

Penggunaan Analisis *Two Step Clustering* untuk Data Campuran

Charles E. Mongi¹

¹PS Matematika FMIPA Universitas Sam Ratulangi Manado, charlesmongi@gmail.com

Abstrak

Penggerombolan adalah proses mengelompokkan objek ke dalam kelompok-kelompok yang memiliki kemiripan. Metode penggerombolan yang sering digunakan adalah metode berhirarki dan tak berhirarki. Kedua metode tersebut mensyaratkan untuk skala data ordinal, interval atau rasio. Permasalahan yang timbul jika peubah bersifat kategorik atau campuran kategorik dan numerik. Untuk mengatasi masalah tersebut digunakan *two step cluster*. Hasil analisis *two step clustering* mengelompokkan menjadi 3 gerombol model kendaraan. Peubah yang berpengaruh terhadap pembentukan gerombol pada Gerombol 1 adalah semua peubah kategorik dan kontinu. Gerombol 2 yang berpengaruh adalah 6 peubah yaitu 1 peubah kategorik dan 5 peubah kontinu. Sedangkan untuk gerombol 3 peubah yang berpengaruh ada 8 yaitu 1 peubah kategorik dan 7 peubah kontinu.

Kata kunci: Peubah Kategorik, Peubah Kontinu, *Two Step Clustering*

Two Step Clustering Analysis for Combination Data

Abstract

Clustering is a process of classifying objects into groups that have similarities. The common clustering methods are hierarchical and non-hierarchical. Ordinal data scale, interval scale and ratio scale are requirements for these methods. Problems appeared if the variables are categoric or categoric-numeric combination. Therefore, two step clustering are used to solve the problem. These analysis took an example of grouping three vehicle models. The results showed variables that influence formation of Clusters 1 are every categoric and continuous variables. In cluster 2, influenced variables are one categoric variable and five continous variables, while in cluster 3, effected variables are one categoric variable and seven continous variables.

Keywords : Categoric variable, Continue variable, *Two Step Clustering*

1. Pendahuluan

Penggerombolan adalah proses mengelompokkan objek ke dalam kelompok-kelompok yang memiliki kemiripan. Saat ini metode penggerombolan yang digunakan secara luas adalah metode penggerombolan berhirarki dan tak berhirarki. Kedua metode ini digunakan untuk data yang berskala interval, rasio atau ordinal. Sedangkan metode *k-rataan* (*k-means*) mensyaratkan peubah dasar penggerombolannya berskala interval.

Permasalahan utama dalam menerapkan analisis gerombol adalah jika peubah dasar penggerombolan bersifat kategorik maupun campuran kategorik dan numerik, contohnya segmentasi desa pada survei potensi desa, segmentasi pelanggan pada riset-riset pemasaran, dan lain-lain. Permasalahan kedua, pada analisis gerombol berhirarki dan *k-rataan* penentuan banyaknya gerombol ditentukan oleh peneliti serta tidak ada uji statistik untuk mengetahui banyaknya gerombol yang pantas (*fit*) dengan data yang dianalisis, sehingga hasil penggerombolan tergantung dari pengetahuan, pengalaman serta subjektifitas peneliti.

Untuk mengatasi masalah penggerombolan tersebut, metode yang dapat digunakan adalah *two step cluster*. Metode *Two Step cluster* merupakan suatu metode penggerombolan yang dapat mengatasi masalah skala pengukuran, data berukuran besar dengan peubah yang memiliki tipe data kategorik dan kontinu, serta mengetahui gerombol optimal yang terbentuk. Gerombol optimal memiliki jarak antar gerombol yang paling jauh, dan jarak antar obyek yang paling dekat [2]. Pada *two step cluster* pendekatan yang dilakukan adalah dengan memilih tipe skala pengukuran tertentu,

dan mentransformasikan peubah-peubahnya sehingga memiliki tipe skala pengukuran yang sama [1].

Metode penggerombolan lain yang dapat digunakan untuk data berskala campuran yaitu metode *Laten Class Model* pada *software Laten GOLD* dan *ALMO*. Dari ketiga metode tersebut, metode *two step cluster* merupakan metode yang paling efisien dari segi waktu, hal ini karena metode *two step cluster* didesain untuk data yang berukuran besar.

2. Analisis Gerombol

Analisis gerombol adalah analisis statistik peubah ganda yang digunakan terhadap n buah individu atau objek yang mempunyai p peubah yang akan dikelompokkan ke dalam k kelompok. Objek yang terletak dalam satu gerombol memiliki kemiripan sifat yang lebih besar dibandingkan dengan objek yang terletak dalam gerombol lain [3].

Menurut [1], terdapat dua metode dalam analisis gerombol yaitu metode berhirarki (*hierarchical clustering method*) dan metode tak berhirarki (*non hierarchical clustering method*). Metode berhirarki digunakan apabila belum ada informasi jumlah gerombol yang akan dibentuk. Sedangkan metode tak berhirarki bertujuan untuk mengelompokkan n objek ke dalam k gerombol ($k < n$) dimana besarnya k telah ditentukan sebelumnya. Pada dasarnya, terdapat dua teknik penggerombolan pada metode berhirarki, yaitu teknik penggabungan (*agglomerative*) dan teknik pembagian (*divisive*), sedangkan metode tak berhirarki antara lain dengan teknik penyekatan (*partitioning*) dan penggunaan grafik.

Ukuran Jarak

Menurut [1], ukuran jarak dibutuhkan untuk setiap pasang objek yang akan dikelompokkan. Beberapa metode pengukuran jarak antar dua objek, yaitu:

1. Jarak Euclidean

Jarak ini merupakan jarak yang umum digunakan dan dapat digunakan apabila semua peubahnya berskala kontinu. Jarak ini harus memenuhi asumsi bahwa peubah-peubah yang diamati tidak berkorelasi dan antar peubah memiliki satuan yang sama. Dalam metode ini, pengukuran jarak dilakukan dengan menghitung akar kuadrat dari penjumlahan kuadrat selisih dari nilai masing-masing peubah. Jarak Euclidean dapat dirumuskan sebagai berikut:

$$d_{i,j} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

dengan:

$d_{i,j}$ = jarak antara objek i dengan objek k

x_{ik} = nilai objek i pada peubah ke- k

x_{jk} = nilai objek j pada peubah ke- k

p = banyaknya peubah yang diamati

2. Jarak Manhattan (*City Block/Minkowski*)

Jarak ini merupakan bentuk umum dari jarak Euclidean. Jarak Manhattan digunakan jika peubah yang diamati berkorelasi atau tidak saling bebas. Dalam metode ini, pengukuran jarak dilakukan dengan menghitung jumlah absolut perbedaan objek untuk masing-masing peubah. Jarak Manhattan dapat dirumuskan sebagai berikut:

$$d_{i,j} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

dengan:

$d_{i,j}$ = jarak antara objek i dengan objek k

x_{ik} = nilai objek i pada peubah ke- k

x_{jk} = nilai objek j pada peubah ke- k

p = banyaknya peubah yang diamati

3. Jarak Mahalonobis

Jarak ini sangat berguna dalam menghilangkan atau mengatasi perbedaan skala pada masing-masing peubah. Jarak Mahalonobis dapat dirumuskan sebagai berikut:

$$d_{i,j} = \sqrt{(x_{ik} - x_{jk})'S^{-1}(x_{ik} - x_{jk})}$$

dengan:

$d_{i,j}$ = jarak antara objek i dengan objek k

x_{ik} = nilai objek i pada peubah ke- k

x_{jk} = nilai objek j pada peubah ke- k

S = matriks kovarian

4. Jarak Log-Likelihood

Jarak ini digunakan untuk peubah berskala kontinu dan kategorik. Jarak antara gerombol j dengan gerombol s dapat dirumuskan sebagai berikut:

$$d(j, s) = \xi_j + \xi_s - \xi_{(j,s)}$$

dengan:

$$\begin{aligned} \xi_j &= -N \left(\sum_{k=1}^{K^A} \frac{1}{2} \log (\hat{\sigma}_k^2 + \hat{\sigma}_{jk}^2) - \sum_{k=1}^{K^B} \sum_{l=1}^{L_k} \frac{N_{jkl}}{N_j} \log \left(\frac{N_{jkl}}{N_j} \right) \right) \\ \xi_s &= -N \left(\sum_{k=1}^{K^A} \frac{1}{2} \log (\hat{\sigma}_k^2 + \hat{\sigma}_{sk}^2) - \sum_{k=1}^{K^B} \sum_{l=1}^{L_k} \frac{N_{skl}}{N_j} \log \left(\frac{N_{skl}}{N_j} \right) \right) \\ \xi_{(j,s)} &= -N \left(\sum_{k=1}^{K^A} \frac{1}{2} \log (\hat{\sigma}_k^2 + \hat{\sigma}_{(js)k}^2) - \sum_{k=1}^{K^B} \sum_{l=1}^{L_k} \frac{N_{(js)kl}}{N_j} \log \left(\frac{N_{(js)kl}}{N_j} \right) \right) \end{aligned}$$

N = banyaknya objek

N_j = jumlah objek di dalam gerombol j

N_{jkl} = jumlah objek di gerombol j untuk peubah kategorik ke- k dengan kategori ke- l

$\hat{\sigma}_k^2$ = ragam dugaan untuk peubah kontinu ke- k untuk keseluruhan objek

$\hat{\sigma}_{jk}^2$ = ragam dugaan untuk peubah kontinu ke- k untuk keseluruhan objek dalam gerombol j

K^A = banyaknya peubah kontinu

K^B = banyaknya peubah kategorik

L_k = banyaknya kategori untuk peubah kategorik ke- k

5. Jarak Chi-Square

Jarak *Chi-square* menghitung antara profil dalam jarak Euclidean terboboti menggunakan invers proporsi rata-rata sebagai bobot. Misal c_j menunjukkan elemen ke- j dari rata-rata profil, yang merupakan kelimpahan proporsi peubah j dalam seluruh kumpulan data. Maka jarak *Chi-Square*, dinotasikan χ , antara dua objek dengan profil $x = [x_1 x_2 \dots x_j]$ dan $y = [y_1 y_2 \dots y_j]$ didefinisikan sebagai [4]:

$$\chi_{x,y} = \sqrt{\sum_{j=1}^J \frac{1}{c_j} (x_j - y_j)^2}$$

3. Metode

Metode *two step cluster* adalah metode yang didesain untuk menangani jumlah objek yang besar, terutama pada masalah objek yang mempunyai peubah campuran, kontinu dan kategorik. Jarak antara dua gerombol didefinisikan sebagai jarak antar pusat dari masing-masing gerombol tersebut. Pusat dari suatu gerombol adalah vektor dari rata-rata masing-masing peubahnya. Jarak yang digunakan dalam metode *two step cluster* adalah jarak *Log-Likelihood* dan jarak Euclidean. Prosedur penggerombolan dengan metode *two step cluster* mempunyai dua tahapan yaitu tahap

preclustering (penggerombolan awal) objek ke dalam *subcluster-subcluster* kecil dan tahap penggerombolan akhir [6][2].

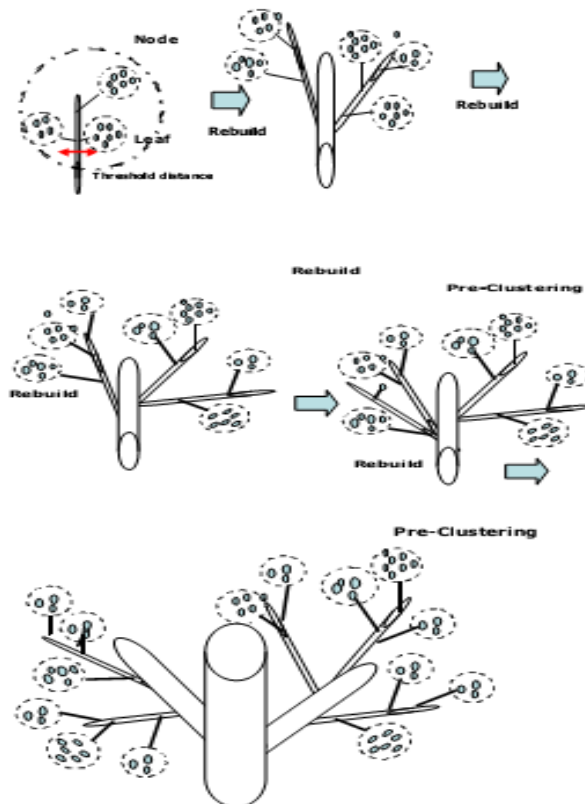
3.1. Penggerombolan Awal (*Preclustering*)

Menurut [5], tahap penggerombolan awal dilakukan dengan pendekatan sekuensial, yaitu objek diamati satu persatu berdasarkan ukuran jarak yang kemudian ditentukan apakah objek tersebut masuk dalam gerombol yang telah terbentuk atau harus membentuk gerombol baru. Pada langkah ini, diimplementasikan dengan pembentukan *Cluster Feature (CF) Tree*.

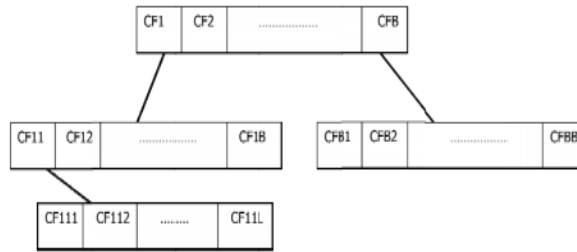
Definisi *CF-tree*

Diberikan N titik objek *d* dimensi pada suatu *cluster* dimana $i = 1, 2, \dots, N$. Vektor *clustering feature* dari *cluster* didefinisikan sebagai *quadriple*: $CF=(N,M,V,K)$ dimana N adalah banyaknya objek pada *cluster*, M menyatakan rata-rata dari peubah kontinu dari N objek, V adalah variansi dari setiap peubah kontinu pada N objek, K adalah banyaknya taraf pada setiap peubah kategorik. *CF-tree* adalah keseimbangan tinggi pohon dengan dua parameter yaitu *branching factor* (B) dan *threshold* (T).

CF-tree terdiri dari beberapa tingkat cabang (nodes) dan masing-masing cabang berisikan individu objek (*entries*) dari gerombol awal. Tingkatan daun atau daun entri yang terdapat pada cabang merepresentasikan anak gerombol (*subcluster-subcluster*). Prosedur *CF-tree* dilakukan dengan memilih satu amatan awal secara acak yang akan diukur jaraknya satu persatu dengan amatan lainnya menggunakan ukuran jarak yang telah ditentukan. Jika besarnya jarak terletak pada daerah penerimaan (*threshold distance*), maka amatan akan menjadi anggota anak gerombol. Jika besarnya jarak terletak di luar wilayah daerah penerimaan, maka amatan tersebut masuk ke dalam gerombol yang telah dibentuk atau akan menjadi cikal bakal daun entri baru.



Gambar 1 Proses pembentukan *CF-tree*



Gambar 2 Contoh CF-tree

Jika dalam kasus data terdapat pencilan, maka ketika dibentuk CF-tree diperiksa apakah dapat dimasukkan dalam gerombol yang sudah terbentuk tanpa harus membentuk CF-tree baru. Untuk mendeteksi ada tidaknya pencilan maka dilakukan perhitungan jarak *Log-Likelihood*. Jika terdapat jarak terbesar antar gerombol yang melebihi titik kritis C , yaitu:

$$C = \log(V)$$

dengan:

$$V = \prod_k R_k \prod_m L_m$$

dan

R_k = range dari peubah kontinu ke- k

L_m = banyaknya kategori untuk peubah kategorik ke- m .

Pada jarak Euclidean, data yang memuat pencilan memiliki prosedur yang sama dengan jarak *Log-Likelihood*. Dikatakan pencilan jika jarak Euclidean terbesar antara gerombol tersebut lebih besar dari titik kritis C , dengan rumus C sebagai berikut :

$$C = 2 \left(\sum_{i=1}^{K^A} \frac{\hat{\sigma}_{kl}^2}{K^A} \right)^{\frac{1}{2}}$$

dengan :

R_k = range dari peubah kontinu ke- k

K^A = banyaknya peubah kontinu

$\hat{\sigma}_{kl}^2$ = ragam dugaan untuk peubah kontinu ke- l dalam gerombol k .

Pembentukan CF-tree terdiri dari dua tahapan. Tahap pertama yaitu tahap penyisipan (*inserting*) dan tahapan yang kedua adalah tahap pembentukan kembali (*rebuilding*). Pada tahap *inserting*, secara random dipilih satu objek lalu diukur jaraknya dengan objek yang lain. Jika jarak tersebut kurang dari jarak maksimum, maka objek tersebut dimasukkan ke dalam satu gerombol. Tetapi jika jarak tersebut melebihi jarak maksimum, maka objek tersebut dianggap pencilan dan begitu seterusnya untuk objek selanjutnya.

Dari objek yang dianggap pencilan tersebut, akan dibuat suatu gerombol yang baru. Tahap ini merupakan tahap *rebuilding*. Jika CF-tree semakin banyak hingga melewati batas ukuran maksimum yang telah ditetapkan, maka batas jarak maksimum harus ditingkatkan sehingga dapat memasukkan lebih banyak objek. Peningkatan jarak ini dapat mengakibatkan objek-objek yang tadinya berasal dari gerombol yang berbeda bergabung menjadi satu gerombol CF, sehingga menghasilkan CF-tree yang berukuran lebih kecil dari semula.

3.2. Penggerombolan Akhir

Pada langkah ini, hasil dari CF-tree digerombolkan dengan analisis gerombol berhierarki dengan metode *agglomerative*. Untuk menghitung banyaknya gerombol dapat dilakukan dengan dua tahapan, yang pertama adalah menghitung *Schwarz's Bayesian Criterion* (BIC) atau *Akaike's Information Criterion* (AIC) untuk tiap gerombol. Rumus BIC dan AIC untuk gerombol J adalah sebagai berikut:

$$BIC(J) = -2 \sum_{j=1}^J \xi_j + m_j \log(N)$$

$$AIC(J) = -2 \sum_{j=1}^J \xi_j + 2m_j$$

dengan:

$$m_j = J \left\{ 2K^A + \sum_{k=1}^{K^B} (L_k - 1) \right\}$$

$$\xi_j = -N \left(\sum_{k=1}^{K^A} \frac{1}{2} \log(\hat{\sigma}_k^2 + \hat{\sigma}_{jk}^2) \sum_{k=1}^{K^B} \sum_{l=1}^{L_k} \frac{N_{jkl}}{N_j} \log \left(\frac{N_{jkl}}{N_j} \right) \right).$$

Solusi banyaknya gerombol yang optimal adalah yang memiliki nilai BIC terkecil, tetapi ada beberapa kasus dalam penggerombolan dimana BIC akan terus meningkat nilainya bila jumlah gerombol semakin meningkat. Maka dalam situasi tersebut, *ratio BIC changes* (rasio perubahan BIC) dan *ratio of distance measure changes* (rasio perubahan jarak) digunakan untuk mengidentifikasi solusi banyaknya gerombol optimal. Solusi untuk banyaknya gerombol optimal akan memiliki *ratio BIC changes* dan *ratio distance measure* yang besar.

Jumlah gerombol yang terbentuk dapat diketahui dengan menggunakan perbandingan antar jarak untuk k gerombol, dengan rumus perbandingannya sebagai berikut:

$$R(k) = d_{k-1}/d_k$$

$$d_k = l_{k-1} - l_k$$

dengan:

$$l_v = (r_v \log n - BIC_v)/2, \text{ atau}$$

$$l_v = (2r_v - AIC_v)/2$$

$$v = k, k - 1$$

$$d_{k-1} = \text{jarak jika } k \text{ gerombol digabungkan dengan } k - 1 \text{ gerombol}$$

$$R(k) = \text{rasio perubahan jarak.}$$

Penentuan peubah yang berpengaruh adalah dengan melihat nilai t dan *chi-square* dengan rumus sebagai berikut

$$t = \frac{\hat{\mu}_k - \hat{\mu}_{jk}}{\hat{\sigma}_{jk}/\sqrt{N_k}}$$

dengan:

$\hat{\mu}_k$: rata-rata peubah numeric ke- k

$\hat{\mu}_{jk}$: rata-rata peubah numeric ke- k kelompok ke- j

$\hat{\sigma}_{jk}$: simpangan baku peubah numeric ke- k pada kelompok ke- j

N_k : jumlah observasi pada peubah numeric ke- k

$$\chi^2 = \sum_{l=1}^{l_k} \left(\frac{N_{jkl} - N_{kl}}{N_{kl}} \right)^2$$

dengan:

N_{jkl} : jumlah observasi pada gerombol ke- j untuk peubah kategorik ke- k dengan kategori ke- l

N_{kl} : jumlah observasi pada gerombol untuk peubah kategorik ke- k dengan kategori ke- l

l_k : jumlah kategori untuk peubah kategorik ke- k

4. Hasil dan Pembahasan

Sebagai ilustrasi untuk analisis gerombol dengan metode *two step cluster*, digunakan data harga dan spesifikasi berbagai model kendaraan dari berbagai perusahaan (Sumber: SPSS 17 for Windows). Data yang digunakan terdiri dari peubah-peubah yang bertipe campuran (kategorik dan kontinu). Peubah-peubah yang digunakan disajikan pada tabel 1 berikut.

Tabel 1 Peubah-peubah yang digunakan dalam analisis gerombol

Peubah	Keterangan
X_1	Tipe (0: Mobil, 1: Truk)
X_2	Harga
X_3	Ukuran mesin
X_4	Tenaga kuda
X_5	Jarak roda
X_6	Lebar
X_7	Panjang
X_8	Berat
X_9	Kapasitas bahan bakar
X_{10}	Efisiensi bahan bakar

Dari hasil analisis dengan menggunakan *software* SPSS 17 for Windows, diperoleh banyaknya gerombol yang dapat terbentuk dengan *auto-clustering* adalah sebanyak 15 gerombol. Nilai BIC, BIC Change, Ratio of BIC Changes, dan Ratio of Distance Measures untuk banyaknya gerombol yang dapat terbentuk seperti dalam tabel 2.

Tabel 2 Nilai BIC dari banyaknya gerombol yang dapat terbentuk

Banyaknya Gerombol	Schwarz's Bayesian Criterion (BIC)	BIC Change	Ratio of BIC Changes	Ratio of Distance Measures
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	.367	2.190
4	897.559	11.635	-.048	1.368
5	931.760	34.201	-.142	1.036
6	968.073	36.313	-.151	1.576
7	1026.000	57.927	-.241	1.083
8	1086.815	60.815	-.253	1.687
9	1161.740	74.926	-.312	1.020
10	1237.063	75.323	-.313	1.239
11	1316.271	79.207	-.330	1.046
12	1396.192	79.921	-.333	1.075
13	1477.199	81.008	-.337	1.076
14	1559.230	82.030	-.341	1.301
15	1644.366	85.136	-.354	1.044

Penentuan banyaknya gerombol akhir yang terbentuk dilihat dari nilai *ratio of distance measure* yang terbesar. Nilai *ratio of distance measure* yang terbesar terdapat pada banyaknya gerombol 3. Oleh karena itu banyaknya gerombol akhir optimal yang terbentuk adalah 3 gerombol. Distribusi anggota dari masing-masing gerombol yang terbentuk dapat dilihat pada Tabel 3.

Tabel 3 Distribusi hasil penggerombolan

Gerombol	N	Persentase dari Total
1	62	39.5%
2	39	24.8%
3	51	32.5%

dengan karakteristik untuk masing-masing gerombol seperti dalam Tabel 4. Dari Tabel 4 dan Tabel 5, dapat diambil kesimpulan yaitu:

- a. Kendaraan yang termasuk ke dalam gerombol 1 adalah kendaraan dengan harga yang murah, ukuran mesin, ukuran kendaraan, dan kapasitas bahan bakar yang kecil, serta efisiensi bahan bakar yang besar. Anggota gerombol 1 sebanyak 62 kendaraan terdiri dari 61 mobil dan 1 truk.
- b. Kendaraan yang termasuk ke dalam gerombol 2 adalah kendaraan dengan harga menengah, dengan bobot kendaraan yang besar, serta kapasitas bahan bakar besar, tetapi efisiensi bahan bakar rendah. Semua anggota gerombol 2 sebanyak 39 kendaraan merupakan tipe truk.
- c. Kendaraan yang termasuk ke dalam gerombol 3 adalah kendaraan dengan harga mahal, ukuran mesin dan ukuran kendaraan besar, kapasitas bahan bakar dan efisiensi bahan bakar sedang. Semua anggota gerombol 3 sebanyak 51 kendaraan merupakan tipe mobil.

Dalam gambar 3 diagram batang yang menunjukkan persentase anggota masing-masing gerombol berdasarkan peubah kategorik yaitu tipe kendaraan.

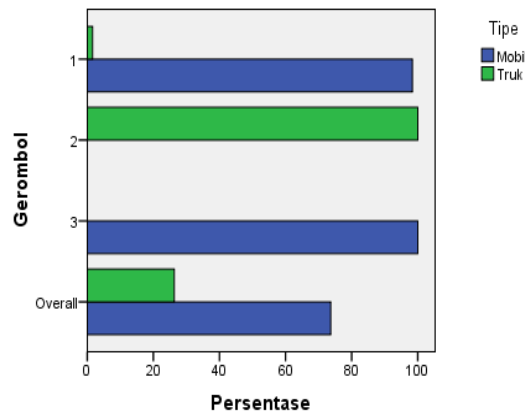
Tabel 4 Karakteristik gerombol

		Gerombol		
		1	2	3
Harga (dalam ribuan dollar)	Mean	19.616	26.561	37.299
	Std. Deviasi	7.6440	10.185	17.381
Ukuran mesin	Mean	2.194	3.559	3.700
	Std. Deviasi	.4238	.9358	.9493
Tenaga kuda	Mean	143.24	187.92	232.96
	Std. Deviasi	30.259	39.049	54.408
Jarak roda	Mean	102.59	112.97	109.02
	Std. Deviasi	4.0799	9.6537	5.7644
Lebar	Mean	68.539	72.744	72.924
	Std. Deviasi	1.9366	4.1781	2.1855
Panjang	Mean	178.23	191.11	194.68
	Std. Deviasi	9.6534	14.441	10.351
Berat	Mean	2.8374	3.9675	3.5789
	Std. Deviasi	.31087	.67176	.29720
Kapasitas bahan bakar	Mean	14.979	22.064	18.443
	Std. Deviasi	1.8699	4.2894	2.0445
Efisiensi bahan bakar	Mean	27.24	19.51	23.02
	Std. Deviasi	3.578	2.910	2.060

Tabel 5 Frekuensi untuk masing-masing gerombol berdasarkan peubah kategorik (tipe)

Gerombol	Mobil		Truk	
	Frekuensi	Persentase	Frekuensi	Persentase
1	61	54.5%	1	2.5%
2	0	.0%	39	97.5%
3	51	45.5%	0	.0%

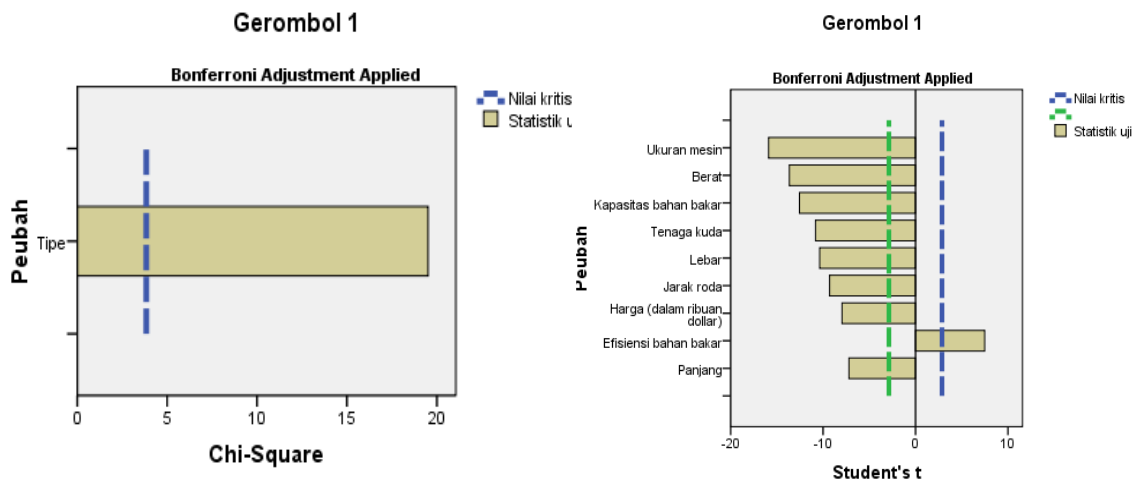
Persentase anggota gerombol berdasarkan tipe kendaraan



Gambar 3 Persentase anggota gerombol berdasarkan tipe kendaraan

Hasil uji *t-Student's* dan uji *Chi-Square* untuk melihat peubah-peubah yang berpengaruh pada pembentukan gerombol adalah sebagai berikut.

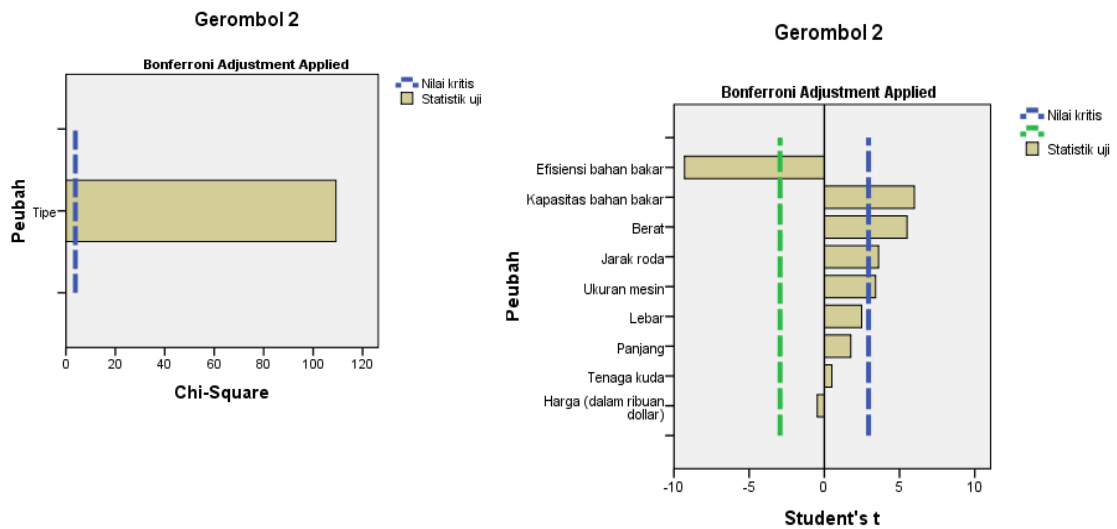
a. Gerombol 1



Gambar 4 Hasil uji peubah kontinu dan peubah kategorik gerombol 1

Hasil uji *t-Student's* dan uji *Chi-Square* untuk gerombol 1 dalam Gambar 4 menunjukkan bahwa peubah kategorik yaitu tipe kendaraan dan semua peubah kontinu berpengaruh terhadap pembentukan gerombol.

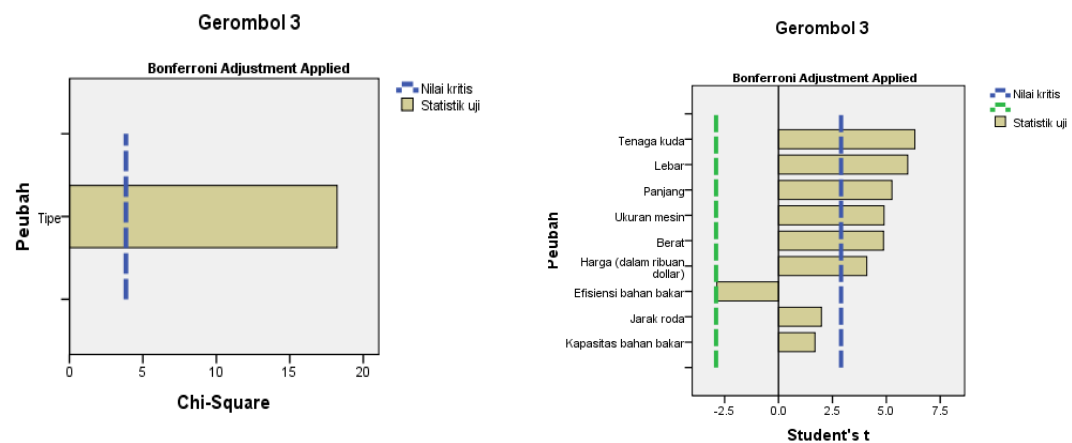
b. Gerombol 2



Gambar 5 Hasil uji peubah kontinu dan peubah kategorik gerombol 2

Hasil uji *t-Student's* dan uji *Chi-Square* untuk gerombol 2 dalam Gambar 5 menunjukkan bahwa peubah kategorik yaitu tipe kendaraan berpengaruh terhadap pembentukan gerombol sedangkan peubah kontinu yang berpengaruh terhadap pembentukan gerombol yaitu efisiensi bahan bakar, kapasitas bahan bakar, berat, jarak roda, dan ukuran mesin.

c. Gerombol 3



Gambar 6 Hasil uji peubah kontinu dan peubah kategorik gerombol 3

Hasil uji *t-Student's* dan uji *Chi-Square* untuk gerombol 3 dalam Gambar 6 menunjukkan bahwa peubah kategorik yaitu tipe kendaraan berpengaruh terhadap pembentukan gerombol sedangkan peubah kontinu yang berpengaruh terhadap pembentukan gerombol yaitu tenaga kuda, lebar, panjang, ukuran mesin, berat, harga, dan efisiensi bahan bakar.

5. Kesimpulan

Analisis *two step clustering* mengelompokkan menjadi 3 gerombol model kendaraan. Peubah yang berpengaruh terhadap pembentukan gerombol pada Gerombol 1 adalah semua peubah kategorik dan kontinu. Gerombol 2 yang berpengaruh adalah 6 peubah yaitu 1 peubah kategorik dan 5 peubah kontinu. Sedangkan untuk gerombol 3 peubah yang berpengaruh ada 8 yaitu 1 peubah kategorik dan 7 peubah kontinu.

6. Daftar Pustaka

- [1] Anderberg MR. 1973. *Cluster Analysis for Application*. Academic Press, New York.
- [2] Bacher J, Wenzig K, Vogler M. 2004. *SPSS Two Step Cluster - A First Evaluation*. [http://www.statisticalinnovations.com/products/Two Step.pdf](http://www.statisticalinnovations.com/products/TwoStep.pdf).
- [3] Dillon WR, and M. Goldstein. 1984. *Multivariate Analysis Method and Applications*. John Wiley & Sons. Canada.
- [4] Michael, <http://www.econ.upf.edu/~michael/stanford/maeb4.pdf>
- [5] SPSS Inc. (2001). *The SPSS TwoStep Cluster Component*. A scalable component to segment your customers more effectively. White paper – technical report, Chicago. <ftp://ftp.spss.com/pub/web/wp/TSCWP-0101.pdf>.
- [6] SPSS Inc. (2004). *TwoStep Cluster Analysis*. Technical report, Chicago. [http://support.spss.com/tech/stat/Algorithms/12.0/twostep cluster.pdf](http://support.spss.com/tech/stat/Algorithms/12.0/twostepcluster.pdf).