-mean	37	16	47	15	38	47	13	35	52	57,67%
		Model	98	2	0	3	97	0	1	0	99	2,00%
		Ward	67	26	7	30	53	17	49	40	11	56,33%
	Tinggi	K-mean	57	20	23	50	35	15	41	18	41	55,67%
		Model	100	0	0	0	100	0	0	0	100	0,00%
		Ward	98	2	0	62	38	0	0	23	77	29,00%
	Rendah	K-mean	57	20	23	50	35	15	41	18	41	55,67%
		Model	98	2	0	2	98	0	0	0	100	1,33%
		Ward	41	5	54	44	50	6	0	43	57	50,67%
Sedang	Sedang	K-mean	32	26	42	5	51	44	38	12	50	55,67%
		Model	100	0	0	0	100	0	0	0	100	0,00%
		Ward	100	0	0	39	61	0	0	9	91	16,00%
	Tinggi	K-mean	72	0	28	0	84	16	0	44	56	29,33%

		Model	100	0	0	0	100	0	0	0	100	0,00%
		Ward	100	0	0	39	61		0	0	100	13,00%
	Rendah	K-mean	87	13	0	10	90	0	0	0	100	7,67%
		Model	100	0	0	0	100	0	0	0	100	0,00%
		Ward	100	0	0	0	100	0	0	0	100	0,00%
Jauh	Sedang	K-mean	85	15	0	12	88	0	8	0	92	11,67%
		Model	100	0	0	0	100	0	0	0	100	0,00%
		Ward	100	0	0	0	100	0	0	0	100	0,00%
	Tinggi	K-mean	90	10	0	11	89	0	11	0	89	10,67%
		Model	100	0	0	0	100	0	0	0	100	0,00%

c. Banyak data tiap *cluster* n=150

Jarak antar pusat *cluster*	**Tingkat korelasi**	**Metode**	***Cluster* 1**			***Cluster* 2**			***Cluster* 3**			**Persentasi salah pengelompokan**
			150	**0**	**0**	**0**	**150**	**0**	**0**	**0**	**150**	
		Ward	66	69	15	8	71	71	71	0	79	52,00%
	Rendah	K-mean	46	72	32	37	76	37	65	30	55	60,67%
		Model	138	10	2	6	144	0	1	1	148	4,44%
		Ward	46	77	27	16	57	77	35	1	114	51,78%
Dekat	Sedang	K-mean	53	73	24	20	73	57	32	64	54	60,00%
		Model	141	8	1	3	147	0	0	0	150	2,67%
		Ward	104	46	0	85	65	0	30	28	92	42,00%
	Tinggi	K-mean	66	20	64	27	56	67	37	41	72	56,89%
		Model	98	2	0	3	97	0	0	0	100	1,11%
		Ward	92	58	0	3	147	0	64	0	86	27,78%
	Rendah	K-mean	118	32	0	40	110	0	35	3	112	24,44%
		Model	149	1	0	3	147	0	0	0	150	0,89%
		Ward	86	64	0	25	125	0	1	47	102	30,44%
Sedang	Sedang	K-mean	83	8	59	22	53	75	25	37	88	50,22%
		Model	149	1	0	0	150	0	0	0	150	0,22%
		Ward	150	0	0	32	118	0	49	14	87	21,11%
	Tinggi	K-mean	65	13	72	26	92	32	30	30	90	45,11%
		Model	150	0	0	0	150	0	0	0	150	0,00%
		Ward	132	18	0	29	121	0	0	0	150	10,44%
	Rendah	K-mean	129	21	0	23	127	0	10	1	139	12,22%
		Model	150	0	0	0	150	0	0	0	150	0,00%
		Ward	150	0	0	0	150	0	0	0	150	0,00%
Jauh	Sedang	K-mean	123	27	0	19	131	0	3	2	145	11,33%
		Model	150	0	0	0	150	0	0	0	150	0,00%
		Ward	150	0	0	0	150	0	0	0	150	0,00%
	Tinggi	K-mean	128	22	0	16	134	0	15	0	135	11,78%
		Model	150	0	0	0	150	0	0	0	150	0,00%

Lampiran 6. Hasil pengelompokan pada kondisi ketiga *cluster* saling tumpang tindih

a. Banyak data tiap *cluster* n=50

Jarak antar pusat *cluster*	Tingkat korelasi	Metode	*Cluster* 1			*Cluster* 2			*Cluster* 3			Persentasi salah pengelompokan
			50	**0**	**0**	**0**	**50**	**0**	**0**	**0**	**50**	
		Ward	22	11	17	19	20	11	21	11	18	60,00%
	Rendah	K-mean	18	9	23	10	19	21	8	18	24	59,33%
		Model	36	13	1	34	14	2	33	14	3	64,67%
		Ward	12	16	22	10	19	21	12	12	26	62,00%
Dekat	Sedang	K-mean	16	12	22	9	21	20	12	16	22	60,67%
		Model	12	18	20	10	19	21	14	18	18	67,33%
		Ward	18	23	9	15	26	9	20	16	14	61,33%
	Tinggi	K-mean	27	10	13	24	14	12	20	15	15	62,67%
		Model	4	43	3	0	45	5	2	12	36	43,33%
		Ward	39	5	6	18	10	22	11	15	24	51,33%
	Rendah	K-mean	24	9	17	9	34	7	15	11	24	45,33%
		Model	34	14	2	20	29	1	26	18	6	54,00%
		Ward	22	21	7	20	23	7	13	1	36	46,00%
Sedang	Sedang	K-mean	25	22	3	21	21	8	4	21	25	52,67%
		Model	2	24	24	4	26	20	3	22	25	64,67%
		Ward	21	20	9	20	30	0	9	15	26	48,67%
	Tinggi	K-mean	15	21	14	14	21	15	13	21	16	65,33%
		Model	44	4	2	2	45	3	4	3	43	12,00%
		Ward	23	18	9	8	35	7	15	3	32	40,00%
	Rendah	K-mean	29	16	5	20	19	11	4	17	29	48,67%
		Model	37	10	3	3	46	1	5	14	31	24,00%
		Ward	18	6	26	2	36	12	11	5	34	41,33%
Jauh	Sedang	K-mean	22	21	7	14	32	4	10	6	34	41,33%
		Model	36	5	9	5	39	6	4	5	41	22,67%
		Ward	38	8	4	1	32	17	1	32	17	42,00%
	Tinggi	K-mean	26	23	1	24	23	3	6	10	34	44,67%
		Model	48	2	0	1	49	0	1	1	48	3,33%

b. Banyak data tiap *cluster* n=100

Jarak antar pusat *cluster*	Tingkat korelasi	Metode	*Cluster* 1			*Cluster* 2			*Cluster* 3			Persentasi salah pengelompokan
			100	**0**	**0**	**0**	**100**	**0**	**0**	**0**	**100**	
		Ward	21	20	59	31	28	41	31	12	57	64,67%
	Rendah	K-mean	36	36	28	39	27	34	27	23	50	62,33%
		Model	44	1	55	39	1	60	36	1	63	64,00%
		Ward	15	34	51	21	43	36	9	32	59	61,00%
Dekat	Sedang	K-mean	28	43	29	27	45	28	26	44	30	65,67%
		Model	16	18	66	14	21	65	10	15	75	62,67%
		Ward	45	11	44	38	23	39	33	19	48	61,33%
	Tinggi	K-mean	31	23	46	31	27	42	31	21	48	64,67%
		Model	65	23	12	9	82	9	7	20	73	26,67%
		Ward	4	32	64	12	73	15	36	29	35	62,67%
	Rendah	K-mean	47	23	30	33	55	12	30	12	58	46,67%
		Model	34	35	31	15	28	57	6	21	73	55,00%
		Ward	68	22	10	68	27	5	31	27	42	54,33%
Sedang	Sedang	K-mean	37	14	49	35	56	9	38	12	50	52,33%
		Model	24	2	74	14	8	78	12	1	87	60,33%
		Ward	59	9	32	47	42	11	37	42	21	59,33%
	Tinggi	K-mean	31	34	35	21	48	31	26	42	32	63,00%

		Model	90	3	7	2	91	7	6	2	92	9,00%
		Ward	50	34	16	5	54	41	3	48	49	49,00%
	Rendah	K-mean	69	19	12	21	60	19	16	9	75	32,00%
		Model	96	3	1	32	55	13	32	5	63	28,67%
		Ward	47	22	31	0	50	50	6	46	48	51,67%
Jauh	Sedang	K-mean	68	31	1	28	42	30	10	24	66	41,33%
		Model	85	8	7	9	85	6	10	8	82	16,00%
		Ward	79	0	21	2	85	13	36	1	63	24,33%
	Tinggi	K-mean	41	57	2	35	61	4	37	0	63	45,00%
		Model	100	0	0	0	100	0	0	0	100	0,00%

c. Banyak data tiap *cluster* n=150

Jarak antar pusat *cluster*	**Tingkat korelasi**	**Metode**	***Cluster* 1**			***Cluster* 2**			***Cluster* 3**			**Persentasi salah pengelompokan**
			150	**0**	**0**	**0**	**150**	**0**	**0**	**0**	**150**	
		Ward	28	52	70	44	59	47	15	37	98	58,89%
	Rendah	K-mean	51	53	46	51	75	24	61	20	69	56,67%
		Model	87	19	44	68	20	62	43	29	78	58,89%
		Ward	35	74	41	24	75	51	24	71	55	63,33%
Dekat	Sedang	K-mean	74	33	43	65	42	43	52	44	54	62,22%
		Model	93	37	20	77	46	27	77	45	28	62,89%
		Ward	63	26	61	32	37	81	33	26	91	57,56%
	Tinggi	K-mean	50	44	56	38	38	74	38	37	75	63,78%
		Model	34	17	99	3	23	124	2	18	130	58,44%
		Ward	77	29	44	17	93	40	14	46	90	42,22%
	Rendah	K-mean	76	53	21	47	75	28	17	47	86	47,33%
		Model	100	1	49	99	16	35	99	1	50	63,11%
		Ward	78	47	25	49	80	21	52	64	34	57,33%
Sedang	Sedang	K-mean	76	66	8	32	60	58	28	39	83	51,33%
		Model	123	9	18	28	100	22	22	8	120	23,78%
		Ward	40	41	69	34	62	54	47	25	78	60,00%
	Tinggi	K-mean	69	41	40	68	41	41	61	40	49	64,67%
		Model	136	5	9	7	139	4	8	3	139	8,00%
		Ward	82	40	28	7	117	26	8	26	116	30,00%
	Rendah	K-mean	95	32	23	24	96	30	9	24	117	31,56%
		Model	123	4	23	35	66	49	16	3	131	28,89%
		Ward	80	4	66	76	62	12	65	0	85	49,56%
Jauh	Sedang	K-mean	56	24	70	51	97	2	46	11	93	45,33%
		Model	125	13	12	8	136	6	14	12	124	14,44%
		Ward	103	44	3	56	91	3	33	3	114	31,56%
	Tinggi	K-mean	68	2	80	37	95	18	69	0	81	45,78%
		Model	147	2	1	1	148	1	1	2	147	1,78%

Lampiran 7. Data *Iris*

Spesies Iris	Panjang Sepal	Lebar Sepal	Panjang Petal	Lebar Petal
setosa	5.1	3.5	1.4	0.2
setosa	4.9	3.0	1.4	0.2
setosa	4.7	3.2	1.3	0.2
setosa	4.6	3.1	1.5	0.2
setosa	5.0	3.6	1.4	0.2
setosa	5.4	3.9	1.7	0.4
setosa	4.6	3.4	1.4	0.3
setosa	5.0	3.4	1.5	0.2
setosa	4.4	2.9	1.4	0.2
setosa	4.9	3.1	1.5	0.1
setosa	5.4	3.7	1.5	0.2
setosa	4.8	3.4	1.6	0.2
setosa	4.8	3.0	1.4	0.1
setosa	4.3	3.0	1.1	0.1
setosa	5.8	4.0	1.2	0.2
setosa	5.7	4.4	1.5	0.4
setosa	5.4	3.9	1.3	0.4
setosa	5.1	3.5	1.4	0.3
setosa	5.7	3.8	1.7	0.3
setosa	5.1	3.8	1.5	0.3
setosa	5.4	3.4	1.7	0.2
setosa	5.1	3.7	1.5	0.4
setosa	4.6	3.6	1.0	0.2
setosa	5.1	3.3	1.7	0.5
setosa	4.8	3.4	1.9	0.2
setosa	5.0	3.0	1.6	0.2
setosa	5.0	3.4	1.6	0.4
setosa	5.2	3.5	1.5	0.2
setosa	5.2	3.4	1.4	0.2
setosa	4.7	3.2	1.6	0.2
setosa	4.8	3.1	1.6	0.2
setosa	5.4	3.4	1.5	0.4
setosa	5.2	4.1	1.5	0.1
setosa	5.5	4.2	1.4	0.2
setosa	4.9	3.1	1.5	0.2
setosa	5.0	3.2	1.2	0.2
setosa	5.5	3.5	1.3	0.2
setosa	4.9	3.6	1.4	0.1
setosa	4.4	3.0	1.3	0.2
setosa	5.1	3.4	1.5	0.2
setosa	5.0	3.5	1.3	0.3
setosa	4.5	2.3	1.3	0.3
setosa	4.4	3.2	1.3	0.2
setosa	5.0	3.5	1.6	0.6
setosa	5.1	3.8	1.9	0.4
setosa	4.8	3.0	1.4	0.3
Setosa	5.1	3.8	1.6	0.2
setosa	4.6	3.2	1.4	0.2
setosa	5.3	3.7	1.5	0.2
setosa	5.0	3.3	1.4	0.2
versicolor	7.0	3.2	4.7	1.4
versicolor	6.4	3.2	4.5	1.5

Spesies Iris	Panjang Sepal	Lebar Sepal	Panjang Petal	Lebar Petal
versicolor	6.9	3.1	4.9	1.5
versicolor	5.5	2.3	4.0	1.3
versicolor	6.5	2.8	4.6	1.5
versicolor	5.7	2.8	4.5	1.3
versicolor	6.3	3.3	4.7	1.6
versicolor	4.9	2.4	3.3	1.0
versicolor	6.6	2.9	4.6	1.3
versicolor	5.2	2.7	3.9	1.4
versicolor	5.0	2.0	3.5	1.0
versicolor	5.9	3.0	4.2	1.5
versicolor	6.0	2.2	4.0	1.0
versicolor	6.1	2.9	4.7	1.4
versicolor	5.6	2.9	3.6	1.3
versicolor	6.7	3.1	4.4	1.4
versicolor	5.6	3.0	4.5	1.5
versicolor	5.8	2.7	4.1	1.0
versicolor	6.2	2.2	4.5	1.5
versicolor	5.6	2.5	3.9	1.1
versicolor	5.9	3.2	4.8	1.8
versicolor	6.1	2.8	4.0	1.3
versicolor	6.3	2.5	4.9	1.5
versicolor	6.1	2.8	4.7	1.2
versicolor	6.4	2.9	4.3	1.3
versicolor	6.6	3.0	4.4	1.4
versicolor	6.8	2.8	4.8	1.4
versicolor	6.7	3.0	5.0	1.7
versicolor	6.0	2.9	4.5	1.5
versicolor	5.7	2.6	3.5	1.0
versicolor	5.5	2.4	3.8	1.1
versicolor	5.5	2.4	3.7	1.0
versicolor	5.8	2.7	3.9	1.2
versicolor	6.0	2.7	5.1	1.6
versicolor	5.4	3.0	4.5	1.5
versicolor	6.0	3.4	4.5	1.6
versicolor	6.7	3.1	4.7	1.5
versicolor	6.3	2.3	4.4	1.3
versicolor	5.6	3.0	4.1	1.3
versicolor	5.5	2.5	4.0	1.3
versicolor	5.5	2.6	4.4	1.2
versicolor	6.1	3.0	4.6	1.4
versicolor	5.8	2.6	4.0	1.2
versicolor	5.0	2.3	3.3	1.0
Versicolor	5.6	2.7	4.2	1.3
versicolor	5.7	3.0	4.2	1.2
versicolor	5.7	2.9	4.2	1.3
versicolor	6.2	2.9	4.3	1.3
versicolor	5.1	2.5	3.0	1.1
versicolor	5.7	2.8	4.1	1.3
virginica	6.3	3.3	6.0	2.5
virginica	5.8	2.7	5.1	1.9
virginica	7.1	3.0	5.9	2.1
virginica	6.3	2.9	5.6	1.8

Spesies Iris	Panjang Sepal	Lebar Sepal	Panjang Petal	Lebar Petal
virginica	6.5	3.0	5.8	2.2
virginica	7.6	3.0	6.6	2.1
virginica	4.9	2.5	4.5	1.7
virginica	7.3	2.9	6.3	1.8
virginica	6.7	2.5	5.8	1.8
virginica	7.2	3.6	6.1	2.5
virginica	6.5	3.2	5.1	2.0
virginica	6.4	2.7	5.3	1.9
virginica	6.8	3.0	5.5	2.1
virginica	5.7	2.5	5.0	2.0
virginica	5.8	2.8	5.1	2.4
virginica	6.4	3.2	5.3	2.3
virginica	6.5	3.0	5.5	1.8
virginica	7.7	3.8	6.7	2.2
virginica	7.7	2.6	6.9	2.3
virginica	6.0	2.2	5.0	1.5
virginica	6.9	3.2	5.7	2.3
virginica	5.6	2.8	4.9	2.0
virginica	7.7	2.8	6.7	2.0
virginica	6.3	2.7	4.9	1.8
virginica	6.7	3.3	5.7	2.1
virginica	7.2	3.2	6.0	1.8
virginica	6.2	2.8	4.8	1.8
virginica	6.1	3.0	4.9	1.8

Spesies Iris	Panjang Sepal	Lebar Sepal	Panjang Petal	Lebar Petal
virginica	6.4	2.8	5.6	2.1
virginica	7.2	3.0	5.8	1.6
virginica	7.4	2.8	6.1	1.9
virginica	7.9	3.8	6.4	2.0
virginica	6.4	2.8	5.6	2.2
virginica	6.3	2.8	5.1	1.5
virginica	6.1	2.6	5.6	1.4
virginica	7.7	3.0	6.1	2.3
virginica	6.3	3.4	5.6	2.4
virginica	6.4	3.1	5.5	1.8
virginica	6.0	3.0	4.8	1.8
virginica	6.9	3.1	5.4	2.1
virginica	6.7	3.1	5.6	2.4
virginica	6.9	3.1	5.1	2.3
virginica	5.8	2.7	5.1	1.9
virginica	6.8	3.2	5.9	2.3
virginica	6.7	3.3	5.7	2.5
virginica	6.7	3.0	5.2	2.3
virginica	6.3	2.5	5.0	1.9
virginica	6.5	3.0	5.2	2.0
virginica	6.2	3.4	5.4	2.3
virginica	5.9	3.0	5.1	1.8