APLIKASI METODE-METODE *AGGLOMERATIVE* DALAM ANALISIS KLASTER PADA DATA TINGKAT POLUSI UDARA

Oleh:

Dewi Rachmatin

Jurusan Pendidikan Matematika, UniversitasPendidikan Indonesia
dewirachmatin@upi.edu

ABSTRAK

Analisis Klaster merupakan analisis pengelompokkan data yang mengelompokkan data berdasarkan informasi yang ditemukan pada data. Tujuan dari analisis klaster adalah agar objek-objek di dalam satu kelompok memiliki kesamaan satu sama lain sedangkan dengan objek-objek yang berbeda kelompok memiliki perbedaan. Analisis klaster dibagi menjadi dua metode yaitu metode hirarki dan metode non-hirarki. Metode hirarki dibagi menjadi dua, yaitu metode *agglomerative* (pemusatan) dan metode *divisive* (penyebaran). Metode-metode yang termasuk dalam metode *agglomerative* adalah *Single Linkage Method*, *Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method* dan *Median Method*. Pada artikel ini dibahas metode-metode *agglomerative* tersebut yang diterapkan pada data tingkat polusi udara. Masing-masing metode tersebut memberikan jumlah klaster yang berbeda.

Kata Kunci : Analisis Klaster, *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method* dan *Median Method*.

ABSTRACT

Cluster analysis is an analysis of the data classification based on information found in the data.The objective of cluster analysis is that the objects in the group have in common with each other, while the different objects have different groups. Cluster analysis is divided into two methods : the method of non-hierarchical and hierarchical methods.Hierarchical method is divided into two methods, namely agglomerative methods (concentration) and divisive methods (deployment). The methods included in the agglomerative method is Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward 's Method, Method and Median Centroid Method. In this article discussed the agglomerative methods were applied to the data rate of air pollution. Each of these methods provides a different number of clusters.

Keywords: Cluster Analysis , Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward 's Method, Method and Median Centroid Method.

I. PENDAHULUAN

Dari sekian banyak metode statistika, analisis multivariat merupakan analisis yang cocok untuk meringkas data dengan peubah yang banyak. Beberapa analisis dalam analisis multivariat yang dapat digunakan untuk memahami dan mempermudah interpretasi data multivariat di antaranya adalah analisis klaster, analisis diskriminan, analisis komponen utama dan analisis faktor.Pada artikel ini analisis multivariat yang akan dibahas adalah analisis klaster.

Analisis klaster pertama kali digunakan oleh Tyron pada tahun 1939. Analisis klaster bertujuan untuk mengalokasikan sekelompok individu pada suatu kelompok-kelompok yang saling bebas sehingga individu-individu di dalam satu kelompok yang sama mirip satu sama lain, sedangkan individu-individu di dalam kelompok yang berbeda tidak mirip.Dalam pengelompokannya digunakan suatu ukuran yang dapat menerangkan keserupaan atau kedekatan antar data untuk menerangkan struktur grup sederhana dari data yang kompleks, yaitu ukuran jarak atau similaritas (lihat Johnson, 1982:538), dan ukuran jarak yang sering digunakan adalah ukuran jarak yang disebut jarak Euclid (Johnson,1982:534).

Saat ini analisis klaster telah banyak digunakan di berbagai bidang ilmu seperti biologi, kimia, ekonomi, psikologi, kesehatan, sosial dan berbagai bidang lainnya. Salah satu contoh, analisis klaster digunakan untuk mengelompokkan daerah-daerah berdasarkan bencana yang sering melanda daerah tersebut seperti banjir, gempa bumi, tsunami, dan bencana letusan gunung berapi.

Tidak seperti halnya dengan analisis multivariat yang lain (contohnya analisis diskriminan) memerlukan asumsi seperti normalitas, dalam analisis klaster asumsi yang harus diperhatikan adalah data bebas dari pencilan dan tidak ada kolinieritas. Dalam melakukan pemilihan objek ke dalam klaster-klaster (kelompok-kelompok), analisis klaster peka terhadap pencilan. Klaster-klaster yang diperoleh akan tidak sesuai dengan struktur yang sebenarnya dari populasi jika pencilan dilibatkan dalam pengolahan data. Sedangkan jika terdapat kolinieritas antar variable sebelum dilakukan analisis klaster, data awal terlebih dahulu ditransformasi melalui teknik komponen utama menjadi zscore.

Analisis klaster dibagi menjadi dua metode yaitu metode hirarki dan metode non-hirarki. Dalam metode hirarki jumlah kelompok yang akan diperoleh belum diketahui, sedangkan dalam metode nonhirarki diasumsikan ada k kelompok terlebih dahulu. Metode hirarki dibagi menjadi dua, yaitu metode *agglomerative* (pemusatan) dan metode *divisive* (penyebaran). Metode-metode yang termasuk dalam metode agglomerative adalah *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method* dan *Median Method*(Everitt, 1974:17).Sedangkan metode yang termasuk metode nonhirarki adalah metode k-means dan *fuzzymethod*.

Hasil dari metode *agglometarive* dapat ditampilkan dalam bentuk diagram yang disebut dendogram (Johnson, 1982:543). Dendogram 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.

Ada beberapa kelebihan dan kelemahan dari analisis klaster (Raharto, 2008:3), yaitu:

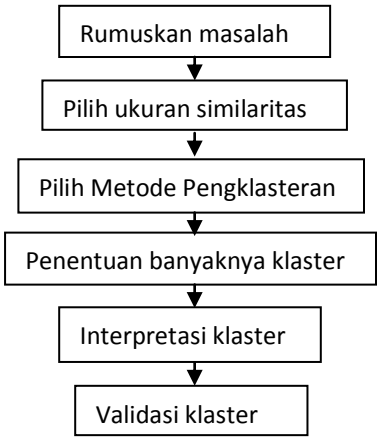
Kelebihan analisis klaster antara lain :
1. Dapat mengelompokkan data observasi dalam jumlah besar dan variabel yang relatif banyak, sedemikian sehingga data yang direduksi dengan kelompok akan mudah dianalisis.
2. Dapat dipakai dalam skala data ordinal, interval dan rasio.

Sedangkan kelemahan analisis klaster antara lain :
1. Pengelompokkan bersifat subjektifitas peneliti karena hanya melihat gambar dendogram.
2. Untuk data yang terlalu heterogen antara objek penelitian yang satu dengan yang lain akan sulit bagi peneliti untuk menentukan jumlah kelompok yang akan dibentuk.
3. Metode-metode yang dipakai memberikan perbedaan yang signifikan sehingga dalam perhitungan biasanya masing-masing metode dibandingkan.
4. Semakin besar observasi, biasanya tingkat kesalahan pengelompokkan akan semakin besar.

Dari kelemahan dan kelebihan analisis klaster tersebut, penulis memandang perlunya untuk mengangkat topik analisis klaster ini karena manfaatnya yang cukup besar, dan sangat jarang peneliti yang mengangkat topik analisis klaster ini dan membandingkan semua metode dalam analisis klaster terutama metode-metode hirarki. Pada bagian berikutnya akan dibahas metode-metode *agglomerative* beserta algoritmanya masing-masing agar dapat menjadi referensi bagi para pembaca yang awam mengenai analisis klaster.

II. ANALISIS KLASTER DAN METODE *AGGLOMERATIVE* DALAM ANALISIS KLASTER

Secara umum, tahapan-tahapan yang harus dilakukan pada analisis klaster atau proses analisis klaster adalah :



Hal terpenting dalam analisis klaster adalah menentukan jumlah klaster. Dalam menentukan banyaknya klaster yang akan terbentuk dari masing-masing metode dapat bergantung pada subjektifitas peneliti dengan hanya melihat dendogram. Hal ini berdampak pada solusi analisis klaster yang menjadi tidak unik.

Dalam melakukan proses analisis klaster, pengujian atas kevalidan atau kesahihan suatu hasil analisis klaster terdapat dua cara, yaitu Pertama *internal test*, suatu cara pengujian dengan membandingkan hasil klaster yang terbentuk dari beberapa metode berbeda yang digunakan; Kedua solusi klaster yang diajukan oleh Sharma (1996:198).

Tahapan validasi dalam analisis klaster yang dilakukan oleh Sharma untuk menguji apakah klaster yang terbentuk dari hasil subjektifitas peneliti telah valid atau tidak, uji validasi terhadap klaster yang terbentuk dilihat dari plot nilai RMSSTD (*Root Mean Square Standard Deviation*) dan nilai CD terhadap jumlah klaster, serta juga dapat dilihat dari plot nilai SPR, dan nilai RS terhadap jumlah klaster. Berikut penjelasan beberapa rumus yang yang terlibat dalam validasi jumlah klaster.

RMSSTD = $\sqrt{\frac{(n-1)\sum_{i=1}^{p} s_i^2}{p(n-1)}}$, di mana $S_i^2 = \frac{\sum_{i=1}^{n} X_{ij}^2}{n-1}$ "variansi untuk variabel ke i" dan X_{ij} adalah *mean corrected* untuk observasi ke i dan variabel ke j, n adalah banyaknya data, dan p adalah banyaknya variabel.

R-Square (RS) = $\frac{SS_B}{SS_T} = \frac{SS_T - SS_W}{SSCP_T}$; SS_w merupakan jumlah kuadrat dalam kelompok; SS_B merupakan jumlah kuadrat antar kelompok, dan SS_Tmerupakan total jumlah kuadrat.

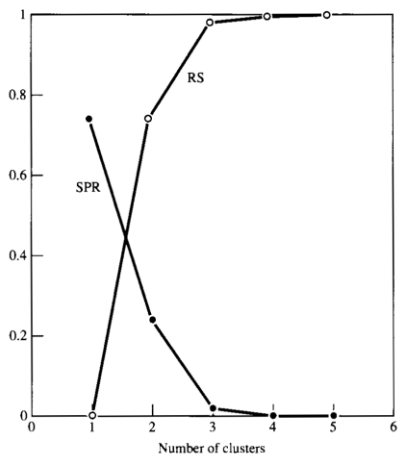
SPR (*Semipartial R-Squared*) = $\frac{SS_W - \sum SS_{W\,(klasteryangbergabung\quad)}}{SS_T}$

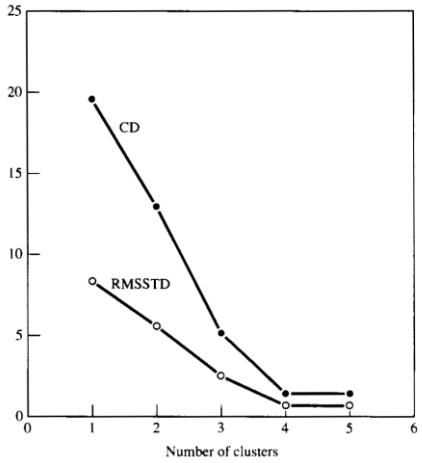
CD (*Distance between two cluster atau Cluster Distance*) merupakan jarak antar dua klaster.

Sharma memberikan konsep dari tentang penentuan solusi klaster yang diberikan pada Tabel 1 berikut ini :

Tabel 1. Statistik, Konsep Ukuran dan Syarat Homogenitas/Heterogenitas Klaster

Statistik	Konsep Ukuran	Syarat
RMSSTD	Homogenitas klaster baru	Nilai harus kecil
SPR	Homogenitas klaster yang bergabung	Nilai harus kecil
RS	Heterogenitas klaster	Nilai harus besar
CD	Homogenitas klaster yang bergabung	Nilai harus kecil





Gambar 1 Plot SPR dan RSGambar 2 Plot RMSSTD dan CD

Pada gambar 1 terlihat bahwa nilai RS meningkat (besar) artinya perbedaan antar klaster tinggi dan dan nilai SPR kecil artinya homogenitas dalam klaster tersebut adalah tinggi. Pada saat nilai RS tinggi dan nilai SPR kecil terjadi pada jumlah klaster 3. Sedangkan pada gambar 2 terlihat bahwa nilai RMSSTD dan CD menurun (kecil), ini berarti tingkat homogenitas antar klaster yang bergabung adalah tinggi. Pada jumlah klaster 4 kedua nilai RMSSTD dan CD ini sama-sama kecil. Jadi berdasarkan kedua gambar tersebut, jumlah klaster yang dapat dipilih adalah 3 atau 4.

Algoritma klaster hirarki *agglomerative* secara umum 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 :

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

$i = 1, 2, \dots, p, ataul = 1,2, \dots, n.$

(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.

Untuk setiap algoritma masing-masing metode *agglomerative* berikut diberikan input data sebagai berikut :

Misalkan diberikan matriks data X_{nxp} , di mana X_{ji} adalah data sampel observasi ke j (j=1,2,…,n) untuk variabel ke i (i=1,2,…,p). Selanjutnya akan diuraikan masing-masing metode *agglomerative* dan algoritma masing-masing metode tersebut.

1. Single Linkage Method

Single Linkage Method adalah proses pengklasteran yang didasarkan pada jarak terdekat antar objeknya. Jika dua objek terpisah oleh jarak yang pendek, maka kedua objek tersebut akan bergabung menjadi satu klaster dan demikian seterusnya. Untuk lebih memahami cara kerja metode ini perhatikan algoritma berikut ini :

(1) Bentuk matriks jarak Euclid untuk matriks data sampel yang diberikan, misalkan

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

(2) Asumsikan setiap data dianggap sebagai klaster, kemudian tentukan klaster yang mempunyai jarak terdekat, misal klaster U dan klaster V mempunyai jarak terdekat kemudian gabungkan, hasil gabungannya adalah klaster UV.
(3) Dari klaster UV yang telah terbentuk cari jarak minimum antar klaster UV dengan klaster (objek) lainnya yang belum bergabung, matriks jarak baru yang diperoleh sebut D(2). Misalkan $d_{(uv)w} = \min(d_{uw}, d_{vw})$, maka klaster yang baru terbentuk adalah (UVW).
(4) Ulangi langkah 2 sampai semua objek bergabung menjadi satu kelompok.

2. *Complete Linkage Method*

Complete Linkage Method adalah metode pengklasteran yang didasarkan jarak terjauh antar objek. Jika dua objek terpisah oleh jarak yang jauh, maka kedua objek tersebut akan digabung menjadi satu klaster, demikian seterusnya.

Langkah ketiga untuk algoritma metode ini berbeda dengan algoritma *Single Linkage Method*, pada langkah ketiga metode ini dari klaster UV yang terbentuk kemudian dicari jarak maksimum antar klaster UV dengan objek-objek berada di luar klaster UV, misalkan $d_{(uv)w}$ = max (d_{uw},d_{vw}). Dari langkah ketiga ini akan diperoleh matriks jarak baru D(2) dan selanjutnya ulangi langkah kedua sampai semua objek bergabung menjadi satu kelompok.

3. *Average Linkage Method*

Average Linkage Method adalah metode pengklasteran yang didasarkan pada jarak rata-rata antar objeknya. Langkah ketiga untuk algoritma metode ini berbeda dengan algoritma *Single Linkage Method*, pada langkah ketiga metode ini dari klaster UV yang terbentuk kemudian dicari jarak rata-rata antar klaster dengan objek lainnya yang belum bergabung, misalkan W. Namakan jarak rata-ratanya adalah

Pada langkah selanjutnya, dari hasil langkah ketiga diperoleh matriks jarak D(2), ditentukan jarak terdekat dari D(2). Objek yang mempunyai jarak terdekat bergabung dan membentuk klaster baru, selanjutnya ulangi langkah kedua sampai semua objek bergabung menjadi satu kelompok.

4. *Ward's Method*

Pada jarak antar dua klaster adalah total jumlah kuadrat dua klaster pada masing-masing variable. Metode ini berbeda dengan metode lainnya karena menggunakan pendekatan analisis varians untuk menghitung jarak antar klaster atau metode ini meminimumkan jumlah kuadrat (ESS). Untuk lebih memahami cara kerja metode ini perhatikan algoritma berikut :

(1) Asumsikan setiap data dianggap sebagai klaster.
(2) Bentuk klaster, di mana sebuah klaster terdiri dari pasangan dua objek sehingga kemungkinan banyaknya klaster C_2^n, kemudian hitung ESS dari semua pasangan klaster dengan rumus : $ESS = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2$.
(3) Pilih nilai ESS yang terkecil kemudian pasangan dari klaster tersebut gabungkan.
(4) Ulangi langkah 2 sampai membentuk satu klaster.

5. *Centroid Method*

Centroid Method atau Metode Centroid merupakan metode pengklasteran dengan memperhatikan rata-rata dari setiap objek yang bergabung berdasarkan jarak

minimum yang diperoleh dari matriks jarak Euclid. Algoritma Metode Centroid sebagai berikut :

(1) Asumsikan setiap data merupakan klaster.
(2) Bentuk matriks jarak dengan menggunakan kuadrat jarak Euclid :

$$D = d_{ij}^2 = (\boldsymbol{x_i} - \boldsymbol{x_j})^{\boldsymbol{T}}(\boldsymbol{x_i} - \boldsymbol{x_j}),$$ dengan $k = 1,2,\ldots,n$,

Sehingga matriks jaraknya adalah :

$$D_{nxn} = \begin{bmatrix} d_{11}^2 d_{12}^2 \ldots d_{1n}^2 \\ d_{21}^2 d_{22}^2 \ldots d_{2n}^2 \\ \vdots\vdots\ddots\vdots \\ d_{n1}^2 d_{n2}^2 \ldots d_{nn}^2 \end{bmatrix}$$

(3) Dari matriks jarak tersebut, pilih jarak terkecil antar klaster lalu gabungkan kedua objek yang memiliki jarak terkecil tersebut. Misalkan klaster U dan klaster V memiliki jarak terdekat, maka U dan V bergabung dalam satu klaster.
(4) Hitung centroid dari U dan V dengan rumus :

$$X_{(UV)} = \frac{(n_U \times \bar{x}_U) + (n_V \times \bar{x}_V)}{n_{UP} + n_{VP}}$$

(5) Bentuk matriks data baru dengan data dari klaster gabungan U dan V yang diperoleh dari langkah keempat.
(6) Ulangi langkah kedua, demikian seterusnya sampai semua data bergabung dengan jumlah klaster yang diinginkan.

6. *Median Method*

Median Method atau Metode Median merupakan metode pengklasteran dengan memperhatikan median dari setiap objek yang bergabung berdasarkan jarak minimum yang diperoleh dari matriks jarak Euclid. Langkah pertama sampai langkah ketiga metode centroid sama dengan metode median, sedangkan :

Langkah keempat : Hitung median dari klaster U dan V dengan menggunakan rumus:

$$m_{UV} = \frac{1}{2}(\bar{x}_U + \bar{x}_V).$$

Langkah kelima : Bentuk matriks data baru dengan data dari klaster gabungan U dan V yang telah diperoleh

Langkah keenam : Ulangi langkah kedua, demikian seterusnya sampai semua data bergabung dalam jumlah klaster yang diinginkan.

III. HASIL PENERAPAN METODE-METODE*AGGLOMERATIVE*

Metode-metode *agglomerative* yang telah dibahas akan diterapkan pada sepuluh data observasi (sepuluh kota) tingkat polusi udara di beberapa kota di Amerika Serikat dengan tujuh variabel sebagai berikut :

X_1 : udara yang berisi SO_2 (mg/m^2)
X_2 : rata-rata suhu (F/tahun)
X_3 : jumlah pabrik yang memperkerjakan lebih dari 20 pekerja
X_4 : jumlah penduduk hasil sensus tahun 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).

Pada dasarnya perhitungan yang banyak tingkat kesalahannya juga akan besar karena kekurangtelitian dalam perhitungan, oleh karena itu untuk contoh penerapan cukup diambil sepuluh data saja sebagai data kasus dalam penerapannya, terutama untuk Ward's method jika terlalu banyak observasi maka tidak memungkinkan seluruh pasangan kombinasi dari seluruh data untuk diujicobakan. Sepuluh data yang dipakai sebagai simulasi dapat dilihat pada Tabel 2 berikut :

Tabel 2. 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

Hasil pengujian bebas pencilan dan pengujian ada tidaknya kolinearitas dapat dilihat pada Tabel 3 dan Tabel 4.

Tabel 3 : Jarak Mahalanobis

Kota	**Jarak Malahalanobis**	$\chi^2_{0,0001;7}$
Phoenix	8,03253	29,8775
Little Rock	7,39328	29,8775
San Fransisco	7,45052	29,8775
Denver	7,88469	29,8775
Harrford	6,93115	29,8775
Wilmington	7,62554	29,8775
Washington	4,64769	29,8775
Jacksonville	4,18771	29,8775
Miami	7,35371	29,8775
Atlanta	1,49318	29,8775

Dapat dilihat pada Tabel 3,karena jarak Malahanobis $<\chi_{p,\alpha}^2 = \chi_{7,0,0001}^2$, hal ini berarti data bebas pencilan.

Tabel 3. Korelasi Variabel untuk Sepuluh Data

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	1	-0,69061	0,236997	-0,42893	0,394402	0,178098	0,495407
[2,]	-0,69061	1	-0,48424	0,176162	-0,4167	0,270984	-0,11312
[3,]	0,236997	-0,48424	1	0,574109	0,300566	-0,39774	-0,10883
[4,]	-0,42893	0,176162	0,574109	1	-0,10387	-0,42653	-0,45636
[5,]	0,394402	-0,4167	0,300566	-0,10387	1	0,550712	0,803022
[6,]	0,178098	0,270984	-0,39774	-0,42653	0,550712	1	0,874897
[7,]	0,495407	-0,11312	-0,10883	-0,45636	0,803022	0,874897	1

Dari Tabel 4 tersebut diketahui bahwa variabel X_5 dengan X_7 mempunyai korelasi yang cukup besar yaitu 0,803022, juga X_6 dengan X_7 mempunyai korelasi yang cukup besar yaitu 0,87897. Karena data mengandung korelasi maka dilakukan proses analisis komponen utama, yaitu dilakukan transformasi data awal menjadi z-score, sehingga selanjutnya data z-score ini yang digunakan dalam penerapan metode-metode *agglomerative*.

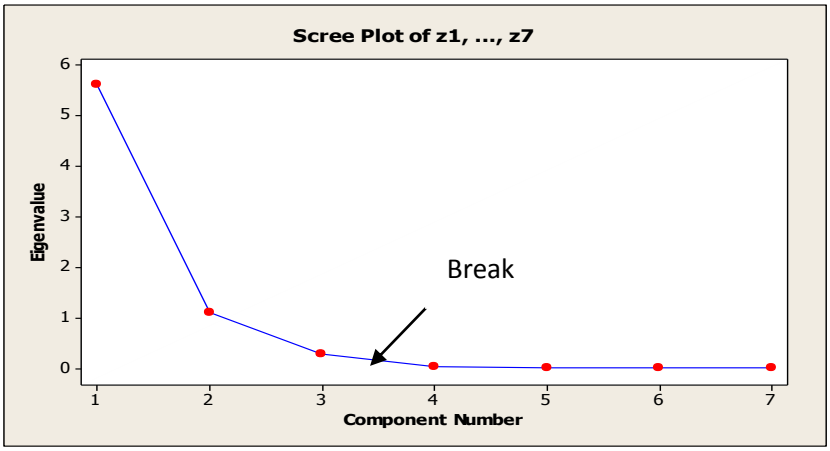
Z-score mentransformasikan p variabel $X_1,X_2,\ldots,X_p$ ke dalam p variable baru yang tidak berkorelasi yaitu $Z_1,Z_2,\ldots,Z_p$dengan rumus $z_i = u_i^T[x - \bar{x}]$ dengan u_i adalah vektor eigen ke i yang diperoleh dari analisis komponen utama (Jackson, 1991:11).

Tabel 5 : *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

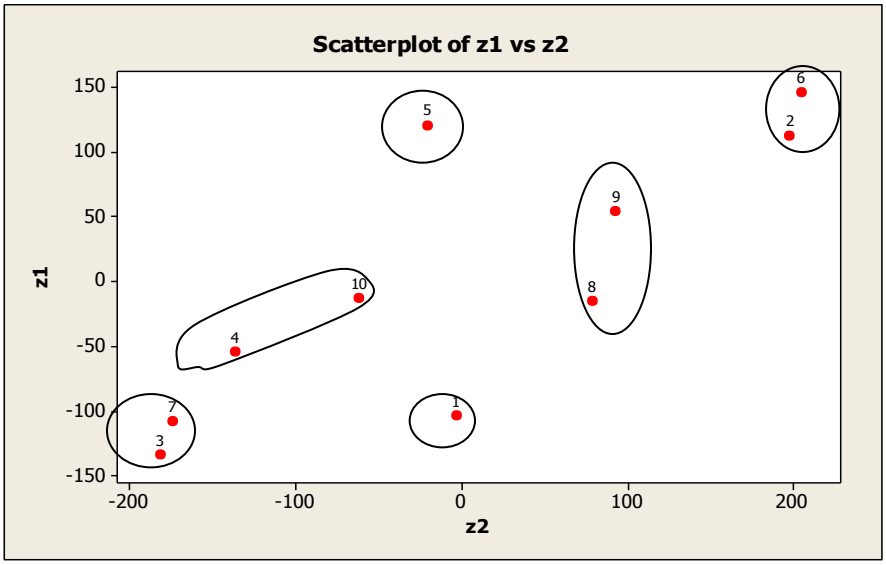
Untuk melihat ada tidaknya pengelompokkan atau mengidentifikasi ada tidaknya klaster yang terbentuk dapat dilihat dari diagram pencar objek-objek di dimensi dua. Plot objek-objek dapat dilakukan di dimensi dua karena dari hasil Scree Plot terdapat break atau patahan di antara komponen kedua dan ketiga, sehingga dapat

disimpulkan bahwa representasi objek dapat dilakukan pada ruang berdimensi dua (bidang).

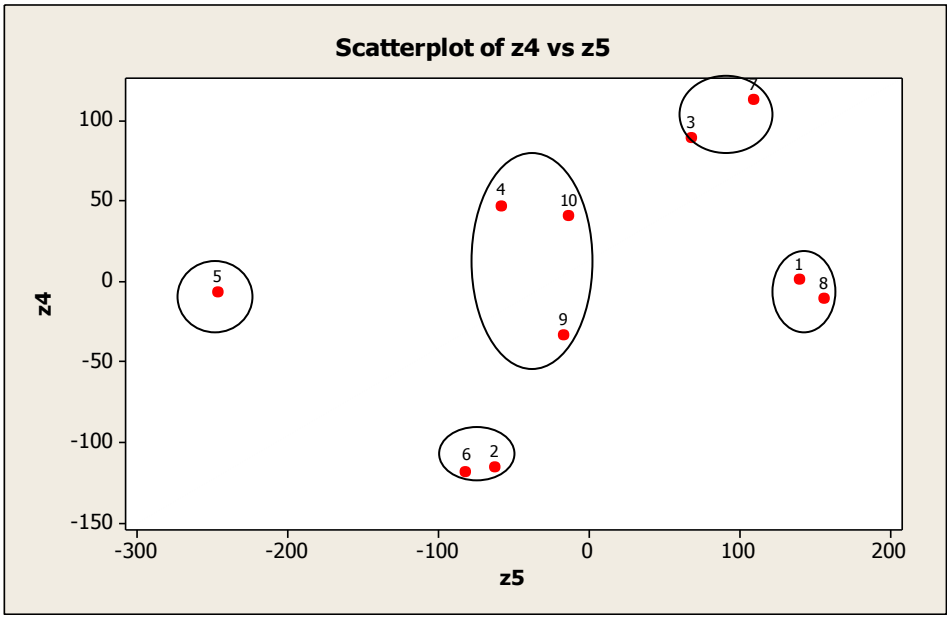


Gambar 1 : *Scree-PlotTest*untuk sepuluh data *z-score*

Jika kita plot data dengan diagram pencar z_1 vs z_2 , z_4 vs z_5 maka dapat diprediksi kemungkinan akan terjadi 5 atau 6 kluster, seperti diperlihatkan pada gambar 2 dan 3 berikut:



Gambar 2 *Scatterplot* z1 dengan z2 untuk melihat ada tidaknya pengelompokkan



Gambar 3 *Scatterplot* z4 dengan z5 untuk melihat ada tidaknya pengelompokkan

Dari gambar 4 dan gambar 5 dapat diidentifikasi bahwa ada pengelompokkan dan kemungkinan akan terdapat 5 atau 6 klaster.

Selanjutnya perhatikan hasil-hasil untuk *Single Linkage Method* berikut ini. Dengan menggunakan software SPSS diperoleh matriks dari langkah pertama dan kedua sbb :

	1	2	3	4	5	6	7	8	9	10
1	000,00	472,55	277,31	255,93	481,20	527,10	294,36	131,96	268,94	198,69
2	472,55	000,00	688,47	529,17	326,29	60,56	713,27	399,98	236,21	458,59
3	277,31	688,47	000,00	202,16	564,94	739,46	67,82	373,02	459,66	241,71
4	255,93	529,17	202,16	000,00	365,17	575,32	245,84	322,84	313,01	99,77
5	481,20	326,29	564,94	365,17	000,00	341,92	600,30	464,97	276,47	343,81
6	527,10	60,56	739,46	575,32	341,92	000,00	764,01	453,47	287,87	507,05
7	294,36	713,27	67,82	245,84	600,30	764,01	000,00	376,04	480,66	268,52
8	131,96	399,98	373,02	322,84	464,97	453,47	376,04	000,00	207,16	234,59
9	268,94	236,21	459,66	313,01	276,47	287,87	480,66	207,16	000,00	299,91
10	198,69	458,59	241,71	99,77	343,81	507,05	268,52	234,59	299,91	000,00

Langkah ketiga menghasilkan sebuah klaster 26 karena min $d_{(2)(6)} = 60,56$.
Langkah keempat : mencari jarak minimum antara klaster 26 dengan klaster lainnya diperoleh matriks jarak baru :

	26	1	3	4	5	7	8	9	10
26	000,00	472,55	688,47	529,17	326,29	713,27	399,98	236,21	458,59
1	472,55	000,00	277,31	255,93	481,20	294,36	131,96	268,94	198,69
3	688,47	277,31	000,00	202,16	564,94	67,82	373,02	459,66	241,71
4	529,17	255,93	202,16	000,00	365,17	245,84	322,84	313,01	99,77
5	326,29	481,20	564,94	365,17	000,00	600,30	464,97	276,47	343,81
7	713,27	294,36	67,82	245,84	600,30	000,00	376,04	480,66	268,52
8	399,98	131,96	373,02	322,84	464,97	376,04	000,00	207,16	234,59
9	236,21	268,94	459,66	313,01	276,47	480,66	207,16	000,00	299,91
10	458,59	198,69	241,71	99,77	343,81	268,52	234,59	299,91	000,00

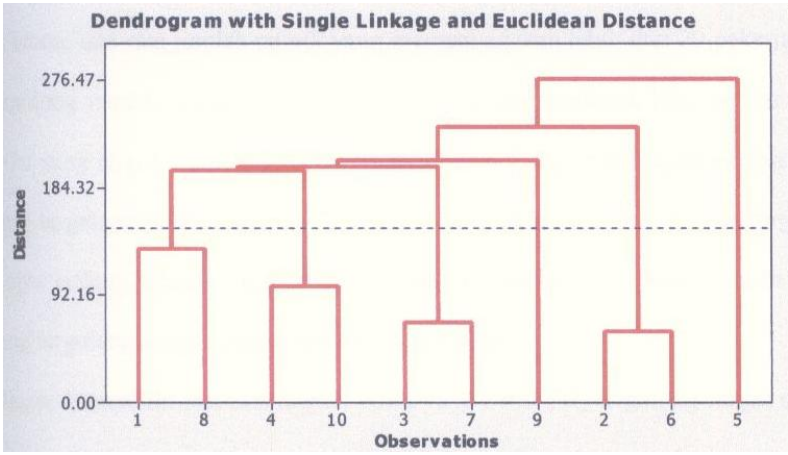
Klaster 3 dan klaster 7 bergabung karena min $d_{(3)(7)} = 67,82$.
Selanjutnya hitung jarak antar klaster 37 dengan klaster lainnya dan diperoleh matriks jarak yang baru :

	37	26	1	4	5	8	9	10
37	000,00	688,47	277,31	202,16	564,94	373,02	459,66	241,71
26	688,47	000,00	472,55	529,17	326,29	399,98	236,21	458,59
1	277,31	472,55	000,00	255,93	481,20	131,96	268,94	198,69
4	202,16	529,17	255,93	000,00	365,17	322,84	313,01	99,77
5	564,94	326,29	481,20	365,17	000,00	464,97	276,47	343,81
8	373,02	399,98	131,96	322,84	464,97	000,00	207,16	234,59
9	459,66	236,21	268,94	313,01	276,47	207,16	000,00	299,91
10	241,71	458,59	198,69	99,77	343,81	234,59	299,91	000,00

Demikian seterusnya, sampai diperoleh hasil terakhir yang diperoleh dari tahap pengklasteran dengan algoritma *Single Lingkage Method*sebagai berikut :

Single Linkage Method			
Tahap	Klaster yang bergabung		Jarak Euclid
1	2	6	60,56
2	3	7	67,82
3	4	10	99,77
4	1	8	131,96
5	18	410	198,69
6	14810	37	202,16
7	1347810	9	207,16
8	13478910	26	236,21
9	1234678910	5	276,47

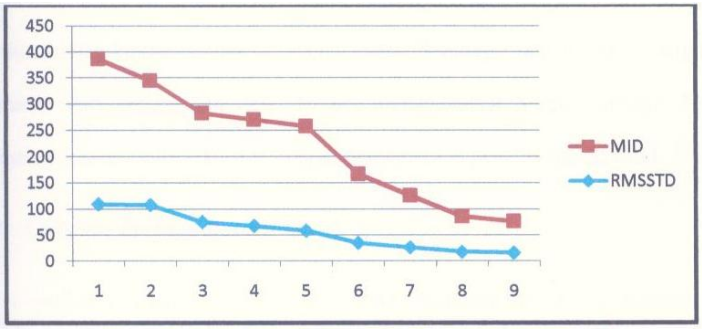
Dengan SPSS dendogram untuk Single Linkage Method untuk sepuluh data tersebut:

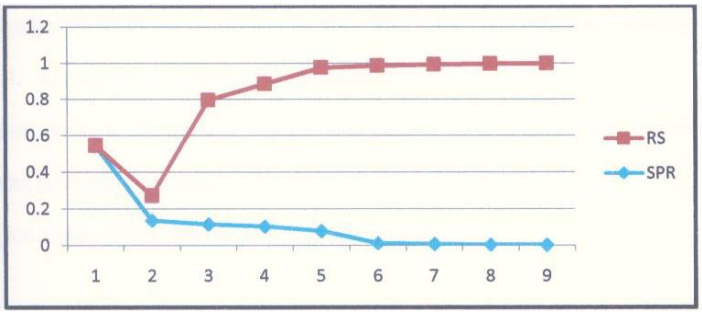


Tabel 6 : Hasil Pengklasteran *Single Linkage Method*untuk Sepuluh Data

Klaster	Anggota klaster	Kota
1	18	Pheonix, Jacksonville
2	410	Denver, Atlanta
3	37	San Fransisco, Washington
4	9	Miami
5	26	Little Rock, Wilmington
6	5	Hartford

Hasil tahap validasi untuk hasil pengklasteran yang diperoleh tersebut diberikan pada gambar 4 berikut.





Gambar 4 Plot RMSSTD dan MID (atas), Plot RS dan SPR (bawah) untuk *Single Lingkage Method*

Perbedaan yang cukup jelas terlihat pada gambar 4, 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 4 atas), demikian pula nilai SPR kecil dan nilai RS besar (gambar 4 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.

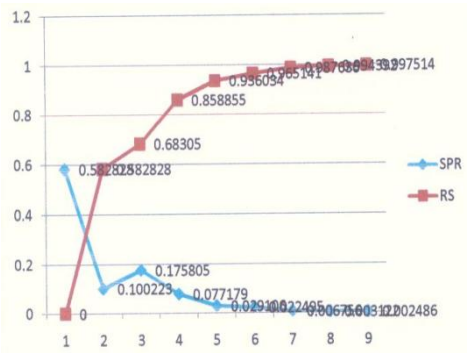
Hasil-hasil penerapan algoritma (hasil pengklasteran) metode-metode *agglomerative* yang lain dapat dilihat pada Tabel 7 dan 8 berikut, untuk uraian setiap langkahnya pada setiap metode dapat dilihat pada Sofyana (2010) dan Asumpta (2010).

Tabel 7 : Hasil Pengklasteran *Complete Linkage Method, Average Linkage Method*dan*Ward's Method*untuk Sepuluh Data

Klaster	Anggota Klaster	Kota
1	18	Phoenix, Jacksonville
2	410	Denver, Atlanta
3	37	San Fransisco, Washington
4	9	Miami
5	26	Little Rock, Wilmington
6	5	Hartford

Tabel 7 : Hasil Pengklasteran *Centroid Method*&*Median Method*untuk Sepuluh Data

Klaster	Anggota Klaster	Kota
1	26	Little Rock, Wilmington
2	37	San Fransisco, Washington
3	410	Denver, Atlanta
4	189	Phoenix, Jacksonville, Miami
5	5	Hartford



Gambar 5 Plot RMSSTD dan CD (kiri), dan Plot SPR dan RS (kanan) untuk *Centroid Method*

Hasil validasi dengan Plot RMSSTD dan CD untuk *Centroid Method*(gambar 5 kiri) menunjukkan untuk jumlah klaster 5 nilai CD cukup besar dan nilai RMSSTD cukup kecil, demikian pula nilai RS cukup besar dan nilai SPR cukup kecil (gambar 5 kanan). Hasil validasi yang sama untuk jumlah klaster 5 diperoleh untuk *Median Method* (Asumpta, 2010:80).

IV. KESIMPULAN

Hasil pengelompokan untuk data uji coba yaitu sepuluh data observasi tingkat polusi udara di sepuluh kota di Amerika Serikat memberikan hasil 5 jumlah klaster seperti ditunjukkan Tabel 8 atau 7 jumlah klaster seperti ditunjukkan pada Tabel 7, hal ini didukung dengan validasi dengan plot Plot RMSSTD dan CD beserta plot RS dan SPR untuk masing-masing metode.

Walaupun berbeda cara dari masing-masing metode *agglomerative* hirarki, akan tetapi semua algoritma metode-metode tersebut mengikuti algoritma umum (algoritma 1). Tidak ada jaminan dalam metode klastering hirarki akan terjadi salah mengelompokkan objek-objek pada tahap awal. Akibatnya konfigurasi akhir dari klaster-klaster harus diperhatikan secara seksama jika hal ini sangat sensitif. Lebih baik memang untuk satu kasus dicobakan beberapa metode *agglomerative* hirarki yang berbeda seperti yang telah penulis lakukan, dan dapat pula dicobakan digunakan ukuran similaritas (lihat Johnson, 1982: 538) sebagai pembanding selain dari ukuran jarak Euclid yang sering digunakan. Jika hasil dari metode-metode hirarki ini konsisten satu sama lain, maka sifat pengelompokkan yang sebenarnya dapat diperoleh (Johnson, 1982:554).

DAFTAR PUSTAKA

Asumpta, E. (2010). *Centroid Method dan Median Method* Pada Analisis Klaster. Tugas Akhir. Bandung : Jurusan Pendidikan Matematika FPMIPA UPI.

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

Jackson, J. E.(1991). *A User's Guide To Principal Components*. Canada : John Wiley&Sons,Inc.

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

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

Sofyana, F. R. (2010). *Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method*Pada Analisis Klaster. Tugas Akhir. Bandung : Jurusan Pendidikan Matematika FPMIPA UPI.

Supranto (2004). *Analisis Multivariat Arti dan Interpretasi*. Jakarta : Rineka Cipta.