Perbandingan Hasil Analisis *Cluster* dengan Menggunakan Metode *Single Linkage* dan Metode *C-Means* (Studi Kasus: Data Tingkat Kualitas Udara Ambien pada Perusahaan Perkebunan di Kabupaten Kutai Barat Tahun 2014)

Comparison From the Result of Cluster Analyse Using Single Linkage Method and C-Means Method (Case Study: Ambient Air Quality Levels in Plantation Company in West Kutai in Year 2014)

Maria Goreti[1], Yuki Novia N[2], Sri Wahyuningsih[2]
[1]Mahasiswa Program Studi Statistika
[2]Dosen Program Studi Statistika
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman
Email: maria_mien_jhun@yahoo.com

Abstract

Cluster analysis is one of the multivariate analysis which is used to classify objects into groups based on similarity of observed variables, in order to obtain the similarity of objects in the same group compared between objects of different groups. Cluster analysis is divided into two methods, they are is hierarchy method that start grouping with two or more objects that have the closest similarity and non-hierarchical method that begin with the process of determining the number of clusters in advance. This study aims is to determine whether there are differences in the results of the cluster grouping formed by using the hierarchy method, that is single linkage method, and non-hierarchical method, that is C-means method. Data, which is taken from the Environment Agency West Kutai, is data Ambient Air Quality Levels in Plantation Company in West Kutai in 2014. The results showed that based on the type of pollutants from all aleven the eleventh plantation companies have different results clusters formed from both methods which were used. With the characteristics of each cluster or groups: single linkage method for the first cluster has good air quality and its members as much as 7 companies, second Cluster both have poor air quality and its members as much as two companies and for the third Cluster have fairly good air quality and its members as much as 2 companies. As for the method of C-means for the first cluster has good air quality and its members as many as four companies, second Cluster both have poor air quality and its members as many as four companies and third Cluster have fairly good air quality and its members as much as 3 companies. For the average value of the ratio of standard deviation in the group (S_w) and between groups (S_b) by using the method of single linkage has a smaller value that is equal to 0.022 while C-means method is equal to 0.063. Thus, in the case of the classification of the ambient air quality in plantation companies in West Kutai 2014, single linkage method better at classifying than C-means method.

Keywords: Cluster analysis,C-means, ambient air quality, single linkage.

Pendahuluan

Analisis *cluster* adalah salah satu analisis multivariat yang digunakan untuk mengelompokkan objek-objek menjadi beberapa *cluster* berdasarkan kemiripan variabel-variabel yang diamati, sehingga diperoleh kemiripan objek dalam *cluster* yang sama dibandingkan antar objek dari *cluster* yang berbeda. Secara umum analisis *cluster* dibagi menjadi dua metode yaitu metode *hierarki* dan metode *non-hierarki*. Di dalam metode *hierarki* sendiri terdapat beberapa metode, di antaranya metode pautan tunggal (*single linkage*), metode pautan lengkap (*complete linkage*), metode antar pusat (*centroid linkage*), metode pautan rata-rata (*average linkage*) dan metode Ward (*Ward's method*), sedangkan metode yang termasuk dalam metode *non-hierarki* adalah metode *K-means (C-means)* (Supranto, 2004).

Pembangunan di Indonesia khususnya pada sektor industri telah mengalami perkembangan yang cukup pesat. Dengan semakin banyak industri yang berkembang di suatu daerah, salah satunya yaitu perusahaan perkebunan, maka kualitas udara ambien (udara bebas) semakin berkurang. Oleh sebab itu, diperlukan pengelompokkan perusahaan berdasarkan kualitas udara ambien untuk mengetahui tingkat kualitas udara dari tiap perusahaan, yaitu kualitas udara yang kurang baik, kualitas udara yang cukup baik dan kualitas udara yang baik. Untuk mengetahui apakah terdapat perbedaan hasil pengelompokkan pada *cluster* yang terbentuk dengan menggunakan metode single linkage dan metode *C-means.*

Berdasarkan uraian di atas, maka penulis tertarik untuk melakukan penelitian dengan judul "Perbandingan Hasil Analisis *Cluster* dengan Menggunakan Metode Single Linkage dan Metode

C-means", dengan Studi Kasus Tingkat Kualitas Udara Ambien pada Perusahaan Perkebunan di Kabupaten Kutai Barat Tahun 2014.

Analisis *Cluster*

Menurut Supranto (2004), analisis *cluster* yaitu analisis untuk mengelompokkan elemen yang mirip sebagai objek penelitian menjadi *cluster* yang berbeda dan *cluster* saling meniadakan (*mutually exclusive*). Analisis *cluster* termasuk dalam analisis statistik multivariat metode interdependen. Analisis *cluster* merupakan salah satu alat analisis yang berguna sebagai peringkas data. Dalam meringkas data ini dapat dilakukan dengan jalan mengelompokkan objek-objek yang hendak diteliti. Tujuan utama analisis *cluster* adalah mengklasifikasi objek (kasus/elemen) seperti manusia, produk (barang), toko, perusahaan ke dalam kelompok-kelompok yang relatif homogen didasarkan pada suatu set variabel yang dipertimbangkan untuk diteliti. Objek di dalam kelompok harus relatif mirip/sama (*relatively similar*). Dinyatakan dalam variabel-variabel dan harus berbeda jauh dengan objek dari kelompok lain. Objek tersebut akan diklasifikasikan ke dalam satu atau lebih *cluster* sehingga objek–objek yang berada dalam satu *cluster* akan mempunyai kemiripan atau kesamaan karakter.

Adapun ciri-ciri *cluster* adalah:

1. Homogenitas (kesamaan) yang tinggi antar anggota dalam satu *cluster* (*within-cluster*).
2. Heterogenitas (perbedaan) yang tinggi antar *cluster* yang satu dengan *cluster* yang lainnya (*between-cluster*).

Dari dua hal di atas dapat disimpulkan bahwa sebuah *cluster* yang baik adalah *cluster* yang mempunyai anggota-anggota yang semirip mungkin satu dengan yang lain, namun sangat tidak mirip dengan anggota-anggota *cluster* yang lain.

Analisis *cluster* merupakan suatu kelas teknik, dipergunakan untuk mengklasifikasi objek atau kasus ke dalam kelompok yang relatif homogen, yang disebut *cluster*. Objek dalam setiap *cluster* cenderung mirip satu sama lain dan berbeda jauh (tidak sama) dengan objek dari *cluster* lainnya. Pengelompokkan dilakukan berdasarkan kemiripan (*similarity*) antar objek. Kemiripan diperoleh dengan meminimalkan jarak antar objek dalam *cluster* (*within-cluster*) dan memaksimalkan jarak antar *cluster* (*between-cluster*)

Proses Analisis *Cluster*

Untuk melakukan analisis *cluster* ada beberapa proses yang harus dilakukan. Proses analisis *cluster* tersebut meliputi :

1. Ukuran Jarak Kemiripan

Dalam Sartono dkk (2003), sesuai prinsip *cluster* yaitu mengelompokkan objek yang mempunyai kemiripan, maka proses pertama pada analisis *culster* adalah mengukur seberapa jauh kesamaan antar objek. Dengan memiliki sebuah ukuran kuantitatif untuk mengatakan bahwa dua objek tertentu lebih mirip dibandingkan dengan objek lain, akan menghilangkan kebingungan dan mempermudah proses formal dalam pengklasifikasian. Salah satu yang jelas bisa menjadi ukuran ketakmiripan adalah fungsi jarak antara objek *a* dan *b*, yang biasa dinotasikan dengan d(*a*,*b*).

Sifat – sifat ukuran ketakmiripan adalah :

1. $d(a,b) \geq 0$
2. $d(a,a) = 0$
3. $d(a,b) = d(b,a)$
4. $d(a,b)$ meningkat seiring semakin tidak mirip kedua objek *a* dan *b*.

Fungsi jarak ini juga memenuhi ketaksamaan segitiga yang menyatakan bahwa $d(a,c) \leq d(a,b) + d(b,c)$.

Dalam Simamora (2005), jarak yang paling umum digunakan adalah jarak *Euclidean*, yang mengukur jarak sesungguhnya menggunakan mata manusia. Jarak *Euclidean* adalah besarnya jarak suatu garis lurus yang menghubungkan antar objek. Misalkan ada dua objek yaitu A dengan koordinat (a_1, b_1) dan B dengan koordinat (a_2,b_2) maka jarak antar kedua objek tersebut dapat diukur dengan rumus :

$$\sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2} \quad (1)$$

Ukuran jarak atau ketidaksamaan antar objek ke-i dengan objek ke-*j*, disimbolkan dengan d_{ij} dan k=1, 2, ...,p. nilai d_{ij} diperoleh melalui perhitungan jarak *Euclidean* sebagai berikut:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(a_{ik} - a_{jk}\right)^2} \quad (2)$$

Memilih Prosedur Pengklasifikasian

Proses *cluster* atau pengelompokkan data bisa dilakukan dengan dua metode, yaitu:

1. Metode *Hierarki*

Metode ini memulai pengelompokkan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian operasi diteruskan ke objek lain yang mempunyai kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk semacam 'pohon' di mana ada *hierarki* (tingkatan) yang jelas antar objek, dari yang paling mirip sampai paling tidak mirip. Metode *hierarki* bisa aglomeratif (*agglomerative)* atau difisif (*divisive)*. Pengklasifikasian aglomeratif dimulai dengan setiap objek dalam suatu *cluster* yang terpisah. *Cluster* dibentuk

dengan mengelompokkan objek yang semakin membesar (semakin banyak elemen atau objek yang menjadi anggotanya). Proses ini dilanjutkan sampai semua objek menjadi anggota dari suatu *cluster* tunggal. Sedangkan pengklasifikasian difisif, dalam prosesnya, merupakan kebalikan dari metode aglomeratif. Metode ini dimulai dengan menempatkan semua objek sebagai satu *cluster*. Lalu, secara bertahap, objek-objek dipisahkan ke dalam *cluster* yang berbeda, dua *cluster*, tiga *cluster*, dan seterusnya, sampai semua objek menjadi *cluster* sendiri-sendiri.

Menurut Simamora (2005), ada lima metode gabungan (Aglomeratif) dalam pembentukan *cluster*, yaitu:

a. Pautan Tunggal (*Single Linkage*)

Metode ini akan mengelompokkan dua objek yang mempunyai jarak terdekat terlebih dahulu. Jika A dan B mempunyai jarak terdekat (misal 4,2) dibanding jarak A dan C (misal 8) atau B dan C (misal 5,6), maka proses hierarki pertama adalah mengelompokkan A-B. Selanjutnya *cluster* A-B akan 'menambah' anggotanya dengan mencari variabel dengan jarak terdekat dengannya. Demikian seterusnya akan terjadi proses pengelompokkan secara hierarki.

Metode *single linkage* akan mengelompokkan dua objek yang mempunyai jarak terdekat dahulu. Jadi pada setiap tahapan, banyaknya *cluster* berkurang satu. Secara formal, dua buah *cluster* B_r dan B_s, jarak antara B_r dan B_s misalkan $h(B_r,B_s)$ didefinisikan sebagai berikut :

$$h(B_r,B_s) = \min\{d(X_i,X_j);\ X_i \text{ anggota } B_r,\ X_j \text{ anggota } B_s\} \quad (3)$$

Hasil berupa *single linkage clustering* dapat disajikan dalam bentuk suatu dendogram atau diagram pohon. Cabang-cabang pohon menunjukkan *cluster*/kelompok. Cabang-cabang tersebut bertemu bersama-sama (menggabung) pada simpul yang posisinya sepanjang suatu sumbu jarak (kemiripan) menunjukkan tingkat dimana penggabungan terjadi.

b. Pautan Lengkap (*Complete Linkage*)

Metode ini akan mengelompokkan dua objek yang mempunyai jarak terjauh dulu. Kemudian proses diteruskan untuk jarak antar variabel yang makin dekat.

c. Pautan Rata-rata (*Average Linkage*)

Metode ini akan mengelompokkan dua objek berdasarkan jarak rata-rata yang didapat dengan melakukan rata-rata semua jarak antar objek terlebih dahulu.

d. Metode Ward (*Ward's Method*)

Pada metode ini, jarak antara dua *cluster* yang terbentuk adalah *sum of squares* di antara dua *cluster* tersebut.

e. Metode Pusat (*Metode Centroid*)

Pada metode ini, jarak antara dua *cluster* adalah jarak di antara dua *centroid clusters* tersebut. *Centroid* adalah rata-rata jarak yang ada pada sebuah *cluster*. Dengan metode ini, setiap terjadi *cluster* baru, segera terjadi perhitungan ulang *centroid*, sampai terbentuk *cluster* yang tetap.

2. Metode *Non-Hierarki*

Metode ini dimulai dengan proses penentuan jumlah *cluster* terlebih dahulu. Salah satu metode *non-hierarki* adalah *C-means*. Metode *C-means* digunakan sebagai alternatif metode *cluster* untuk data dengan ukuran yang besar karena memiliki kecepatan yang lebih tinggi dibandingkan metode *hierarki*. Mac Queen dalam Sartono dkk (2003) menyarankan penggunaan *C-means* untuk menjelaskan algoritma dalam penentuan suatu objek ke dalam *cluster* tertentu berdasarkan rataan terdekat.

Menurut Agusta (2007), analisis *cluster* metode *C-means clustering* adalah prosedur analisis *cluster* yang digunakan untuk mengidentifikasi kelompok kasus atau objek secara relatif sama yang didasarkan pada karakteristik-karakteristik yang sudah dipilih. *C-means* merupakan salah satu metode pengelompokkan data *non-hierarki* yang berusaha menggabungkan data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini menggabungkan data ke dalam kelompok sehingga data berkarateristik sama dimasukkan ke dalam kelompok yang sama. Proses pengklasifikasian dengan metode *C-means* adalah sebagai berikut:

a. Menentukan besarnya c, yaitu banyaknya *cluster*.
b. Menentukan *centroid* secara acak (terserah peneliti)
c. Menghitung jarak tiap objek dengan setiap *centroid* menggunakan jarak *Euclidean*
d. Mengelompokkan setiap data berdasarkan jarak terdekat data tersebut dengan setiap *centroid*.
e. Menentukan posisi *centroid* yang baru pada proses iterasi dengan menghitung nilai rata-rata dari data yang ada pada *centroid* yang sama menggunakan persamaan:

$$C_{hj} = \frac{\sum_{k=1}^{n_i} X_{kj}}{n_h} \quad (4)$$

f. Mengulangi langkah 3 jika posisi *centroid* baru tidak sama dengan *centroid* yang sebelumnya.

Pengecekan *konvergensi* dilakukan dengan melihat pada iterasi sebelumnya dengan iterasi yang sedang berjalan. Jika hasilnya sama maka proses *clustering* sudah *konvergen*, tetapi jika

berbeda maka belum *konvergen* sehingga perlu dilakukan iterasi berikutnya.

Menentukan Banyaknya *Cluster*

Isi pokok/utama dalam analisis *cluster* adalah menentukan berapa banyaknya *cluster*. Sebetulnya tidak ada aturan yang baku untuk menentukan barapa banyaknya *cluster*, namun demikian, menurut Simamora (2005), terdapat beberapa petunjuk yang bisa dipergunakan yaitu:

1. Pertimbangan teoritis, konseptual, praktis, mungkin bisa diusulkan/disarankan untuk menentukan beberapa banyaknya *cluster* yang sebenarnya.
2. Di dalam pengklasifikasian *hierarki*, jarak dimana *cluster* digabungkan bisa digunakan sebagai kriteria.
3. Di dalam pengklasifikasian *non hierarki*, rasio jumlah varian dalam *cluster* dengan jumlah varian antar *cluster* dapat diplotkan terhadap banyaknya *cluster.*

Interpretasi Hasil *Cluster*

Setelah *cluster* terbentuk, baik dengan metode *hierarki* maupun *non-hierarki*, langkah selanjutnya adalah melakukan interpretasi terhadap *cluster* yang terbentuk, yang pada intinya memberi nama spesifik untuk menggambar isi *cluster* tersebut. Pengelompokkan tidak bermanfaat apabila kita tidak mengetahui profil setiap kelompok. Untuk menginterpretasi *cluster* dan membuat profil semua *cluster*, digunakan rata-rata setiap *cluster* pada setiap variabel (Simamora, 2005).

Melakukan Validasi *Cluster*.

Menurut Santoso (2014), untuk menguji validasi *cluster* digunakan uji parsial F dan signifikansi yang terdapat pada tabel ANOVA. Tabel ANOVA digunakan untuk melihat tingkat signifikansi antar *cluster*. Teknik ANOVA akan menguji variabilitas dari observasi masing-masing kelompok dan variabilitas antar mean kelompok. Melalui kedua variabilitas tersebut, akan dapat ditarik kesimpulan mengenai mean populasi. Jika diketahui bahwa x_j merupakan variabel pembeda pengklasifikasian dan *c* adalah banyaknya *cluster* yang terbentuk. Nilai F diperoleh dari rata-rata jumlah kuadrat (*mean square)* dengan rata-rata jumlah kuadrat dalam kelompok. Menurut Sudjana (2005) rumus ANOVA adalah sebagai berikut:

$$F = \frac{means\ square\ cluster}{means\ square\ error} = \frac{JKA/(c-1)}{JKD/\sum(n-c)} \quad (5)$$

Hipotesis yang digunakan dalam pengujian ANOVA adalah :

H_0: variabel X_j bukan variabel pembeda dalam pengklasifikasian.

H_1: variabel X_j merupakan variabel pembeda dalam pengklasifikasian.

Dasar dari pengambilan keputusan adalah:

Jika F hitung $\geq$ F tabel, maka H_0 ditolak

Jika F hitung < F tabel, maka H_0 diterima

Dengan menggunakan df_1 = c-1 dan df_2 = n-c

Evaluasi Hasil Pengelompokkan

Menurut Bunkers dan Miller (1996), untuk melihat kinerja kedua metode atau untuk melihat kualitas pengelompokkan yang terbaik, digunakan kriteria nilai simpangan baku, yaitu simpangan baku dalam *cluster* (S_w) dan simpangan baku antar *cluster* (S_b). Untuk menghitung nilai S_w dan nilai S_b digunakan persamaan berikut:

$$S_{wj} = \frac{1}{c}\sum_{k=1}^{c} S_{kj} \quad (6)$$

$$S_{bj} = \left[\frac{1}{c-1}\sum_{k=1}^{c}\left(\overline{X}_{kj} - \overline{X}\right)^2\right]^{1/2} \quad (7)$$

Semakin kecil nilai S_w dan semakin besar nilai S_b, maka metode tersebut memiliki kinerja yang baik, artinya mempunyai homogenitas yang tinggi. Sehingga digunakan rasio antara S_w dan S_b, nilai rataan rasio S_w/S_b yang terkecil menunjukkan ketepatan pengelompokkan yang paling baik.

Metodologi Penelitian

Penelitian ini dilaksanakan pada bulan Januari sampai dengan Mei 2015. Tempat pengambilan data diambil dari BLH (Badan Lingkungan Hidup) Kabupaten Kutai Barat dan pengolahan data dilakukan di Laboratorium Statistika Terapan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman (FMIPA UNMUL). Dalam penelitian ini menggunakan rancangan *kausal komparatif* yang bersifat *ex post facto*, artinya data dikumpulkan setelah semua kejadian yang dipersoalkan berlangsung atau lewat (Sugiyono, 2005).

Dalam penelitian ini data yang digunakan adalah data tingkat kualitas udara ambien pada perusahaan perkebunan di Kabupaten Kutai Barat tahun 2014. Variabel dan objek yang digunakan dalam penelitian ini yaitu ada 11 objek yang merupakan nama perusahaan perkebunan di Kabupaten Kutai Barat, yaitu PT. Borneo Persada Energi Jaya (X_1) , PT. Borneo Surya Mining Jaya (X_2), PT. Citra Agro Kencana (X_3), PT. Fanggiono Agro Plantation (X_4), PT. Farinda Bersaudara (X_5), PT. Ketapang Hijau Lestari (X_6), PT. Ketapang Agro Lestari (X_7), PT. Maha Karya Bersama (X_8), PT. Munte Waniq Jaya Perkasa (X_9), PT. Perkebunan Sentawar Membangun (X_{10}) dan PT. Teguh Swakarsa Sejahtera (X_{11}). Sedangkan variabel yang digunakan yaitu ada 4 variabel, dimana merupakan jenis polutan yang meliputi *Sulfur dioksida* (SO_2) (X_{12}), *Nitrogen dioksida* (NO_2) (X_{13}), *Carbonmonoksida* (CO) (X_{14}) dan debu (TSP = *Total Suspended Particulate*) (X_{15}).

Menerapkan analisis *cluster* pada data penelitian yaitu:

1. Menghitung ukuran jarak kemiripan atau *similarity* dengan menggunakan perhitungan jarak *Euclidean* dengan menggunakan persamaan (2).
2. Membuat *cluster* dengan dua metode, yaitu metode *single linkage* dan metode *C-means*.
   a. Menghitung dengan menggunakan metode *single linkage* yaitu dengan menentukan jarak ketakmiripan yang digunakan persamaan (3).
   b. Menghitung dengan menggunakan metode *C-means* yaitu dengan:
      - Menentukan nilai *c*, yaitu banyaknya *cluster* dan menentukan *centroid* di setiap *cluster.*
      - Mengalokasikan data dalam suatu kelompok secara acak.
      - Menghitung nilai *centroid* masing-masing *cluster* dengan menggunakan persamaan (4)
      - Menghitung jarak tiap-tiap data ke *centroid* dengan menggunakan persamaan dan mengaloksikan masing-masing data ke suatu kelompok berdasarkan jarak terdekat.
      - Kembali ke langkah c, apabila masih terdapat perpindahan data dari satu *cluster* ke *cluster* yang lain.
3. Mengevaluasi hasil pengelompokkan melalui nilai simpangan baku dalam kelompok (S_w) dan nilai simpangan baku antar kelompok (S_b), maka metode pengelompokkan memiliki kinerja semakin baik. Agar mempermudah dalam membandingkan nilai tersebut maka digunakan nilai rataan rasio antara S_w/S_b, nilai terkecil menunjukkan ketepatan pengelompokkan yang terbaik. Untuk menghitung nilai S_w dengan menggunakan persamaan (6) dan nilai S_b dengan menggunakan persamaan (7).
4. Kesimpulan diperoleh dengan melihat hasil analisis yang dilakukan, kemudian hasil tersebut diinterpretasikan.

Hasil Dan Pembahasan

1. Menentukan Ukuran Jarak Kemiripan

Jarak tiap objek (perusahaan) yang dihitung dengan jarak *euclidean.* Perhitungan jarak *Euclidean* dengan menggunakan persamaan (2). Jarak antara X_1 dengan X_2, yaitu sebesar 37,002. Jarak antara X_1 dengan X_3 menggunakan perhitungan yang sama yaitu sebesar 0,362. Hal ini menunjukkan bahwa X_1 lebih mirip karakteristiknya dengan X_3, dari pada X_1 dengan X_2. Demikian seterusnya untuk penafsiran objek-objek yang lain. Semakin kecil nilai jarak antara dua objek, maka semakin mirip kedua objek tersebut.

2. Memilih Prosedur Pengklasifikasian

Proses *cluster* atau pengelompokkan data adalah sebagai berikut:

- Pengklasifikasian dengan *Single Linkage*

Proses pengklasifikasian dengan metode *single linkage* yaitu pada awal proses ada sebelas *cluster*, tahap pertama yang dilakukan adalah menentukan jarak yang terdekat antara dua objek dari sekian banyak kombinasi jarak dari kesebelas objek dengan menggunakan jarak *euclidean*. Jarak antara X_9 dan X_{10} merupakan jarak terdekat yaitu sebesar 0 sehingga kedua perusahaan tersebut menjadi satu *cluster*. Sekarang tersisa sepuluh *cluster*.

Kemudian dilakukan perbaikan matriks jarak menggunakan *single linkage* dengan persamaan (3), sehingga terjadi perubahan jarak yang melibatkan *cluster* baru (*cluster* yang anggotanya X_9 dan X_{10}) sehingga diperoleh matrik jarak yang baru. Sebagai contoh perhitungan antara (X_9, X_{10}) dengan (X_1) sebagai berikut:

$$h\{(X_9, X_{10}),(X_1)\} = \min\{d(X_9, X_1), d(X_{10}, X_1)\}$$
$$= \min\{1.353{,}256, 1.353{,}256\}$$
$$= 1.353{,}256$$

Demikian seterusnya pada objek yang lain. Kemudian dilakukan perbaikan matriks jarak menggunakan metode *single linkage* seperti pada contoh di atas. Kemudian menentukan jarak terdekat pada objek berdasarkan matriks jarak yang baru. Dengan demikian hasil akhir matriks jarak tersisa 3 *cluster* seperti pada Tabel 1 berikut :

Tabel 1. Matriks Jarak 3 *Cluster*

Cluster	A	B	C
A	0	1.399,527	1.125,886
B	1.399,527	0	283,828
C	1.125,886	283,828	0

dengan

A = $(\boldsymbol{X_1, X_2, X_3, X_4, X_6, X_7, X_8})$

B = $(\boldsymbol{X_5, X_1})$

C = $(\boldsymbol{X_9, X_1})$

Untuk memperjelas proses penggabungan satu demi satu dapat digambarkan dalam bentuk *dendogram* pada Gambar 1.

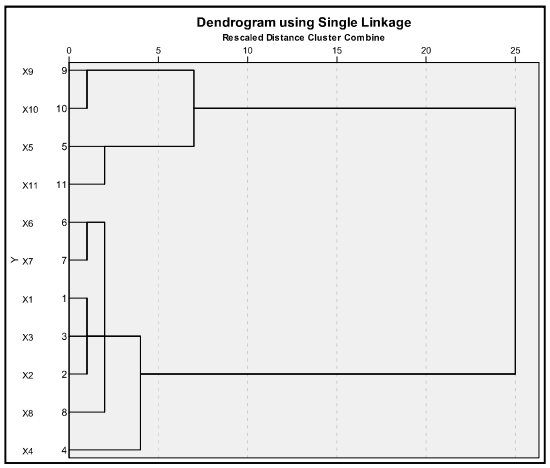
Dari ketiga *cluster* yang terbentuk sehingga dapat mengklasifikasikan sebagai berikut:

Cluster pertama : memiliki rata- rata nilai variabel (polutan) yang paling kecil yang berarti bahwa *cluster* pertama memiliki kualitas udara yang dapat digolongkan baik dan anggotanya adalah X_1, X_2, X_4, X_6, X_7 dan X_8.

Cluster kedua : memiliki rata- rata nilai variabel (polutan) yang terbesar yang berarti bahwa *cluster* kedua memiliki kualitas udara yang kurang baik. Karena perusahaan-perusahaan dalam

cluster ini menghasilkan polutan yang lebih banyak dan anggotanya adalah X_5 dan X_{11}.

Cluster ketiga : memiliki rata- rata nilai variabel (polutan) yang lebih besar dari *cluster* pertama dan lebih kecil dari *cluster* kedua, yang berarti bahwa *cluster* ketiga memiliki kualitas udara yang dapat digolongkan cukup baik dan anggotanya adalah X_9 dan X_{10}.



Gambar 1**.** *Dendogram Single Linkage*

- Pengklasifikasian dengan *C-means*

Proses pengklasifikaisn dengan metode *C-means*, pada tahap awal yaitu :

a. Menentukan besar *c* yaitu sebesar 3.
b. Mengalokasikan data dalam suatu kelompok secara acak. Hasil kelompok secara acak adalah sebagai berikut:

Tabel 2. Mengalokasikan Data Secara Acak

Perusahaan	X_{12}	X_{13}	X_{14}	X_{15}	*Cluster*
X_1	0	0,211	192	0,017	**1**
X_2	0	0,58	229	0,026	**3**
X_3	0	0,569	192	0,071	**2**
X_4	0,292	0,734	29,72	0,019	**3**
X_5	30,6	134,7	1874,7	172,7	**2**
X_6	0	0,499	331	0,031	**3**
X_7	0	0,388	314	0,035	**1**
X_8	0	0,552	420	0,158	**2**
X_9	13,7	73	1542	57,6	**3**
X_{10}	13,7	73	1542	57,6	**2**
X_{11}	27,3	121,6	1806,1	148,5	**2**

c. Menghitung nilai *centroid* masing-masing *cluster* dengan mengunakan persamaan (4).
d. Setelah mendapat nilai *centroid* dari masing-masing *cluster*, lalu menghitung jarak setiap objek ke *centroid* dengan rumus jarak *Euclidean*, lalu mengalokasikan masing-masing objek ke suatu *cluster* berdasarkan jarak terdekat/terkecil.
e. Kembali ke langkah c, apabila masih terdapat perpindahan objek dari satu *cluster* ke *cluster* yang lain.

Sehingga hasil akhirnya adalah sebagai berikut:

Tabel 3. Keseluruhan Jarak Tiap Objek ke *Centroid*

Perusahaan	c_1***	c_2***	c_3***
X_1	**31,322**	1.506,661	163,000
X_2	**68,320**	1.469,823	126
X_3	**31,320**	1.506,633	163
X_4	**130,960**	1.668,180	325,280
X_5	1.728,182	**197,403**	1.535,656
X_6	170,32	1.368,397	**24,000**
X_7	153,320	1.385,304	**41,000**
X_8	259,32	1.279,952	**65,000**
X_9	1.384,485	**160,410**	1.190,683
X_{10}	1.384,485	**160,410**	1.190,683
X_{11}	1.656,759	**123,418**	1.463,946

Pada Tabel 3 dapat dilihat bahwa anggota dari c_1: X_1, X_2, X_3, X_4, untuk c_2: X_5, X_9, X_{10}, X_{11}, dan c_3: X_6, X_7, X_8.

Setelah *cluster* terbentuk, tahap selanjutnya yaitu memberi nama spesifik dan menentukan rata-rata jumlah polutan untuk menggambarkan isi *cluster* tersebut.

Cluster pertama : memiliki rata- rata nilai variabel (polutan) yang paling kecil yang berarti bahwa *cluster* pertama memiliki kualitas udara yang dapat digolongkan baik dan anggotanya adalah X_1, X_2, X_3 dan X_4.

Cluster kedua : memiliki rata- rata nilai variabel (polutan) yang terbesar yang berarti bahwa *cluster* kedua memiliki kualitas udara yang kurang baik. Karena perusahaan-perusahaan dalam *cluster* ini menghasilkan polutan yang lebih banyak dan anggotanya adalah X_5, X_9, X_{10} dan X_{11}.

Cluster ketiga :memiliki rata- rata nilai variabel (polutan) yang lebih besar dari *cluster* pertama dan lebih kecil dari *cluster* kedua, yang berarti bahwa *cluster* ketiga memiliki kualitas udara yang dapat digolongkan cukup baik dan anggotanya adalah X_6, X_7 dan X_8.

3. Validasi pada hasil *cluster*

Hasil validasi pada analisis *cluster* adalah sebagai berikut:

a. Validasi pada metode *single linkage*

Tabel 4. Uji F pada *Single Linkage*

Polutan	Cluster		Error		F	Signifikansi
	Mean Square	df	Mean Square	df		
X_{12}	692,799	2	0,690	8	1.004,405	0,000
X_{13}	14.265,931	2	10,746	8	1.327,534	0,000
X_{14}	2.710.253,157	2	12.168,361	8	222,730	0,000
X_{15}	20.439,371	2	36,604	8	558,385	0,000

Hipotesis:

H_0: variabel X_j bukan variabel pembeda dalam pengklasifikasian.

H_1: variabel X_j merupakan variabel pembeda dalam pengklasifikasian

(di mana X_j : X_{12}, X_{13}, X_{14}, X_{15})

Taraf Signifikansi

$\alpha = 5\% = 0,05$

Kriteria Pengujian

Tolak H_0 jika $F_{hitung} > F_{((0,05),(2),(8))}$ atau tolak H_0 jika nilai signifikansi < α

Kesimpulan

Dipilih α = 0,05 dan nilai n = 11, c = 3 sehingga dari tabel distribusi F diperoleh nilai $F_{((0,05),(2),(8))}$ adalah 4,46. Nilai F hitung pada tabel 3 di atas, nilai F_{hitung} variabel X_{12} = 1.004,405, X_{13} = 1.327,534, X_{14} = 222,730, dan X_{15} = 558,385 > $F_{((0,05),(2),(8))}$ = 4,46. Karena nilai $F_{hitung} > F_{((0,05),(2),(8))}$ maka H_0 ditolak, dengan demikian variabel X_{12}, X_{13}, X_{14}, X_{15} merupakan variabel pembeda dalam pengklasifikasian.

Dari tabel 3, juga dapat dilihat nilai signifikansi variabel X_{12} = 0,000, X_{13} = 0,000, X_{14} = 0,000 X_{15} = 0,000 < α = 0,05 sehingga H_0 ditolak yang berarti variabel X_{12}, X_{13}, X_{14}, X_{15} merupakan variabel pembeda dalam pengklasifikasian.

b. Validasi pada metode *C-means*

Tabel 5. Uji F pada *C-means*

Polutan	Cluster		Error		F	Signifikan
	Mean Square	df	Mean Square	df		
X_{12}	576,522	2	29,759	8	19,373	0,001
X_{13}	12.745,172	2	390,936	8	32,602	0,000
X_{14}	2.698.097,81	2	15.207,196	8	177,422	0,000
X_{15}	15.134,873	2	1.362,729	8	11,106	0,005

Hipotesis:

H_0: variabel X_j bukan variabel pembeda dalam pengklasifikasian.

H_1: variabel X_j merupakan variabel pembeda dalam pengklasifikasian

(di mana X_j : X_{12}, X_{13}, X_{14}, X_{15})

Taraf Signifikansi

$\alpha = 5\% = 0,05$

Kriteria Pengujian

Tolak H_0 jika $F_{hitung} > F_{((0,05),(2),(8))}$ atau tolak H_0 jika nilai signifikansi < α

Kesimpulan

Dipilih α = 0,05 dan nilai n = 11, c = 3 sehingga dari tabel distribusi F diperoleh nilai $F_{((0,05),(2),(8))}$ adalah 4,46. Nilai F hitung pada tabel 3 di atas, nilai F_{hitung} variabel X_{12} = 19,373, X_{13} = 32,602, X_{14} = 177,422, dan X_{15} = 11,106 > $F_{((0,05),(2),(8))}$ = 4,46. Karena nilai $F_{hitung} > F_{((0,05),(2),(8))}$ maka H_0 ditolak, dengan demikian variabel X_{12}, X_{13}, X_{14}, X_{15} merupakan variabel pembeda dalam pengklasifikasian.

Dari tabel 3, juga dapat dilihat nilai signifikansi variabel X_{12} = 0,001, X_{13} = 0,000, X_{14} = 0,000, dan X_{15} = 0,005 < α = 0,05 sehingga H_0 ditolak yang berarti variabel X_{12}, X_{13}, X_{14}, X_{15} merupakan variabel pembeda dalam pengklasifikasian.

4. Evaluasi Hasil Pengklasifikasian

Untuk melihat kinerja kedua metode tersebut digunakan kriteria nilai simpangan baku, yaitu simpangan baku dalam *cluster* (S_w) dan simpangan baku antar *cluster* (S_b). Nilai rataan rasio S_w/S_b yang terkecil menunjukkan ketepatan pengelompokkan yang paling baik. Hasil perhitungannya adalah sebagai berikut:

a. Evaluasi ada *single linkage*

Variabel X_{12} = S_{w1}/S_{b1} = 0,815/244,6 = 0,003

Variabel X_{13} = S_{w2}/S_{b2} = 3,143/190,428 = 0,017

Variabel X_{14} = S_{w3}/S_{b3} = 58,111/1.485,216 = 0,039

Variabel X_{15} = S_{w4}/S_{b4} = 5,721/190,778 = 0,030

Rataan Rasio = $\frac{0,003 + 0,017 + 0,039 + 0,030}{4}$

= **0,022**

b. Evaluasi pada *C-means*

Variabel X_{12} = S_{w1}/S_{b1} = 3,018/253,232 = 0,012

Variabel X_{13} = S_{w2}/S_{b2} = 10,864/227,654 = 0,048

Variabel X_{14} = S_{w3}/S_{b3} = 106,835/1.050,220 = 0,102

Variabel X_{15} = S_{w4}/S_{b4} = 20,127/226,031 = 0,089

Rataan Rasio = $\frac{0,012 + 0,048 + 0,102 + 0,089}{4}$

= **0,063**

Nilai rataan rasio simpangan baku dalam *cluster* (S_w) dan antar *cluster* (S_b) dengan menggunakan metode *single linkage* memiliki nilai yang lebih kecil yaitu sebesar 0,022 sedangkan metode *C-means* yaitu sebesar 0,063.

Dengan demikian, untuk kasus pengklasifikasin kualitas udara ambien pada perusahaan perkebunan di Kabupaten Kutai Barat tahun 2014, metode *single linkage* lebih baik dalam melakukan klasifikasi dibandingkan metode *C-means*.

Kesimpulan

Kesimpulan yang diperoleh dari hasil penelitian ini adalah sebagai berikut:

Hasil analisis *cluster* dari 11 perusahaan perkebunan berdasarkan jenis polutan (SO_2, NO_2, CO dan TSP) dapat disimpulkan bahwa terdapat perbedaan hasil *cluster* yang terbentuk dari kedua metode yang digunakan. Dengan ciri-ciri dari tiap *cluster* yaitu: Metode *single linkage* untuk *Cluster* pertama memiliki kualitas udara yang baik dan anggotanya sebanyak 7 perusahaan, untuk *Cluster* kedua memiliki kualitas udara yang kurang baik dan anggotanya sebayak 2 perusahaan dan untuk *Cluster* ketiga memiliki kualitas udara yang cukup baik dan anggotanya sebanyak 2 perusahaan. Sedangkan untuk metode *C-means* untuk *Cluster* pertama memiliki kualitas udara yang baik dan anggotanya sebanyak 4 perusahaan, untuk *Cluster* kedua memiliki kualitas udara yang kurang baik dan anggotanya sebayak 4 perusahaan dan untuk *Cluster* ketiga memiliki kualitas udara yang cukup baik dan anggotanya sebanyak 3 perusahaan.

Nilai rataan rasio simpangan baku dalam kelompok (S_w) dan antar kelompok (S_b) dengan metode *single linkage* lebih kecil yaitu sebesar 0,022 dari pada metode *C-means* yaitu sebesar 0,063. Dengan demikian, untuk kasus pengklasifikasian kualitas udara ambien pada perusahaan perkebunan di Kabupaten Kutai Barat tahun 2014 metode *single linkage* lebih baik dalam melakukan pengklasifikasian daripada metode *C-means.*

Daftar Pustaka

Agusta, Yudi. 2007. *K-Mean-Penerapan, Permasalahan dan Metode Terkait.Jurnal Sistem dan Informatika* Vol.3 (Februari 2007), hal. 47-60

Bunkers M.J. dan J.R. Miller 1996.*Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique. Journal of Climate*, No. 9, hal 130-146

Santoso, Singgih. 2014. *Statistik Multivariat Edisi Revisi Konsep dan Aplikasi dengan SPSS*. Jakarta: Gramedia

Sartono, B., Farid M. Affendi, Utami Dyah Syafitri, I. Made Sumertajaya. dan Yenni Angraeni. 2003. *Analisis Peubah Ganda*. Bogor: IPB

Simamora, Bilson. 2005. *Analisis Multivariat Pemasaran*. Jakarta: PT. Gramedia Pustaka Utama

Sudjana. 2005. *Metode Statistika*. Bandung: Tarsito

Sugiyono. 2005. *Statistika Untuk Penelitian*. Bandung: Alfabeta

Supranto. 2004. *Analisis Multivariat: Arti dan Interprestasi*. Jakarta: PT. Rineka Cipta