Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

Pembangunan Sistem Geodemografi di Pulau Pinang: Proses Pemilihan Variabel dengan Menggunakan Analisis Komponen Utama (PCA)

Development of Geodemographic System in Penang: Variable Process Selection by Using the Principal Component Analysis (PCA)

Kamarul Ismail[1]*, Siti Naeilah Ibrahim[1] & Ruslan Rainis[2]
[1]Department of Geography and Environment, Sultan Idris Education University, 35900 Tanjong Malim, Perak, Malaysia
[2]Geography Division, Centre for Humanities Studies, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia
*[1]*Corresponding author: kamarul.ismail@fsk.upsi.edu.my*

Abstrak

Secara umum istilah geodemografi boleh ditafsirkan sebagai kajian mengenai manusia dan hubungan dengan lokasi tempat mereka tinggal. Salah satu daripada tumpuan kajian dalam bidang ini ialah pengkelasan kawasan yang melibatkan komponen-komponen utama seperti data digital, perlombongan data dan paparan dalam persekitaran sistem maklumat geografi. Sumber utama data digital bagi tujuan pembangunan sistem geodemografi ialah data banci yang melibatkan pengumpulan data secara komprehensif maklumat demografi penduduk di sesebuah kawasan. Sebagai contoh, pangkalan data banci penduduk dan perumahan yang dibangunkan oleh Jabatan Perangkaan Malaysia pada tahun 2000 mengandungi lebih daripada 190 variabel banci yang boleh digunakan sebagai input untuk membangunkan sistem ini. Walau bagaimanapun jumlah variabel yang besar ini tidak boleh digunakan sebagai input untuk membangunkan sistem pengkelasan disebabkan oleh pelbagai masalah. Oleh itu, proses pemilihan variabel perlu dijalankan terlebih dahulu bagi memastikan variabel yang digunakan dalam pembentukan kluster tidak mengandungi maklumat yang berulangan. Salah satu kaedah pemilihan variabel yang boleh digunakan ialah dengan menjalankan analisis komponen utama (PCA) yang membahagikan data-data banci ini dalam beberapa kumpulan kecil. Selain daripada membahagikan variabel dalam komponen berasingan, analisis PCA juga boleh digunakan untuk memilih variabel individu berdasarkan kepada nilai Eigen yang dihasilkan.

Kata kunci *Sistem klasifikasi, geodemografi, data banci, Analisis Komponen Utama (PCA)*

Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

Abstract

In general, geodemographics can be defined as the study of people ang its realtion to where they live. One of the thrusts in this field is area classification which involves major components such as digital data, data mining and geographic information system. The main source of digital data for the development of geodemographics system is census data which involves comprehensive data collection of demographic information in a particular area. For example, the data base of population and housing census developed by the Department of Statistics in 2000, has more than 190 variables that can be used as inputs to develop this classification system. However, this large amount of variables can not be used as input to develop a classification system due to various problems. Thus, variable selection process has been carried out in advance to ensure the variables used in the cluster formation do not contain repeated information. One of the variable selection methods that can be used is the Principal Component Analysis (PCA), which divides the census data into small groups. Apart from dividing variables into separate components, the PCA analysis can also be used to select individual variables based on Eigen value produced.

Keyword *Classification system, geodemographic, census data, Principal Component Analysis (PCA)*

Pengenalan

Sistem pengkelasan geodemografi menurut Sleight (1997), Vickers (2006), dan Vickers dan Rees (2007) merupakan salah satu daripada tema kecil bidang kajian pengkelasan kawasan. Bidang kajian ini mengkelaskan setiap kawasan geografi kepada beberapa kumpulan yang berasingan berasaskan kepada persamaan ciri-ciri penduduk yang tinggal dalam sesebuah kawasan. Istilah geodemografi merupakan gabungan daripada dua bidang kajian iaitu geografi dan demografi. Salah satu daripada tumpuan dalam bidang geografi ialah kajian mengenai aktiviti manusia di atas permukaan bumi. Manakala bidang demografi pula mengkaji secara saintifik mengenai penduduk di sesebuah kawasan terutamanya dalam aspek saiz, taburan, struktur umur, jantina, aspek sosioekonomi, kadar kelahiran, kematian dan migrasi (O'Malley *et al.*, 1995). Oleh kerana bidang kajian ini melibatkan data-data berkaitan dengan ciri-ciri penduduk dan tempat mereka tinggal, sumber data utama yang digunakan untuk membangunkan sistem pengkelasan geodemografi ialah data banci penduduk dan perumahan negara.

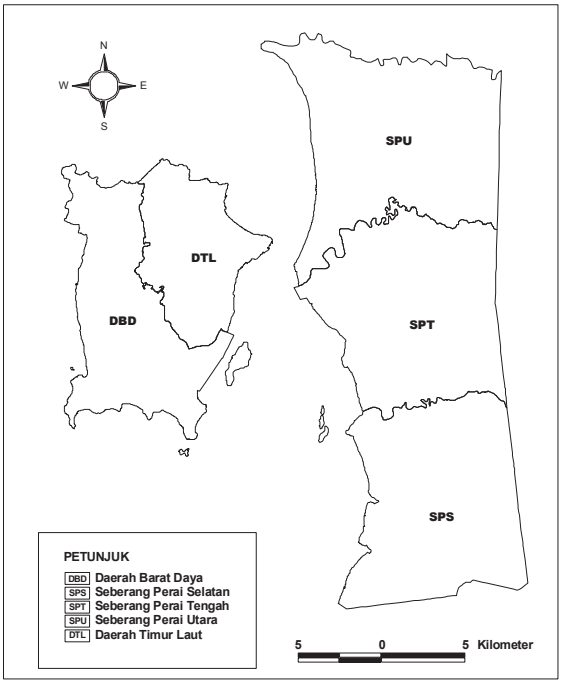
Pernyataan Masalah

Proses pengutipan data banci dijalankan secara berterusan bagi setiap sepuluh tahun merupakan kumpulan data yang paling komprehensif yang direkodkan di Malaysia. Proses merekod dan membangunkan pangkalan data banci ini melibatkan implikasi kewangan yang sangat besar. Banci pada tahun 2000 dan 2010 sebagai contoh masing-masing melibatkan kos sejumlah RM198 juta dan RM200 juta. Oleh itu adalah tidak wajar data-data yang dikutip ini hanya disimpan dalam pangkalan data dan tidak dimanfaatkan sepenuhnya kerana peruntukan untuk menjalankan banci ini melibatkan

kos yang sangat besar. Kajian pembangunan sistem geodemografi ini bukan sahaja bertujuan untuk menyusun data banci secara sistematik, mencerap maklumat yang terlindung melalui proses perlombongan data tetapi kajian ini juga bertepatan dengan polisi penyebaran data yang dilaksanakan oleh Jabatan Perangkaan Malaysia (JPM).

Kawasan Kajian

Pulau Pinang terbahagi kepada lima daerah pentadbiran yang berbeza iaitu daerah Timur Laut, daerah Barat Daya, daerah Seberang Prai Utara, daerah Seberang Prai Tengah dan daerah Seberang Prai Selatan. Kedudukan koordinat bagi negeri Pulau Pinang adalah di antara 5°8' Utaraan hingga 5°35' Utaraan dan 100°8' Timuran hingga 100°32' Timuran. Berdasarkan kepada Banci Penduduk dan Perumahan Tahun 2000, jumlah penduduk di Pulau Pinang adalah seramai 1.3 juta orang. Luas keseluruhan negeri Pulau Pinang ialah 1,043 kilometer persegi dengan kepadatan penduduk seramai 1,180 orang bagi setiap kilometer persegi.



Rajah 1 Peta kawasan kajian

Semakan literatur

Banci penduduk dan perumahan di negara ini telah mula dikutip sejak dari awal abad ke 20. Banci terperinci pertama yang dilakukan pada tahun 1891 yang meliputi beberapa buah negeri seperti Pulau Pinang, Melaka, Singapura, Perak, Pahang, Negeri Sembilan dan Selangor. Banci kedua pula dilakukan pada tahun 1901 dan

Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

diikuti dengan banci ketiga pula dijalankan pada tahun 1911. Banci tahun 1931, 1947 dan 1957 pula dijalankan oleh kerajaan British. Banci tahun 1957 merupakan banci terakhir dijalankan oleh kerajaan British sebelum Persekutuan Tanah Melayu mencapai kemerdekaan pada 31 Ogos 1957. Selepas penubuhan negara Malaysia pada tahun 1963, kerajaan Malaysia telah menjalankan banci pertama pada tahun 1970 dan diikuti dengan banci tahun 1980, 1991 dan tahun 2000. Banci tahun 2000 merupakan banci penduduk dan perumahan keempat yang dijalankan oleh kerajaan Malaysia (Abdul Rahman dan Norfariza, 2007). Data banci yang digunakan dalam kajian ini secara keseluruhannya diperolehi daripada data banci penduduk dan perumahan Malaysia tahun 2000. Selain daripada maklumat asas mengenai sesebuah kawsan, data banci juga mempunyai pelbagai maklumat lain seperti demografi, komposisi isi rumah, maklumat perumahan, ciri-ciri sosioekonomi, pekerjaan dan sektor industri. Abdul Rahman dan Norfariza (2007) menjelaskan bahawa beberapa kriteria perlu diikuti bagi menentukan bidang yang bersesuaian untuk dijadikan sebagai soalan banci. Soalan yang dikemukakan perlu mengambilkira kesesuaian keperluan pengguna yang pelbagai, mempunyai perbandingan secara maksimum dengan negara lain, maklumat yang boleh dimanfaatkan oleh masyarakat umum dan kemampuan kewangan untuk menjalankan banci.

Jadual 1 Perbandingan soalan banci tahun 1970, 1980, 1991 dan 2000

		1970	1980	1991	2000
1.	Tempat individu ditemui pada hari banci	x	x	x	-
2.	Tempat tinggal biasa pada hari banci	-	-	x	x
3.	Jantina	x	x	x	x
4.	Umur	x	x	x	x
5.	Tarikh lahir	x	x	x	x
6.	Status perkahwinan	x	x	x	x
7.	Bangsa	x	x	x	x
8.	Agama	x	x	x	x
9.	Status warganegara	x	x	x	x
10.	Warna kad pengenalan	x	x	-	-
11.	Dialek	x	x	-	-
12.	Kecacatan	x	x	-	x
13.	Bilangan kelahiran hidup kanak-kanak	x	x	-	x
14.	Bilangan kanak-kanak hidup	x	x	-	x
15.	Umur pada perkahwinan pertama	-	x	-	-
16.	Bilangan perkahwinan	x	x	-	-
17.	Tahun perkahwinan	x	-	-	-
18.	Tempat lahir	x	x	x	x
19.	Tempoh masa tinggal di Malaysia	x	x	-	-
20.	Tempoh masa tinggal di lokasi semasa	x	x	-	-
21.	Tempat tinggal terdahulu	-	x	-	-

Jadual 1 (*samb.*)

22.	Alasan bermigrasi	-	x	-	-
23.	Tempat tinggal lima tahun lepas	-	-	x	x
24.	Tahun sampai ke Malaysia	x	-	-	x
25.	Celik huruf	x	x	-	x
26.	Bersekolah	x	x	x	x
27.	Peringkat pendidikan tertinggi	x	x	x	x
28.	Kelayakan tertinggi diterima	x	x	x	x
29.	Latihan vokasional	-	x	-	x
30.	Bidang pendidikan	-	-	-	x
31.	Tempat belajar	-	-	-	x
32.	Jenis aktiviti ekonomi (dalam tempoh satu minggu terakhir)	x	x	x	x
33.	Bilangan jam berkerja (dalam tempoh satu minggu terakhir)	-	-	x	-
34.	Jenis aktiviti ekonomi (dalam tempoh satu tahun terakhir)	x	x	-	-
35.	Pekerjaan	x	x	x	x
36.	Industri	x	x	x	x
37.	Status pekerjaan	x	x	x	x
38.	Sektor	-	-	-	x
39.	Pertalian dengan ketua isi rumah	x	x	x	x
40.	Bilangan penghuni	x	x	x	x
41.	Jenis tempat tinggal	x	-	-	-
42.	Sewa (berperabot/tidak berperabot)	x	-	-	x
43.	Minyak masak utama	x	-	-	-
44.	Kelengkapan isi rumah	x	x	x	x
45.	Pendapatan isi rumah	x	x	-	-
46.	Lokasi tempat tinggal	x	x	x	x
47.	Jenis tempat tinggal	x	x	x	x
48.	Bahan binaan asas	x	-	-	-
49.	Bahan binaan dinding luar	x	x	x	x
50.	Bahan binaan untuk atap	x	x	-	-
51.	Keadaan kediaman	x	x	-	-
52.	Tahun pembinaan	x	x	x	-
53.	Jenis pemilikan	x	x	x	x
54.	Jenis bekalan air	x	x	x	x
55.	Jenis lampu	x	x	x	x
56.	Jenis kemudahan tandas	x	x	x	x
57.	Kemudahan bilik mandi	x	x	-	-
58.	Kemudahan memasak	x	x	-	-
59.	Bilangan bilik tidur	x	-	x	x

Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

Jadual 1 (*samb.*)

60.	Kemudahan pengutipan sampah	-	-	-	x
61.	Bilangan penghuni	x	x	x	x
62.	Bilangan isi rumah	x	x	x	x

(Sumber: Abdul Rahman dan Norfariza, 2007)

Perbandingan pemilihan soalan yang dimasukkan dalam borang banci pada tahun 1970, 1980, 1991 dan 2000 adalah seperti yang ditunjukkan dalam Jadual 1. Berdasarkan kepada jadual yang ditunjukkan, soalan-soalan utama yang dimasukkan dalam borang banci dibahagikan kepada tiga bahagian, iaitu maklumat populasi yang merekodkan tentang ciri-ciri geografi, ciri-ciri demografi dan sosial, kadar kelahiran dan kematian, ciri-ciri migrasi dan ciri-ciri pendidikan, maklumat mengenai isi rumah yang mengumpul data mengenai ciri-ciri isi rumah dan maklumat mengenai perumahan. Walaubagaimanapun, terdapat beberapa perbezaan soalan yang dimasukkan dalam borang banci. Sebagai contoh soalan mengenai kecacatan, celik huruf, jenis aktiviti ekonomi dan kemudahan kutipan sampah yang tidak dimasukkan dalam banci tahun 1991 dimasukkan dalam soalan banci tahun 2000.

Melalui penelitian yang dilakukan terhadap sektor-sektor utama yang terdapat dalam data banci, dapat disimpulkan di sini bahawa maklumat-maklumat yang diperolehi ini adalah sangat penting untuk melaksanakan sebarang bentuk pembangunan. Data-data banci boleh digunakan untuk mengeluarkan laporan, penemuan serta merangka polisi bagi kerajaan tempatan dan peringkat kebangsaan. Selain itu variabel-variabel banci yang merangkumi pelbagai skop soalan juga boleh digunakan untuk membuat keputusan sama ada untuk membuka atau menutup kemudahan-kemudahan seperti sekolah, hospital dan klinik (Boyle dan Dorling, 2004). Walaubagaimanapun, dalam proses pembangunan sistem pengekelasan geodemografi (SPG) tidak semua variabel-variabel banci ini sesuai digunakan. Dramowicz (2004) menegaskan bahawa penggunaan variabel yang berlebihan perlu dielakkan kerana ia memberi kesan kepada ketepatan kluster yang dihasilkan. Fowlkes dan Mallows (1983) merujuk variabel yang tidak relevan ini dengan menggunakan istilah *masking variable*, ia merupakan masalah utama kepada analisis gunaan kerana variabel ini menghalang pencarian struktur kluster sebenar dalam pangkalan data dan secara tidak langsung menyebabkan hasil analisis menjadi tidak tepat.

Milligan (1996) juga menyokong pernyataan ini dengan menyatakan bahawa variabel hanya boleh digunakan dalam pembangunan sistem sekiranya ia mempunyai alasan yang kuat untuk dimasukkan. Memasukkan variabel yang tidak relevan dalam analisis kluster hanya akan menimbulkan masalah untuk mencari sturktur kluster yang sebenar. Vickers (2006) pula menyatakan bahawa matlamat utama pemilihan variabel dilakukan adalah untuk memilih seminimum mungkin variabel yang boleh mewakili keseluruhan dimensi yang terdapat dalam sesuatu set data. Analisis kluster bergantung sepenuhnya kepada variabel yang digunakan sebagai input, algoritma analisis kluster itu sendiri sebenarnya tidak boleh menentukan sama ada variabel yang dimasukkan itu relevan atau tidak. Oleh itu prosedur pemilihan variabel yang rapi perlu dilakukan. Ini bertepatan dengan pandangan Bailey et al. (2000) dan Vickers dan Rees (2007)

yang menyatakan bahawa matlamat utama pemilihan variabel dilakukan adalah untuk membolehkan penggunaan variabel paling minimum bagi mewakili keseluruhan dimensi data banci untuk mengutip sebanyak mungkin maklumat daripada variabel-variabel tersebut.

Analisis PCA sering disamakan dengan analisis faktor, walaubagaimanapun terdapat perbezaan yang ketara antara kedua-dua analisis ini. Sebagai contoh, hasil analisis PCA sebagai adalah dalam bentuk komponen, manakala analisis faktor pula menghasilkan faktor. Selain itu, PCA mengira semua jumlah variabel yang terdapat dalam kumpulan data, manakala analisis faktor pula hanya mengira varian yang dikongsi dalam sesuatu kumpulan data (Gaur dan Gaur, 2009). Penentuan saiz komponen dalam analisis PCA pula dilakukan dengan mengira hubungan linear antara variabel yang terdapat dalam matrik korelasi (R-matrix). Secara teori, saiz komponen dalam matrik korelasi adalah bersamaan dengan jumlah keseluruhan variabel yang digunakan dalam analisis. Namun begitu, sebahagian daripada komponen tersebut yang mempunyai nilai eigen kecil adalah tidak sesuai digunakan dalam sistem pengkelasan. Nilai eigen yang dikira dalam analisis PCA menerangkan jumlah varian yang terdapat dalam sesuatu komponen. Lebih besar nilai eigen sesuatu komponen, lebih banyak varian yang diterangkan oleh komponen tersebut.

Metodologi

Berdasarkan kepada pernyataan dalam bahagian sebelum ini, kertas kerja ini mengemukakan beberapa kaedah pengurangan data yang terdapat dalam analisis komponen utama (PCA). Dalam bahagian metodologi kajian ini, penerangan secara ringkas proses-proses yang berlaku semasa pemilihan data, penggunaan kaedah pemerhatian secara grafik, menguji kesesuaian kaedah kriteria kaiser dan akhirnya menentukan kaedah pemilihan variabel yang paling sesuai digunakan untuk membangunkan SPG dijalankan secara berperingkat.

Kaedah Scree Plot

Kaedah pemerhatian secara grafik yang dikemukakan oleh Cattell (1966) merupakan kaedah pertama yang digunakan untuk menentukan bilangan komponen yang dikehendaki. Kaedah ini memaparkan graf bagi setiap nilai eigen (paksi Y) dan jumlah faktor atau komponen (paksi X). Graf ini menurut Field (2009) dikenali sebagai scree plot yang merupakan adaptasi istilah daripada bidang geologi untuk menggambarkan permukaan batu yang mempunyai pecahan-pecahan batu kecil didasarnya. Plot yang dihasilkan mempunyai kecerunan yang tinggi di permulaan dan mendatar pada bahagian yang lain. Ini kerana hanya sebilangan komponen awal sahaja yang mempunyai nilai eigen yang tinggi, manakala komponen-komponen lain mempunyai nilai yang lebih rendah.

Kaedah Kriteria Kaiser

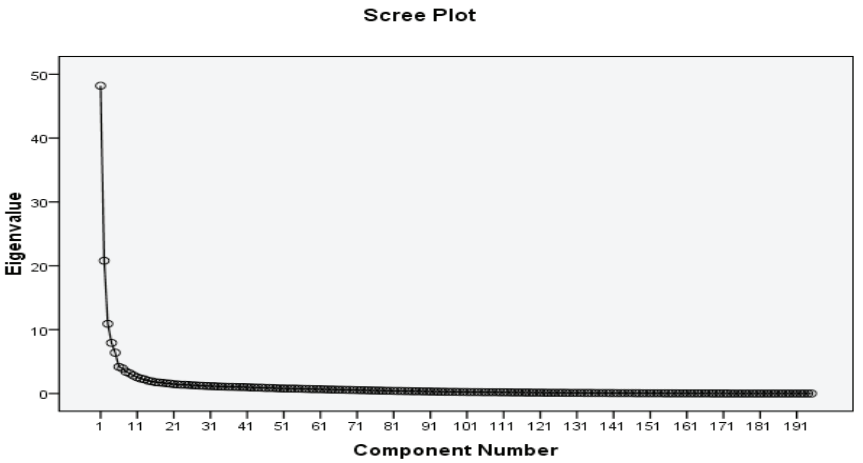
Kriteria Kaiser yang diperkenalkan oleh Kaiser (1960) menganggap komponen yang mempunyai niai eigen lebih kecil daripada satu menunjukkan bahawa maklumat yang terdapat dalam variabel tersebut adalah lebih rendah berbanding dengan variabel-

Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

variabel lain. Oleh itu, komponen-komponen yang mempunyai nilai eigen lebih daripada satu dikekalkan, manakala komponen yang mempunyai nilai eigen yang kecil daripada satu pula dikeluarkan daripada analisis. Namun begitu, Jolliffe (1972) menyatakan bahawa kaedah yang diperkenalkan oleh Kaiser untuk mengehadkan penentuan jumlah komponen dengan hanya menggunakan nilai eigen yang lebih daripada satu merupakan proses pemilihan yang terbatas. Untuk mendapatkan maklumat maksimum dalam sesuatu pangkalan data, beliau mencadangkan supaya komponen yang mempunyai nilai eigen lebih daripada 0.7 turut dikekalkan. Walaubagaimanapun, Field (2009) menyatakan bahawa kaedah yang diperkenalkan oleh Joliffe (1972) ini jarang digunakan dalam kajian kerana kaedah ini menyebabkan penghasilan komponen yang berlebihan.

Analisis Data

Analisis data dengan menggunakan kaedah scree plot seperti yang ditunjukkan dalam rajah 2 adalah merupakan hasilan yang diperolehi daripada analisis PCA dengan menggunakan perisian SPSS. Dalam rajah tersebut, nilai eigen tertinggi ditunjukkan oleh komponen 1 dan diikuti dengan niai eigen bagi komponen 2 dan 3. Sekiranya kaedah penentuan titik peralihan seperti yang dikemukakan oleh Cattell (1966) ini digunakan sebagai asas penetapan jumlah komponen, hanya sejumlah 6 komponen sahaja yang boleh digunakan.



Rajah 2 Scree plot

Seperti yang ditunjukkan dalam Jadual 2, hanya 46.8% jumlah varian dikekalkan berbanding dengan jumlah keseluruhan varian dalam kumpulan data sebenar. Peratus jumlah varian yang kecil ini menunjukkan bahawa sejumlah 53.2% maklumat yang terdapat dalam kumpulan data secara automatik. Ini bertepatan dengan kajian Field (2009) yang menyatakan bahawa walaupun kaedah pemerhatian secara grafik ini sangat berguna dalam proses pemilihan variabel, ia tidak boleh dijadikan sebagai kaedah utama dalam proses penentuan jumlah komponen disebabkan oleh kesukaran untuk

menetapkan titik peralihan. Selain itu, kaedah ini juga hanya mengekalkan sebahagian kecil varian yang terdapat dalam kumpulan data asal.

Jadual 2 Penentuan saiz komponen dengan menggunakan scree plot

Komponen	Nilai Eigen Awalan			Jumlah Putaran Memuatkan Kuasa Dua		
	Jumlah	Varian (%)	Kumulatif (%)	Jumlah	Varian (%)	Kumulatif (%)
1	48.193	24.715	24.715	42.929	22.015	22.015
2	20.798	10.666	35.380	17.487	8.968	30.983
3	10.910	5.595	40.975	12.400	6.359	37.342
4	7.926	4.065	45.040	8.899	4.564	41.905
5	6.390	3.277	48.317	5.526	2.834	44.739
6	4.176	2.142	50.458	4.041	2.072	46.812

Analisis PCA yang dijalankan terhadap 195 variabel berpotensi dengan menggunakan kaedah kriteria Kaiser pula menghasilkan sejumlah 40 komponen yang mewakili sejumlah 79.4% peratus jumlah varian daripada keseluruhan kumpulan data. Seperti yang ditunjukkan dalam Jadual 3, nilai muatan komponen 1 adalah sebanyak 42.93 unit yang mewakili jumlah varian tertinggi iaitu 22%. Ini diikuti oleh komponen 2 dengan jumlah varian sebanyak 8.97% dan komponen 3 dengan muatan sebanyak 12.40 unit atau bersamaan dengan 6.36% jumlah varian. Peratusan jumlah varian yang ditunjukkan dalam hasilan Jumlah Putaran Memuatkan Kuasa Dua ini seterusnya semakin mengecil dan susut pada nilai 0.58% dalam komponen ke 40. Berdasarkan kepada analisis dalam jadual tersebut, secara umumnya dapat disimpulkan bahawa setiap komponen mempunyai jumlah varian berbeza dan hanya beberapa komponen utama sahaja yang mempunyai majoriti jumlah varian. Komponen 1 contohnya mempunyai jumlah varian tertinggi, diikuti dengan komponen 2 dan komponen 3.

Jadual 3 Saiz komponen dengan menggunakan kaedah kriteria Kaiser

Komponen	**Nilai Eigen Awalan**			**Jumlah Putaran Memuatkan Kuasa Dua**		
	Jumlah	**Varian (%)**	**Kumulatif (%)**	**Jumlah**	**Varian (%)**	**Kumulatif (%)**
1	48.19	24.71	24.71	42.93	22.01	22.01
2	20.80	10.67	35.38	17.49	8.97	30.98
3	10.91	5.59	40.98	12.40	6.36	37.34
4	7.93	4.06	45.04	8.90	4.56	41.91
5	6.39	3.28	48.32	5.53	2.83	44.74
6	4.18	2.14	50.46	4.04	2.07	46.81
7	3.92	2.01	52.47	3.86	1.98	48.79

Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

Jadual 3 (*samb.*)

8	3.39	1.74	54.21	3.58	1.84	50.63
9	3.14	1.61	55.82	2.99	1.53	52.16
10	2.82	1.45	57.27	2.45	1.26	53.42
11	2.53	1.30	58.56	2.40	1.23	54.65
12	2.33	1.20	59.76	2.34	1.20	55.85
13	2.21	1.13	60.89	2.22	1.14	56.99
14	1.99	1.02	61.91	2.19	1.12	58.11
15	1.90	0.98	62.89	2.07	1.06	59.18
16	1.77	0.91	63.80	2.06	1.05	60.23
17	1.72	0.88	64.68	2.05	1.05	61.28
18	1.66	0.85	65.53	2.00	1.02	62.31
19	1.60	0.82	66.35	1.96	1.00	63.31
20	1.55	0.79	67.15	1.93	0.99	64.30
21	1.46	0.75	67.90	1.86	0.96	65.26
22	1.43	0.73	68.63	1.80	0.92	66.18
23	1.38	0.71	69.34	1.75	0.90	67.08
24	1.37	0.70	70.04	1.69	0.87	67.94
25	1.34	0.69	70.72	1.67	0.86	68.80
26	1.29	0.66	71.39	1.67	0.85	69.65
27	1.27	0.65	72.04	1.60	0.82	70.48
28	1.23	0.63	72.67	1.56	0.80	71.28
29	1.21	0.62	73.29	1.54	0.79	72.07
30	1.18	0.60	73.89	1.44	0.74	72.81
31	1.14	0.59	74.48	1.44	0.74	73.54
32	1.14	0.58	75.06	1.40	0.72	74.26
33	1.10	0.57	75.63	1.35	0.69	74.95
34	1.09	0.56	76.19	1.30	0.67	75.62
35	1.07	0.55	76.73	1.28	0.66	76.27
36	1.05	0.54	77.27	1.25	0.64	76.91
37	1.05	0.54	77.81	1.25	0.64	77.55
38	1.02	0.52	78.33	1.22	0.62	78.18
39	1.01	0.52	78.85	1.18	0.60	78.78
40	1.00	0.51	79.36	1.13	0.58	79.36

Variabel-variabel yang terdapat dalam komponen-komponen awal menurut Vickers (2006) adalah merupakan variabel yang penting. Variabel-variabel ini memiliki jumlah varian yang besar dan mempunyai pengaruh yang kuat dalam pembentukan sistem pengkelasan. Manakala variabel-variabel dalam komponen yang lain pula tidak memainkan peranan yang penting dalam pembangunan sistem pengkelasan kerana mempunyai jumlah varian yang kecil. Variabel-variabel ini perlu dikeluarkan daripada

analisis kerana ia tidak memberi sumbangan yang besar terhadap keseluruhan sistem pengkelasan.

Jadual 4 menunjukkan senarai 52 variabel yang termasuk dalam komponen pertama yang merupakan komponen yang paling penting kerana mempunyai jumlah varian terbesar berbanding dengan komponen-komponen yang lain. Antara variabel-variabel penting yang termasuk dalam komponen ini ialah variabel v101, v131, v138 dan v140 yang masing-masing mempunyai jumlah muatan sebanyak 0.96. Nilai muatan yang tinggi ini menunjukkan bahawa korelasi antara variabel-variabel tersebut dengan komponen satu adalah tinggi, nilai muatan 0.96 ini mewakili sejumlah 92% ((0.96^2) x 100) varian variabel v101, v131, v138 dan v140 ditunjukkan dalam komponen tersebut. Ini menunjukkan bahawa variabel-variabel ini mempunyai pengaruh yang kuat dalam komponen pertama dan seterusnya memberi kesan yang kuat terhadap penghasilan kluster dalam proses pembangunan sistem pengkelasan.

Jadual 4 Senarai variabel dalam komponen pertama

No	Nama Variabel (Komponen 1)	Muatan	No	Nama Variabel (Komponen 1)	Muatan
v131	Didiami	0.96	v082	Jumlah Isirumah 6 orang	0.84
v140	Elektrik 24 jam sehari	0.96	v049	Sekolah Menengah Rendah	0.83
v138	Air paip yang dirawat	0.96	v048	Sekolah Rendah	0.83
v101	Televisyen	0.96	v080	Jumlah Isirumah 4 orang	0.83
v098	Peti Sejuk	0.95	v057	Sijil PMR	0.83
v100	Radio / Hi-fi	0.94	v103	Telefon Talian Tetap	0.82
v027	Berkahwin	0.94	v047	Pra Sekolah	0.79
v181	Pertalian Suami / isteri	0.94	v012	Penduduk berumur 40 hingga 44 tahun	0.78
v093	1 Motosikal	0.93	v050	Sekolah Menegah Atas	0.78
v039	Warga Warganegara Malaysia	0.93	v079	Jumlah Isirumah 3 orang	0.78
v002	Jantina Lelaki	0.92	v102	Video / VCD / DVD	0.77
v001	Jumlah Penduduk	0.92	v056	Tiada Sijil	0.73
v081	Jumlah Isirumah 5 orang	0.92	v083	Jumlah Isirumah 7 orang	0.70
v097	Mesin Basuh	0.92	v009	Penduduk berumur 25 hingga 29 tahun	0.70
v182	Pertalian Anak belum kahwin	0.91	v095	Basikal	0.69
v005	Penduduk berumur 05 hingga 09 tahun	0.91	v013	Penduduk berumur 45 hingga 49 tahun	0.68
v180	Pertalian Ketua IR	0.89	v030	Agama Islam	0.66
v003	Jantina Perempuan	0.88	v058	Sijil SPM	0.66
v004	Penduduk berumur 00 hingga 04 tahun	0.87	v020	Etnik Bangsa Melayu	0.66

Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

Jadual 4 (*samb.*)

v006	Penduduk berumur 10 hingga 14 tahun	0.87	v190	Etnik ketua isirumah Melayu	0.66
v011	Penduduk berumur 35 hingga 39 tahun	0.86	v154	Kerja Juruteknik & Separa Profesional	0.61
v133	Milik Individu	0.86	v170	Pengangkutan, Penyimpanan & Perhubungan	0.58
v026	Belum Pernah Berkahwin	0.86	v084	Jumlah Isirumah 8 orang	0.58
v046	Tidak Pernah Bersekolah	0.85	v014	Penduduk berumur 50 hingga 54 tahun	0.57
v010	Penduduk berumur 30 hingga 34 tahun	0.84	v155	Pekerja Perkeranian	0.55
v090	1 Motokar	0.84	v078	Jumlah Isirumah 2 orang	0.52

Perbincangan

Berdasarkan kepada analisis yang dijalankan terhadap sejumlah 195 variabel banci yang berpotensi dimasukkan sebagai input dalam proses pembangunan sistem geodemografi, proses pemilihan variabel dengan menggunakan analisis PCA sepenuhnya tidak mampu untuk menyelesaikan masalah pengurangan data. Analisis PCA dengan menggunakan scree plot menunjukkan bahawa kaedah pemerhatian secara grafik yang dikemukakan oleh Cattell (1966) ini gagal untuk menyelesaikan masalah pengurangan data disebakan oleh kesukaran untuk menentukan titik peralihan bagi penentuan jumlah komponen. Selain itu nilai varian bagi 6 komponen yang ditetapkan hanya sebanyak 46.8% berbanding jumlah varian secara keseluruhan.

Untuk mengatasi masalah jumlah varian yang kecil dengan menggunakan kaedah scree plot, analisis PCA dengan menggunakan kaedah kriteria kaiser dilakukan terhadap kumpulan data yang diperolehi daripada JPM ini. Hasil daripada analisis yang dijalankan menunjukkan bahawa, kaedah kriteria kaiser ini berjaya mengekalkan 76.4% jumlah varian berbanding dengan jumlah keseluruhan (100%) varian yang terdapat dalam kumpulan data. Namun begitu, daripada 40 komponen yang dihasilkan jumlah varian terbesar hanya tertumpu dalam komponen-komponen awal. Sebagai contoh, 22% varian terdapat dalam komponen, sejumlah 8.97% varian terdapat dalam komponen 2 dan 6.36% varian terdapat dalam Komponen 3. Secara keseluruhan, jumlah varian yang terdapat dalam komponen-komponen utama ini mewakili sejumlah 48.4% berbanding dengan jumlah varian yang terdapat dalam komponen-komponen lain.

Jumlah varian yang tinggi dalam komponen utama menunjukkan bahawa variabel-variabel yang dominan dikelompokkan dalam komponen yang sama. Sebagai contoh, komponen 1 yang dihasilkan dengan menggunakan kaedah kriteria kaiser mengandungi 52 variabel yang berkorelasi dan mengulang maklumat yang sama. Variabel-variabel seperti, v131, v140, v138 dan v101 merupakan antara variabel yang mempunyai

jumlah varian tertinggi dalam Komponen 1. Namun begitu, hubungan korelasi antara variabel-variabel ini juga adalah tinggi. Untuk menyelesaikan masalah ini, variabel-variabel yang terdapat dalam setiap komponen ini perlu dianalisis sekali lagi dengan menggunakan analisis pekali korelasi perason. Selain daripada membuang salah satu daripada pasangan variabel yang berkorelasi tinggi, variabel umur dan jumlah isi rumah juga digabungkan bagi menghasilkan variabel baru.

Kesimpulan

Walaupun penggunaan komponen yang dihasilkan dengan menggunakan analisis PCA juga boleh digunakan sebagai input dalam analisis kluster, kaedah pemilihan variabel yang lebih sesuai dilaksanakan seperti yang disarankan oleh Vickers et. al (2003) ialah dengan melakukan pemilihan terhadap variabel asal. Penelitian terhadap hasil analisis daripada variabel asal adalah lebih mudah berbanding dengan pengunaan komponen. Analisis PCA memainkan peranan yang penting dalam proses pemilihan variabel, variabel-variabel dengan nilai muatan yang tinggi adalah lebih berpotensi untuk dipilih dan digunakan sebagai input dalam analisis kluster. Walaubagaimanapun, penentuan kekuatan pengaruh variabel dengan mengggunakan analisis PCA sahaja tidak boleh menyelesaikan masalah ini, terutamanya apabila wujud hubungan berbagai (*multicolinearity*) antara variabel.

Rujukan

Abdul Rahman Hassan & Norfariza Hanim Kasim. (2007). Malaysian Population Census: Review of Enumeration Strategies and Topics. *Journal of Department of Statistic Malaysia 1*(2007), 51–60.

Bailey, S., Charlton, J., Dollamore, G., and Fitzpartick, J. (2000). Families, Groups and Clusters of Local and Health Authorities: Revised for Authorities in 1999. *Population Trends*, *99,* 37–52.

Boyle, P. and Dorling, D. (2004). Guest Editorial: the 2001 UK Census: Remarkable Resource or Bygone Legacy of The 'Pencil And Paper Era'?. *Area, 36(2),* 101–110.

Cattell, R.B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research, 1*, 245–76.

Dramowics, K. (2004, November). A Primer on How to Create a Customized Segmentation System. *Directions Magazine. Diakses daripada h*ttp://www.directionsmag.com/articles/a-primer-on-how-to-create-a-customized-segmentation-system/123592. pada 24 Jun 2008.

Field, A. (2009). *Discovering Statistics using SPSS (And Sex and Drugs and Rock'n'roll)*. London: Thousand Oaks.

Fowlkes, E. B., and Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association, 78,* 553–569.

Gaur, A. S. and Gaur S. S. (2009). Statistical Methods For Practice And Research: A Guide To Data Analysis Using SPSS. Edisi Kedua. Place??:Sage Publications

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I. *Artificial DataJournal of the Royal Statistical Society.* Series C (Applied Statistics),Vol. 21, No. 2 (1972), 160–173.

Jurnal Perspektif Jil. 5 Bil. 1 (83-96)
ISSN 1985 - 496X

Miligan, G. W. (1996). Clustering validation: Results and implications for applied analyses, in Arabie, P., Hubert, L. J. dan De Soete, G. (eds), *Clustering and Classification*, World Scientific, Singapore.

O'Malley, L., Patterson, M and Evans, M., (1995). Retailing Applications of Geodemographics: A Preliminary Investigation. *Marketing Intelligence & Planning*, Vol. 13 Iss: 2, pp.29 – 35.

Sleight, P. (1997). *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*. Edisi Kedua. Thames: NTC Publications Ltd.

United Nations. (2007). Principles and Recommendations for Population and Housing Censuses. Revision 2. New York: Publisher???.

Vickers, D. (2006). Multi-level Integrated Classifications Based on the 2001 Census (Unpublished thesis). Department of Geography, University of Leeds.

Vickers, D. and Rees, P. (2007). Creating the UK national statistics 2001 output area classification. *Journal of the Royal Statistical Society, Series A 170,*379–403.

Vickers, D.W., Rees, P.H. and Birkin, M. (2003). A New Classification of UK Local Authorities Using 2001 Census Key Statistics. Working Paper 03/3, School of Geography, University of Leeds, Leeds diaksess daripada http://www.geog.leeds.ac.uk/wpapers/03-3.pdf.