PERBANDINGAN ANTARA METODE AGGLOMERATIF, METODE DIVISIF DAN METODE K-MEANS DALAM ANALISIS KLASTER

Dewi Rachmatin[1] dan Kania Sawitri[2]

Jurusan Pendidikan Matematika FPMIPA UPI
dewirachmatin@upi.edu
Jurusan Teknik Elektro ITENAS
kania@itenas.ac.id

Abstrak. Proses pengelompokan data dalam analisis klaster dapat dilakukan dengan dua metode yaitu: metode hirarki dan metode non-hirarki. Metode hirarki (*hierarchical methods)* adalah metode pengklasteran yang membentuk konstruksi hirarki berdasarkan tingkatan tertentu seperti struktur pohon. Metode ini dibagi menjadi dua yaitu metode agglomeratif (pemusatan) dan metode divisif (penyebaran). Metode-metode yang termasuk metode agglomeratif di antaranya : *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method dan Median Method.* Keenam metode agglomeratif (hirarki) tersebut telah dibahas pada penelitian terdahulu (Rachmatin, 2012). Dalam metode hirarki jumlah kelompok yang akan diperoleh belum diketahui, sedangkan metode non-hirarki memulai dengan mengasumsikan ada k kelompok terlebih dahulu. Metode yang termasuk metode non-hirarki adalah metode k-means dan metode fuzzy *k*-means. Pada artikel ini dibahas hasil kajian teoritis yang merupakan perbandingan metode hirarki (metode agglomeratif dan metode divisif) dengan metode non-hirarki yang diwakili oleh metode *k*-means. Hasil kajian teoritis tersebut diterapkan pada sebuah data yaitu data tingkat polusi udara (Gunawan dkk., 2010; Rachmatin, 2012).

***Kata Kunci** : Metode Agglomeratif, Metode Divisif, dan Metode K-Means.*

1. PENDAHULUAN

Dalam perkembangannya analisis klaster telah dipergunakan dalam berbagai disiplin ilmu seperti biologi, ekonomi, psikologi, pemasaran dan lain-lain. Sebagai contoh dalam bidang pemasaran, analisis klaster bertujuan untuk membuat segmen pasar dalam mengelompokkan jumlah pembeli berdasarkan keuntungan pembelian barang, memahami perilaku pembeli dalam mengelompokkan tempat belanja, mengenali peluang produk baru dalam mengelompokkan merek suatu produk, memilih uji pasar dalam mengelompokkan jenis kota [8].

Dillon dan Goldstein [2] mengatakan bahwa analisis klaster adalah statistik multivariat yang digunakan apabila ada *n* buah individu atau objek yang mempunyai *p* variabel dan ingin dikelompokkan ke dalam *k* klaster berdasarkan sifat-sifat yang diamati sehingga individu atau objek yang terletak dalam satu klaster memiliki kemiripan yang lebih besar dibandingkan dengan objek yang terletak dalam klaster lain.

Prinsip dasar dalam analisis klaster adalah mengelompokkan objek (observasi) pada suatu klaster yang memiliki kemiripan sangat besar dengan objek lain dalam klaster yang sama (*similarity*), tetapi sangat tidak mirip dengan objek lain pada klaster yang berbeda (*dissimilarity*). Hal ini berarti bahwa klaster yang baik akan mempunyai homogenitas (kesamaan) yang tinggi antar anggota dalam satu klaster (*within-cluster*) dan heterogenitas (perbedaan) yang tinggi antar klaster yang satu dengan yang lainnya (*between-cluster*).

Sebelum melakukan proses analisis klaster, hendaknya dilakukan pengujian asumsi terlebih dahulu. Asumsi-asumsi dalam analisis klaster yaitu data bebas dari *outliers* (pencilan) dan multikolinieritas. Pencilan merupakan suatu data observasi yang menyimpang dari sekumpulan data yang lain. Adanya pencilan dapat mengubah struktur sebenarnya dari populasi sehingga klaster-klaster yang terbentuk tidak representatif. Ini berarti bahwa klaster tersebut tidak mencerminkan karakteristik populasi yang sebenarnya. Sedangkan multikolinieritas berarti terdapat hubungan linear di antara beberapa atau semua variabel. Oleh karena itu, variabel-variabel yang bersifat multikolinieritas dalam analisis klaster perlu dipertimbangkan secara seksama.

Proses pengelompokan data dalam analisis klaster dapat dilakukan dengan dua metode yaitu: metode hirarki dan metode non-hirarki. Metode hirarki (*hierarchical methods)* adalah metode pengklasteran yang membentuk konstruksi hirarki berdasarkan tingkatan tertentu seperti struktur pohon. Metode ini dibagi menjadi dua yaitu metode *agglomeratif* (agglomeratif/pemusatan) dan metode *divisif* (divisif/penyebaran). Metode-metode yang termasuk metode agglomeratif di antaranya : *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method dan Median Method.* Keenam metode agglomeratif (hirarki) tersebut telah dibahas pada penelitian terdahulu [9]. Dalam metode hirarki jumlah kelompok yang akan diperoleh belum diketahui, sedangkan metode non-hirarki memulai dengan mengasumsikan ada k kelompok terlebih dahulu. Metode yang termasuk metode non-hirarki adalah metode k-means dan metode fuzzy k-means.

Pada artikel ini dibahas hasil kajian teoritis yang merupakan perbandingan metode hirarki (metode agglomeratif dan metode divisif) dengan metode non-hirarki yang diwakili oleh metode k-means. Hasil kajian teoritis tersebut diterapkan pada sebuah data yaitu data tingkat polusi udara pada sepuluh kota besar di negara Amerika Serikat (lihat [4] dan [9]).

2. METODE AGGLOMERATIF, METODE DIVISIF DAN METODE K-MEANS

Analisis klaster terdiri dari metode hirarki dan non-hirarki. Metode hirarki digunakan apabila belum ada informasi jumlah klaster yang akan dipilih. Metode hirarki ini secara umum dibedakan menjadi dua yaitu metode agglomeratif (penggabungan) dan metode divisif (pemecahan). Metode-metode yang termasuk metode agglomeratif di antaranya : *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method dan Median Method* [3]. Keenam metode agglomeratif dan metode divisif ini telah dibahas pada penelitian terdahulu [9].

Metode non-hirarki bertujuan untuk mengelompokkan n objek ke dalam k klaster ($k < n$), di mana nilai k telah ditentukan sebelumnya. Metode yang termasuk metode non-hirarki adalah k-means method [8] dan *fuzzy method.*

Metode divisif merupakan kebalikan dari metode agglomeratif dalam analisis klaster. Pada setiap langkahnya dalam metode divisif terjadi penambahan kelompok ke dalam nilai dua nilai terkecil sampai akhirnya semua elemen bergabung. Metode divisif merupakan proses pengklasteran yang didasarkan pada persamaan nilai rata-rata antar objek. Ini berarti bahwa klaster hirarki dibangun dalam $n-1$ langkah ketika data mengandung n objek (lihat [4]dan [9]).
Metode k-means diperkenalkan oleh James B MacQueen (1967), dasar pengelompokan dalam metode ini adalah menempatkan objek berdasarkan rata-rata klaster terdekat [8]. Oleh karena itu, metode ini bertujuan untuk meminimumkan *error* akibat partisi n objek ke dalam k klaster. Kelemahan metode adalah jumlah klaster harus ditentukan sebelumnya dan tidak menjamin solusi klaster yang unik karena metode ini sulit mencapai global optimum.

Algoritma metode hirarki agglomeratif secara umum (lihat [1], [8], [9], dan [12] untuk mengelompokkan N objek adalah sebagai berikut :

1. Mulai dengan N klaster, setiap klaster mengandung unsur tunggal dan sebuah matriks simetris D = $\{d_{jl}\}$ adalah jarak Euclid dengan rumus :

$$d_{jl} = \left\{(\boldsymbol{x_l} - \boldsymbol{x_j})'(\boldsymbol{x_l} - \boldsymbol{x_j})\right\}^{\frac{1}{2}} = \sqrt{\sum_{k=1}^{i}(x_{lk} - x_{jk})^2}$$

   $i = 1, 2, \ldots, p, atau\ l = 1,2, \ldots, n.$

2. Tentukan jarak untuk pasangan klaster yang terdekat. Misalkan jarak antara klaster U dan V adalah d_{UV}.
3. Gabungkan klaster U dan V. Tandai klaster baru yang terbentuk dengan (UV). Hitung kembali matriks jarak baru dengan cara :
   i. Hapus baris dan kolom yang bersesuaian dengan klaster U dan V.
   ii. Tambahkan baris dan kolom yang memberikan jarak-jarak antara klaster (UV) dan klaster-klaster yang tersisa.
4. Ulangi langkah 2 sebanyak (N-1) kali, sampai semua objek akan berada dalam klaster tunggal.

Algoritma metode-metode agglomeratif, yaitu *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method* dan *Median Method* dapat dilihat pada [1], [9], dan [12].

Hasil dari metode agglomeratif dapat ditampilkan dalam bentuk diagram yang disebut dendogram [6]. Dendrogram menggambarkan proses pembentukan klaster yang dinyatakan dalam bentuk gambar. Garis mendatar di atas dendogram menunjukkan skala yang menggambarkan tingkat kemiripan, semakin kecil nilai skala menunjukkan semakin mirip individu tersebut.

Metode divisif atau *divisive analysis method* disingkat metode DIANA ([4] dan [9]) merupakan proses pengklasteran yang didasarkan pada persamaan nilai rata-rata antar objek. Jika sebuah objek memiliki persamaan nilai rata-rata terbesar maka objek tersebut akan terpisah dan berubah menjadi *splinter group* [7]. Pada metode divisif ini perhitungan dilihat dari perbedaan atau selisih antara persamaan nilai rata-rata dengan nilai elemen matriks yang telah menjadi *splinter group*. Jika selisih nilai antara persamaan nilai rata-rata dengan nilai elemen matriks *splinter group* bernilai negatif, maka perhitungan terhenti sehingga harus dibuat matriks baru untuk mendapatkan klaster yang lain. Perhitungan ini terus dilakukan sedemikian sehingga semua objek terpisah [7]. Untuk memahami metode ini, selanjutnya akan diberikan algoritmanya.

Algoritma Metode DIANA (lihat [4] dan [9]) :

1. Bentuk suatu matriks jarak dengan menggunakan jarak Euclid. Asumsikan setiap data dianggap sebagai klaster, sehingga diperoleh matriks jarak sebagai berikut :

$$D(1)_{nxn} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}.$$

2. Hitung nilai rata-rata setiap objek dengan objek lainnya.
3. Tentukan objek yang memiliki nilai rata-rata yang terbesar, objek yang memiliki nilai rata-rata yang terbesar akan dipisah dan berubah menjadi *splinter group.*
4. Hitung selisih nilai antara elemen matriks *splinter group* dengan nilai rata-rata setiap objek yang tersisa.

5. Tentukan objek yang memiliki nilai selisih terbesar antara elemen matriks *splinter group* dengan nilai rata-rata. Jika nilai selisih tersebut bernilai positif, maka objek yang memiliki nilai selisih terbesar bergabung dengan *splinter group.*
6. Ulangi langkah satu sampai lima sedemikian sehingga semua nilai selisih antara elemen matriks *splinter group* dengan nilai rata-rata bernilai negatif dan klaster terbagi menjadi dua klaster baru.

Perbedaan metode agglomeratif dengan divisif yang dapat dikaji dari kedua algoritmanya adalah kedua metode mengkonstruksi hirarki dalam arah yang berlawanan, sehingga memungkinkan diperoleh hasil yang berbeda [7]. Metode agglomeratif mulai dengan menganggap ada n klaster, kemudian pada setiap langkah dua klaster bergabung hingga akhirnya ada satu klaster. Sedangkan metode divisif mulai dengan menganggap setiap objek bersatu (hanya ada satu klaster), dan kemudian pada setiap langkah berikutnya sebuah klaster berpisah hingga akhirnya hanya ada satu klaster.

Pada metode k-means, pengelompokan dimulai dengan penentuan jumlah klaster (k) dan centroid awal yang dipilih secara acak. Centroid merupakan rata-rata observasi yang berada dalam satu klaster. Observasi dikelompokkan ke dalam klaster berdasarkan jarak minimum dari observasi ke centroid klaster tersebut. Centroid terus diupdate selama terjadi perubahan observasi dalam klaster. Proses ini berhenti setelah mencapai kriteria konvergensi yang ditandai dengan tidak adanya observasi yang berpindah lagi.

Algoritma *K-Means* dalam pembentukan klaster [8] :

1. Misalkan diberikan matriks data $X = \{x_{ij}\}$ berukuran $n \times p$ dengan $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$ dan asumsikan jumlah klaster awal K
2. Tentukan centroid.
3. Hitung jarak setiap objek ke setiap centroid dengan menggunakan jarak Euclid atau dapat ditulis sebagai berikut:
$$d(x_i, c_i) = \sqrt{(x_i - c_i)^2}$$
4. Setiap objek disusun ke centroid terdekat dan kumpulan objek tersebut akan membentuk klaster.
5. Tentukan centroid baru dari klaster yang baru terbentuk, di mana centroid baru itu diperoleh dari rata-rata setiap objek yang terletak pada klaster yang sama.
6. Ulangi langkah 3, jika centroid awal dan baru tidak sama.

Dalam pemilihan jumlah klaster awal, menurut Rencher [10] dapat ditentukan melalui pendekatan salah satu metode hirarki. Oleh karena itu k (jumlah klaster awal) dapat ditentukan terlebih dahulu dengan metode hirarki, kemudian dilanjutkan pengelompokan objek-objek dengan metode partitioning yaitu dengan metode k-means. Walaupun terkadang jumlah klaster yang ditentukan tergantung pada subjektifitas seseorang [8].

Dalam melakukan setiap proses analisis klaster perlu dilakukan pengujian atas kevalidan atau kesahihan suatu hasil analisis. Terdapat beberapa cara yang dapat dilakukan untuk menguji kesahihan penentuan jumlah klaster (lihat [4] dan [9]), yaitu :

1. *Internal Test, internal test* merupakan salah satu cara menguji kesahihan penentuan jumlah klaster dengan cara membandingkan hasil klaster yang terbentuk dari masing-masing metode. Dalam hal ini, perbandingan hasil antara metode agglomeratif (dari hasil penelitian sebelumnya) dan metode divisif dapat dibandingkan.
2. Solusi Klaster (*Cluster Solution*), pada validasi penentuan jumlah klaster hasil analisis klaster dengan menggunakan *Cluster Solution* digunakan beberapa statistik, yaitu RMSSTD, SPR, RS dan CD (lihat [1], [8], [9], dan [12]).

3. HASIL PENERAPAN METODE AGGLOMERATIF, METODE DIVISIF DAN METODE K-MEANS

Untuk memberikan ilustrasi hasil penerapan kajian teoritis metode agglomeratif, metode divisif, dan metode k-means, berikut ini diberikan contoh penerapannya pada data tingkat polusi udara di beberapa kota di Amerika Serikat (lihat [4] dan [9]). Pada dasarnya perhitungan yang banyak, tingkat kesalahannya akan besar karena ketidaktelitian dalam perhitungan. Oleh karena itu akan diambil sepuluh data observasi untuk diolah sebagai contoh untuk menerapkan masing-masing algoritma dari masing-masing metode, dengan tujuh variabel yang terlibat adalah sebagai berikut : X_1 = Udara yang berisi SO_2 (mg/m^2); X_2 = Rata-rata suhu (^{o}F/tahun); X_3 = Jumlah pabrik yang mempekerjakan lebih dari 20 pekerja; X_4 = Jumlah penduduk hasil sensus 1970 dalam ribuan orang; X_5 = Rata-rata kecepatan angin (mil/jam); X_6 = Rata-rata curah hujan (inci); X_7 = Rata-rata jumlah hari dengan curah hujan (per tahun).

Tabel 1. Sepuluh Data Tingkat Polusi Udara Di Kota Amerika Serikat

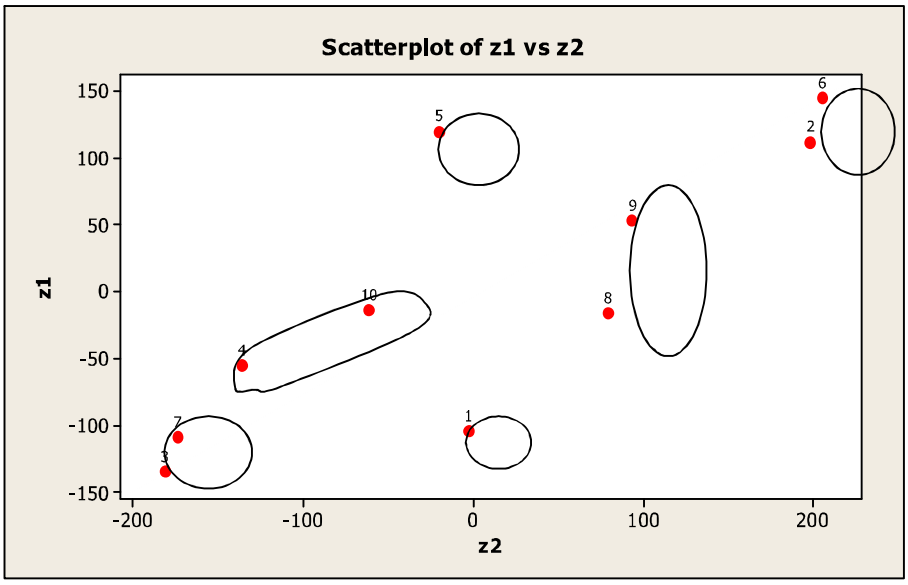
No.	Kota	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	Phoenix	10	70,3	213	582	6	7,05	36
2	Little Rock	13	61	91	132	8,2	48,52	100
3	San Francisco	12	56,7	453	716	8,7	20,66	67
4	Denver	17	51,9	454	515	9	12,95	86
5	Hartford	56	49,1	412	158	9	43,37	127
6	Wilmington	36	54	80	80	9	40,25	114
7	Washington	29	57,3	434	757	9,3	38,89	111
8	Jacksonville	14	68,4	136	529	8,8	54,47	116
9	Miami	10	75,5	207	335	9	59,8	128
10	Atlanta	24	61,5	368	497	9,1	48,34	115

Dari tabel korelasi (lihat [4] dan [9]) diketahui bahwa data mengandung kolinearitas. Korelasi antara variabel X_5 dengan variabel X_7 sebesar 0,803022 dan antara variabel X_6 dengan variabel X_7 sebesar 0,874897. Korelasi antara variabel tersebut dikatakan sangat kuat. Karena data mengandung korelasi, maka dilakukan transformasi dari data awal menjadi *z-score* dengan menggunakan program S-PLUS 2000 diperoleh data *z-score* pada tabel 2.

Tabel 2. *Z-score* sepuluh data

No.	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7
1	-104,377	-3,01824	49,6479	0,288578	140,1948	-10,9209	18,79304
2	112,2992	198,2421	-233,691	-116,01	-62,4245	-34,5032	39,4952
3	-134,097	-180,97	214,6247	88,5078	69,04649	40,79587	-21,351
4	-55,254	-136,186	94,86769	46,38724	-57,8049	23,48428	-9,44223
5	120,1657	-20,1308	-127,265	-7,47223	-245,987	6,672011	24,74572
6	145,3963	205,0331	-277,61	-118,567	-81,0943	-50,0775	39,56648
7	-108,793	-173,975	242,0781	112,3141	110,4211	35,29677	-50,4768
8	-15,5218	78,96966	23,65385	-11,5695	157,0643	-16,4857	-22,0113
9	53,42911	93,40841	-59,5932	-34,0326	-16,5095	-13,5163	-2,47179
10	-13,2472	-61,3731	73,28669	40,15393	-12,9069	19,25473	-16,8473

Data pada tabel 2 merupakan data *z-score* yang diperoleh dari perhitungan dengan menggunakan rumus : $z_i = u_i^T[x - \bar{x}]$ dengan u_i adalah vektor eigen ke i yang diperoleh dari hasil analisis komponen utama [4]. Untuk melihat ada tidaknya pengelompokan dari data tersebut, maka sepuluh data *z-score* diplot di ruang berdimensi dua dengan bantuan software Minitab 14. Hal ini disebabkan dari hasil analisis komponen utama yaitu dari *Scree-Plot Test* terdapat break (patahan) antara komponen utama kedua dan ketiga, sehingga dapat digunakan ruang berdimensi dua untuk merepresentasikan data [5]. Jika data diplot dengan diagram pencar z_1 *vs* z_2 , z_5 *vs* z_6 (lihat [4] dan [9]) maka dapat diprediksi kemungkinan akan terjadi 5 atau 6 kluster, seperti diperlihatkan pada gambar 1.



Gambar 1. *Scatterplot* z_1 dengan z_2 untuk melihat ada tidaknya pengelompokan

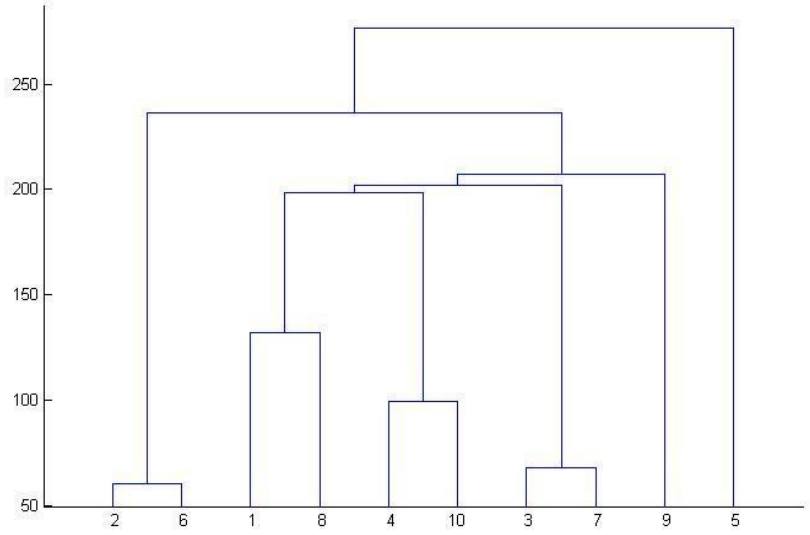
Hasil penerapan metode agglomeratif dan *k*-means yang dikerjakan dengan program SAS versi 6.1.2 dan program MATLAB R2010a serta metode divisif yang dikerjakan secara manual. Metode ini masih dikerjakan secara manual dikarenakan belum ada pada kedua program, hasilnya ditunjukkan pada tabel 3 dan tabel 4. Sedangkan dendrogram yang dihasilkan dari metode agglomeratif diperlihatkan pada gambar 2.

Tabel 3. Hasil Pengklasteran dengan Metode Agglomeratif, Divisif dan K-Means untuk k=5

Klaster	Anggota klaster	Kota
1	2, 5 , 6, 9	Willmington, Little Rock, Harrford, Miami
2	1, 8	Jacksonville, Phoenix
3	4, 10	Atlanta, Denver
4	3	San Fransisco
5	7	Washington

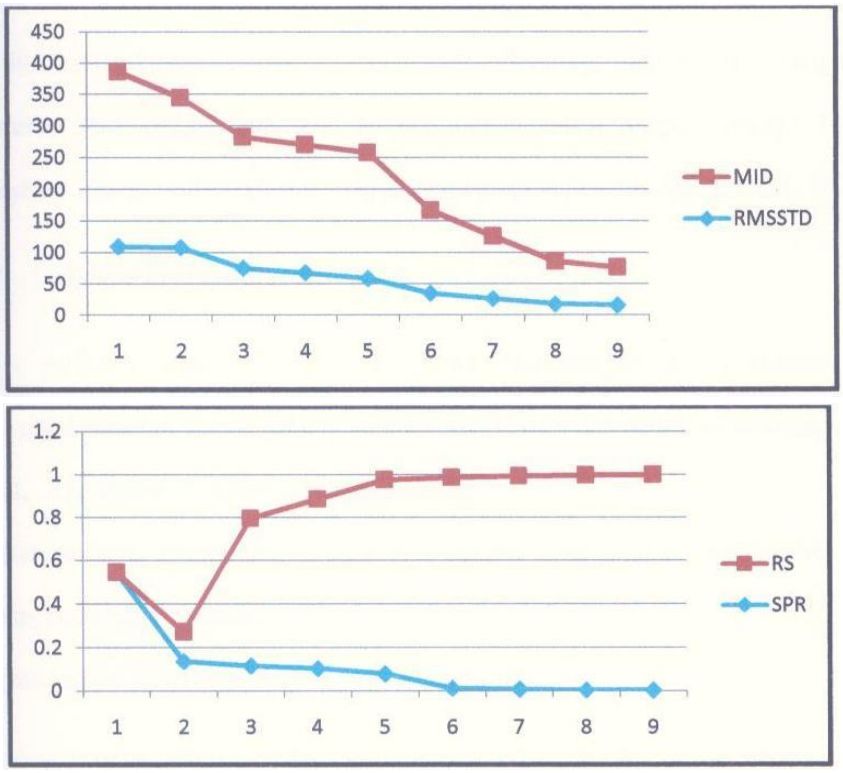
Tabel 4. Hasil Pengklasteran Metode Agglomeratif, Divisif dan K-Means untuk k=6

Klaster	Anggota klaster	Kota
1	4, 10	Atlanta, Denver
2	1, 8	Jacksonville, Phoenix
3	3, 7	San Fransisco, Washington
4	9	Miami
5	2, 6	Willmington, Harrford
6	5	Little Rock



Gambar 2. Dendrogram yang Dihasilkan dari Metode Agglomeratif

Hasil validasi dengan program SAS versi 6.1.2 untuk pengklasteran dengan salah satu metode agglomeratif, yaitu metode Single Linkage ditunjukkan pada gambar 3, sedangkan hasil-hasil kelima metode agglomeratif yang lain dan metode k-means mempunyai kemiripan dengan hasil metode Single Linkage ini, lihat [1] dan [12].



Gambar 3. Plot RMSSTD dan MID (atas), Plot RS dan SPR (bawah) untuk Metode Single Lingkage

Perbedaan yang cukup jelas terlihat pada gambar 3, mulai jumlah klaster 5 atau 6 terdapat perbedaan pada plot titik-titiknya baik plot RMSSTD dan MID maupun plot RS dan SPR. Mulai jumlah klaster 5 nilai MID besar dan nilai RMSSTD kecil (gambar 3 atas), demikian pula nilai SPR kecil dan nilai RS besar (gambar 3 bawah), sehingga dapat disimpulkan dari hasil plot ini jumlah klaster yang valid dapat dipilih 5 atau 6 klaster. Jadi jumlah klaster sebanyak 5 atau 6 valid atau dapat dipercaya.

Untuk penamaan masing-masing klaster yang diperoleh dapat berdasarkan karakteristik masing-masing klaster digunakan statistik rata-rata [3].

4. KESIMPULAN

Hasil pengelompokan untuk keenam metode agglomeratif, divisif dan metode *k*-means pada contoh data sampel : sepuluh data observasi tingkat polusi udara di Amerika Serikat memberikan hasil 5 atau 6 jumlah klaster. Hal ini didukung dengan validasi dengan plot RMSSTD dan CD beserta plot RS dan SPR untuk masing-masing metode aggolmeratif (lihat [1], [8], [9], dan [12]). Pengelompokan kesepuluh kota ditunjukkan pada tabel 3 dan 4, untuk jumlah klaster (k) = 5 atau 6.

Perbedaan metode agglomeratif dengan divisif dari algoritmanya adalah kedua metode mengkonstruksi hirarki dalam arah yang berlawanan, sehingga memungkinkan diperoleh hasil yang berbeda [7]. Metode agglomeratif mulai dengan menganggap ada *n* klaster, kemudian pada setiap langkah dua klaster bergabung hingga akhirnya ada satu klaster. Sedangkan metode divisif mulai dengan menganggap setiap objek bersatu (hanya ada satu klaster), dan kemudian pada setiap langkah berikutnya sebuah klaster berpisah hingga akhirnya hanya ada satu klaster.

Oleh karena metode hirarki (agglomeratif dan divisif) tidak efisien jika digunakan dalam mengelompokkan observasi dalam jumlah besar, maka untuk mengatasi masalah ukuran sampel yang sangat besar dianjurkan agar pengelompokan menggunakan metode *k*-means. Akan tetapi disamping kelebihan tersebut, metode k-means memiliki kelemahan terutama solusi klaster yang hanya mencapai lokal optimum daripada global optimum. Hal ini sangat dipengaruhi oleh pemilihan centroid secara acak [8].

DAFTAR PUSTAKA

[1] Asumpta, Rachmatin, dan Suherman. (2010). *Centroid Method dan Median Method Pada Analisis Klaster*, Jurusan Pendidikan Matematika FPMIPA UPI, Bandung.

[2] Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis : Methods and Applications*. John Wiley & Sons Inc, New Jersey.

[3] Everitt, B. (1974). *Cluster Analysis*, Social Science Research Council.

[4] Gunawan, Rachmatin, dan Suherman. (2010). *Pengklasteran Data dengan Menggunakan Divisive Analysis Method (DIANA),* Jurusan Pendidikan Matematika FPMIPA UPI, Bandung.

[5] Jackson, J. E.(1991). *A User's Guide To Principal Components,* John Wiley & Sons Inc, New Jersey.

[6] Johnson, R. A. and Wincern, D. W. (1982). *Applied Multivariate Statistical Analysis*, Prentice Hal Inc, New Jersey.

[7] Kaufman, et. all. (2005). *Finding Groups in Data.* John Wiley&Sons, New Jersey.

[8] Nuningsih, Rachmatin, dan Suherman. (2010). *K-Means Clustering*, Jurusan Pendidikan Matematika FPMIPA UPI, Bandung.

[9] Rachmatin, D. (2012). *Aplikasi Metode Agglomeratif Dalam Analisis Klaster Pada Data Tingkat Polusi Udara*, Jurusan Pendidikan Matematika FPMIPA UPI, Bandung.

[10] Rencher, A. C. (2002). *Method of Multivariate*. John Wiley&Sons, Canada.

[11] Sharma, S.(1996). *Applied Multivariate Technique*, John Wiley&Sons, Canada.

[12] Sofyana, Rachmatin, dan Suherman. (2010). *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method Pada Analisis Klaster*, Jurusan Pendidikan Matematika FPMIPA UPI, Bandung.