Analisis Pemilihan Pemboleh Ubah dalam Proses Pembentukan Sistem Geodemografi

Selection Analysis of Variables to Produces Geodemography System

Kamarul Ismail, Mohamad Suhaily Yusri Che Ngah, Yazid Saleh, Nasir Nayan, Mohmadisa Hashim & Siti Naielah Ibrahim

Jabatan Geografi dan Alam Sekitar, Fakulti Sains Kemanusiaan, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak.
e-mel: kamarul.ismail@fsk.upsi.edu.my

Abstrak

Sistem pengkelasan geodemografi (SPG) merujuk kepada proses pengkelasan penduduk berdasarkan kepada kepada lokasi atau tempat mereka tinggal. Pernyataan ini merujuk kepada kecenderungan manusia yang lebih selesa untuk bergaul, membentuk komuniti dan tinggal dalam kawasan penduduk yang mempunyai persamaan dengan mereka. Konsep kajian ini mempunyai persamaan dengan hukum pertama geografi yang dikemukakan oleh Tobler pada tahun 1970, iaitu semua pemboleh ubah adalah berkaitan dengan pemboleh ubah yang lain, tetapi pemboleh ubah yang lebih dekat mempunyai hubungan yang lebih kuat berbanding dengan pemboleh ubah yang lebih jauh. Data banci penduduk dan perumahan Malaysia adalah merupakan pangkalan data yang paling komprehensif yang dikutip secara berterusan sejak dari tahun 1870 sehingga tahun 2010. Analisis pemilihan data dilakukan kerana bilangan pemboleh ubah yang banyak akan menyebabkan berlakunya pengulangan maklumat yang sama. Melalui kajian ini, analisis pemilihan data seperti analisis komponen utama (PCA), analisis korelasi dan analisis varian dijalankan bagi memastikan pemboleh ubah yang benar-benar perlu sahaja yang digunakan dalam membentuk sistem pengkelasan geodemografi di Perak. Sejumlah 179 pemboleh ubah dalam data banci penduduk dan perumahan negeri Perak pada tahun 2000 digunakan untuk membangunkan sistem pengkelasan geodemografi (SPG) dan daripada analisis yang dilakukan, hanya 69 pemboleh ubah berpotensi yang dipilih untuk dijalankan analisis pengklusteran. Pengurangan jumlah pemboleh ubah melalui analisis pemilihan pemboleh ubah akan memudahkan proses analisis kluster dilakukan.

Kata kunci Geodemografi, Korelasi Pearson, Kriteria Kaiser, Nilai Eigen

Abstract

Geodemography classification system (SPG) is refers to the classification of the population according to the location or where they live. This statement refers to the tendency of people who are more comfortable to mix, forming communities and residents who live in the area have in common with them. The concept of this study are similar to the first law of geography presented by Tobler in 1970, that all things are related to other things, but the things that are closer have a stronger relationship than with more distant things. Data population and housing census is the most comprehensive database of the collected continuously from 1870 until 2010. The analysis of data selection is done because many of the variables that will cause the repetition of the same information. Through this study, analysis of data selection such as principal component analysis (PCA), correlation analysis and analysis of variance carried out to ensure that the variables that really should be only used in forming the classification system geodemografi in Perak. 179 variables in a population and housing census data Perak in 2000 to develop a classification system geodemography (SPG) and from the analysis, only 69 potential variables selected to run the cluster analysis. Reducing the number of variables in the analysis variable selection will simplify the process of cluster analysis was performed.

Keywords Geodemograhy, Pearson Correlation, Kaiser Criteria, The Eigen Values

PENGENALAN

Sistem geodemografi merujuk kepada suatu sistem yang mengkelaskan penduduk kepada lokasi atau tempat tinggal mereka berdasarkan ciri sosioekonomi yang sama. Menurut Harris *et al.* (2005), Vickers (2006) dan Vickers dan Rees (2007), sistem geodemografi merupakan salah satu daripada bidang kecil kajian pengkelasan kawasan. Goss (1995) pula menjelaskan sistem geodemografi merupakan satu sistem pengkelasan yang menggunakan teknologi maklumat bagi memudahkan para peniaga meramal tindak balas perlakuan pengguna berasaskan kepada model statistik berkaitan dengan identiti dan lokasi tempat tinggal mereka. Istilah geodemografi merupakan gabungan daripada 2 bidang kajian iaitu geografi dan demografi. Geografi merupakan kajian tentang ciri semulajadi permukaan bumi termasuklah iklim, tanah, tumbuhan dan aktiviti manusia, terhadap alam (*Kamus Dewan Edisi Keempat*) manakala bidang geodemografi pula merupakan kajian saintifik mengenai populasi di sesebuah kawasan seperti ciri penduduk, saiz, taburan geografi, struktur umur, jantina, aspek sosioekonomi, kadar kelahiran, kematian dan migrasi. Oleh itu, data utama yang digunakan dalam kajian ini adalah data banci penduduk dan perumahan negara kerana data tersebut mengandungi maklumat berkaitan ciri-ciri penduduk dan tempat tinggal mereka.

Pernyataan Masalah

Dewasa ini, para pengguna dan penyelidik tidak lagi menghadapi masalah kekurangan data, sebaliknya mereka berhadapan dengan masalah data yang berlebihan. Keadaan ini jelas ditunjukkan melalui kajian yang dijalankan oleh Naisbitt melalui pernyataan bahawa kita dilimpahi oleh data tetapi kekurangan maklumat untuk difahami pada tahun 1982. Pernyataan ini jelas membuktikan masalah sebenar yang sedang dihadapi oleh pengguna bukan ketiadaan data tetapi jumlah data yang semakin bertambah sehingga menimbulkan kesukaran untuk memanfaatkan data-data tersebut (Larose, 2005).

Data banci Penduduk dan Perumahan Negara merupakan sumber data utama dalam proses pembentukan sistem geodemografi bersifat rencam dan dikutip secara berterusan dalam selang masa yang telah ditetapkan. Bagi negara-negara maju, proses pengutipan data banci dilaksanakan dalam tempoh masa lima tahun, manakala di negara-negara sedang membangun kerja-kerja merekodkan maklumat mengenai penduduk ini dijalankan setiap sepuluh tahun. Data banci penduduk dan perumahan mengandungi pelbagai maklumat mengenai ciri-ciri populasi seperti umur, jantina, status perkahwinan, pendapatan dan pendidikan. Data-data ini disimpan dalam pangkalan data banci dan merupakan sumber data yang lengkap dan paling komprehensif bagi kebanyakan negara yang sedang membangun (Smith, 2003).

Proses pengutipan data secara berterusan akan menyukarkan para pengguna kerana bilangan data semakin bertambah setiap kali banci dilakukan dan pertambahan ini akan menyukarkan pengguna seandainya data itu tidak dianalisis terlebih dahulu. Hal ini kerana, data tersebut akan disesuaikan dengan maklumat penduduk bagi memastikan ketepatan maklumat penduduk di suatu kawasan. Bagi mengatasi masalah ini, kaedah analisis yang boleh meringkaskan maklumat banci tersebut perlu dilakukan kepada bentuk yang lebih ringkas agar boleh memudahkan capaian kepada pengguna dan mudah untuk difahami. Oleh itu, pengkelasan sistem geodemografi dengan menggunakan algoritma pengklusteran dalam penyelidikan ini merupakan kaedah yang berkesan untuk mengurangkan jumlah data dan menyusun data tersebut dalam kumpulan kecil yang lebih mudah difahami dan diuruskan. Di dalam konteks kajian ini, sejumlah 179 pemboleh ubah berpotensi dan 1,884 unit zon banci digunakan sebagai input dalam pembentukan sistem ini.

Vickers (2006) menyatakan bahawa masalah utama untuk menganalisis data banci penduduk dan perumahan adalah disebabkan oleh kerencaman data banci itu sendiri kerana jumlah data banci yang besar dan mengandungi berbagai maklumat menyebabkan ianya sukar untuk difahami. Oleh itu, salah satu kaedah yang boleh digunakan untuk mengubah data yang kompleks itu menjadi sebuah data yang mudah difahami adalah melalui penghasilan sistem geodemografi. Dalam proses pembangunan sistem geodemografi , tidak semua pemboleh ubah sesuai digunakan. Dramowicz (2004) menegaskan bahawa penggunaan pemboleh ubah yang berlebihan perlu dielakkan kerana ianya akan memberikan kesan kepada ketepatan kluster yang dihasilkan. Fowlkes dan Mallows (1983) menyatakan pemboleh ubah yang tidak relevan ini sebagai pemboleh ubah tersembunyi kerana ianya merupakan masalah utama kepada analisis gunaan. Hal ini kerana pemboleh ubah ini akan menghalang pencarian struktur kluster sebenar dalam pangkalan data dan secara tidak langsung akan menyebabkan hasil analisis menjadi tidak tepat.

Semakan Literatur

Sejak awal abad ke-19, data banci penduduk dan perumahan di negara ini telah mula dikutip yang mana banci terperinci pertama dilakukan pada tahun 1871 yang meliputi beberapa buah negeri seperti Pulau Pinang, Melaka, dan Singapura dan data penduduk yang dikumpul oleh pihak British diterbitkan dalam *Strait Settlements* (Toru, 2006). Data banci penduduk bagi negeri Selangor, Perak, Sungai Ujung dan Pahang pula mula dibanci pada tahun 1981 dan ianya dibanci secara asing dari kawasan *Strait Settlements*. Banci mengenai penduduk diterbitkan pada tahun 1901 dan 1911 setelah penggabungan beberapa negeri di bawah Negeri Melayu Bersekutu pada tahun 1896. Pada zaman pemerintahan kerajaan British, banci penduduk telah dijalankan pada tahun 1931, 1947 dan 1957 oleh pihak kerajaan British. Selepas penubuhan negara Malaysia pada tahun 1963, kerajaan Malaysia telah menjalankan banci pertama pada tahun 1970 dan diikuti dengan banci pada tahun 1980, 1991, 2000 dan 2010.

Menurut Jabatan Perangkaan Malaysia (2001), data banci penduduk dan perumahan tahun 2000 digunakan sepenuhnya untuk menyediakan dan memantau Rancangan Malaysia Kelapan (RMK-8) yang bermula pada tahun 2001 hingga tahun 2005. Malahan, antara salah satu objektif utama banci dilakukan adalah untuk menyediakan maklumat terperinci berkaitan ciri-ciri demografi, sosial dan ekonomi pada peringkat kawasan statistik terkecil. Objektif banci juga cuba mencapai matlamat penggunaan data yang tidak terhad kepada pihak kerajaan sahaja, tetapi boleh dimanfaatkan oleh sektor-sektor lain seperti usahawan, penganalisis industri, ahli akademik dan ahli politik (Malaysia, 2001). Kesedaran untuk memanfaatkan sepenuhnya data banci dalam jumlah yang besar bukanlah perkara yang baharu kerana perkara ini telah diperkatakan sejak tahun 1980 lagi. Han dan Kamber (2006) menggambarkan keadaan masyarakat pada hari ini sebagai kaya dengan data tetapi pengetahuan berkurang yang merujuk kepada kerja pengutipan data yang berterusan sehingga menyebabkan pembangunan pangkalan data secara besar-besaran yang melampaui keupayaan manusia untuk menganalisis data tersebut.

Sistem pengelasan geodemografi (SPG) menurut Vickers dan Rees (2007) merupakan salah satu daripada bidang kecil kajian pengkelasan kawasan. Bidang kajian ini mengkelaskan setiap kawasan geografi kepada beberapa kumpulan yang berasingan berasaskan kepada persamaan ciri-ciri populasi yang tinggal dalam kawasan tersebut. Goss (1995) merujuk sistem pengkelasan ini sebagai satu sistem pengkelasan yang menggunakan teknologi maklumat bagi membolehkan peniaga meramalkan tindak balas perlakuan pengguna berasaskan kepada model statistik berkaitan dengan identiti dan lokasi tempat tinggal mereka. Namun begitu, penggunaan SPG tidak terhad dalam bidang

perniagaan semata-mata, tetapi sebaliknya boleh digunakan dalam pelbagai bidang yang lain (Harris *et al.,* 2005). Analisis kluster merupakan teknologi penting dalam pembentukan perlombongan data. Ianya akan membahagikan set data kepada beberapa kumpulan kluster untuk menunjukkan struktur semulajadi set data tersebut. Terdapat beberapa kaedah yang biasa digunakan untuk algoritma kluster seperti k-means, STING, CLIQUE, CURE, CLARA, ISODATA, CLARANS, BIRCH, GRIDCLUS dan DBSCAN (Khaled *et al.,*2009). SPG amat berguna dalam mengenalpasti kawasan jenayah di bandar Tshwane, Afrika Selatan setelah bandar tersebut telah dibelengu peningkatan pelbagai jenis jenayah. Hasil kajian yang dilakukan oleh Breetzke. dan Horn (2009) mendapati kawasan yang mengalami peningkatan jenayah akan dikecualikan untuk dijalankan pembangunan sosial dan ekonomi bagi mencegah kadar jenayah dan kawasan tersebut akan dihuni oleh rakyat Afrika yang berkulit hitam. Kaedah ini secara tidak langsung akan menyebabkan kerajaan yang memerintah menilai semula penguatkuasaan undang-undang dan mencari inisiatif pencegahan jenayah di bandar Tshwane. Dalam bidang kesihatan, SPG juga mampu membantu dalam mengenalpasti kawasan atau golongan yang mengalami ketidakseimbangan kesihatan. Gerdin *et al.* (2008) telah mengaplikasikan pengkelasan geodemografi dengan memaparkan corak ketidakseimbangan kesihatan dalam kalangan kanak-kanak Sweden yang berumur antara 4 hingga 10 tahun berdasarkan kepada taraf sosioekonomi mereka. Berdasarkan pengelasan yang telah dilakukan, didapati bahawa masalah obesiti di kalangan kanak-kanak Sweden tidak hanya tertumpu di kalangan kanak-kanak bandar tetapi ianya lebih menyeluruh bagi semua peringkat sosio ekonomi. Namun begitu, dari segi masalah karies gigi, didapati bahawa kanak-kanak Sweden yang tinggal di kawasan luar bandar lebih cenderung untuk mengalami masalah tersebut jika dibandingkan dengan kanak-kanak bandar.

Kawasan Kajian

Perak merupakan negeri kedua terbesar di Semenanjung Malaysia dengan keluasannya adalah kira-kira 21,035 kilometer persegi. Kedudukan koordinat bagi negeri Perak ialah di antara 5° 53' Utaraan hingga 3° 42' Utaraan dan 101° 40' Timuran hingga 100° 22' Timuran. Jumlah penduduk di negeri Perak berdasarkan Banci Penduduk dan Perumahan Malaysia 2000 adalah kira-kira 1.97 juta orang dengan bilangan penduduk lelaki dan perempuan adalah hampir seimbang iaitu 49.99 peratus lelaki dan 50.01 peratus perempuan. Terdapat sembilan daerah di negeri Perak iaitu Batang Padang, Manjung, Kinta, Kerian, Kuala Kangsar, Larut dan Matang, Hilir Perak, Ulu Perak dan Perak Tengah.

Rajah 1 Peta lokasi kajian

METODOLOGI

Kajian ini mengemukakan kaedah yang digunakan untuk analisis pemilihan pemboleh ubah bagi memastikan pemboleh ubah yang dipilih tidak mengulangi penindanan maklumat. Setelah data diperolehi, data tersebut perlu dibersihkan terlebih dahulu bagi mengelakkan kesilapan berlaku semasa melaksanakan prosedur perlombongan data yang akhirnya membawa kepada masalah ketidaktepatan hasilan (Gopalan dan Sivaselvan, 2009). Data-data tersebut akan dikurangkan dengan menggunakan analisis korelasi, analisis komponen utama (PCA) dan analisis varian menerusi perisian *Statistical Package for the Social Sciences* (SPSS) yang mana ianya dilaksanakan pada peringkat awal analisis kluster bagi memastikan pemboleh ubah yang benar-benar relevan sahaja dipilih dan digunakan dalam pembentukan pangkalan data.

a) Analisis korelasi

Analisis korelasi berfungsi untuk menyenaraikan pemboleh ubah yang berkolaborasi sesama sendiri mengikut hierarki. Pengiraan nilai pekali korelasi lazimnya dilakukan untuk mengukur kekuatan hubungan linear antara dua pemboleh ubah yang mana nilai pekali korelasi Pearson (r) adalah di antara nilai -1.00 hingga 1.00. Dengan menggunakan korelasi, pemboleh ubah yang paling tinggi korelasinya akan dapat dikesan. Malahan, kategori setiap pemboleh ubah turut dapat dikenal pasti. Agarwal dan Rao (2006) menjelaskan bahawa pemboleh ubah yang berkorelasi tinggi berkemungkinan mengulangi maklumat yang sama, oleh itu adalah penting supaya kumpulan pemboleh ubah tersebut diubah menjadi kumpulan pemboleh ubah baharu dalam proses membentuk

komponen asas. Pemboleh ubah yang mempunyai perhubungan yang ketara dan sangat tinggi akan dibuang bagi mengelakkan ianya bertindan. Petunjuk korelasi di antara dua pemboleh ubah akan merujuk kepada petunjuk Guilford sepertimana jadual 1.

Jadual 1 Petunjuk korelasi Guilford (1959)

Nilai Korelasi	Petunjuk	Perhubungan
Kurang 0.20	Korelasi sangat rendah	Perhubungan yang sangat lemah, hampir diabaikan
0.20 hingga 0.40	Korelasi rendah	Perhubungan yang jelas, tetapi kecil
0.40 hingga 0.70	Korelasi sederhana	Perhubungan yang sederhana besar
0.70 hingga 0.90	Korelasi tinggi	Perhubungan yang ketara
0.90 hingga 1.00	Korelasi sangat tinggi	Perhubungan yang sangat boleh dipercayai

Mengikut Guilford (1959), jika nilai korelasi melebihi 0.7071, salah satu pemboleh ubah data tersebut akan dibuang Pengurangan data perlu dilakukan kerana data yang terdapat di dalam pangkalan data sangat banyak dan mengambil masa yang lama untuk dianalisis. Kaedah pengurangan bertujuan untuk mengecilkan saiz kumpulan pemboleh ubah dan mendapatkan beberapa pemboleh ubah yang baru dengan menggunakan kaedah pengurangan dimensi dan pemilihan pemboleh ubah.

b) Analisis Komponen Utama (PCA)

Komponen asas merupakan kumpulan pemboleh ubah berkorelasi tinggi yang diubah menjadi kumpulan data baru yang tidak berhubungan sesama sendiri kerana pemboleh ubah yang berkorelasi tinggi akan mengulang maklumat yang sama (Agarwal dan Rao, 2006). Faktor atau komponen asas bagi setiap kumpulan data diperoleh dengan menjalankan analisis komponen utama (PCA) kerana kaedah ini secara umumnya akan menggabungkan pemboleh ubah-pemboleh ubah yang berkorelasi menjadi beberapa kumpulan atau komponen asas. Menurut Friendly (2008), komponen asas merupakan kombinasi linear jumlah pemberat yang diperolehi daripada pemboleh ubah asal. Skor komponen ini boleh digunakan dalam analisis bagi menggantikan kumpulan data yang asal seandainya komponen utama mewakili sebahagian besar daripada varian yang terdapat dalam data. Perwakilan matematik bagi mengira setiap komponen ini adalah seperti berikut:

$$\mathbf{C1=b_{11}(X_1) + b_{12}(X_2) + ...b_{1p}(X_p)}$$ Persamaan 1.1

(sumber: Friendly, 2008)

di mana,

C1 = skor subjek dalam komponen utama

b_{1p} = pekali atau pemberat pemboleh ubah *p* yang dikira, digunakan untuk membentuk komponen pertama

X_p = skor subjek pada pemboleh ubah *p*

Hair *et al.* (2009) menyatakan bahawa PCA mengira pekali korelasi dalam suatu kumpulan kecil data dan membentuk kumpulan pemboleh ubah yang saling berhubungan yang dikenali sebagai faktor atau komponen. Komponen yang mengandungi kumpulan pemboleh ubah yang saling berhubungan dianggap sebagai mewakili dimensi yang terdapat dalam kumpulan data tersebut. Kaedah analisis pelbagai pemboleh ubah (*multivariate*) yang melibatkan penggunaan pemboleh ubah dalam jumlah yang besar akan menyebabkan berlaku pertindihan maklumat (Hair *et al.*,2009). Keadaan ini menyebabkan pengkaji mencari kaedah yang sesuai untuk menguruskan pemboleh ubah-

pemboleh ubah tersebut seperti kaedah menggabungkan pemboleh ubah yang berkorelasi tinggi, melabelkan kumpulan pemboleh ubah dan membentuk pemboleh ubah komposit. Pernyataan pengkaji-pengkaji terdahulu seperti Everitt *et al.* (2001), Harris *et al.* (2005), Shepherd (2006) dan Vickers (2006) telah membuktikan bahawa PCA merupakan kaedah yang sangat sesuai digunakan dalam proses pemilihan pemboleh ubah dalam membangunkan sistem pengkelasan. PCA sangat berguna untuk menyelesaikan masalah pemilihan pemboleh ubah sekiranya analisis korelasi tidak mempu untuk mengatasi masalah tersebut.

Terdapat beberapa kaedah yang digunakan dalam analisis PCA antaranya adalah kaedah pemerhatian secara grafik dan kaedah kriteria Kaiser. Kaedah pemerhatian secara grafik dikemukan oleh Cattell (1996) merupakan kaedah pertama yang digunakan untuk menentukan bilangan komponen yang dikehendaki. Setiap nilai eigen akan dipaparkan pada paksi Y manakala jumlah faktor atau komponen akan dipaparkan pada paksi X. Menurut Field (2009), graf ini dikenali sebagai '*scree plot'* yang mana ianya merupakan istilah daripada bidang geologi yang menggambarkan permukaan batu yang mempunyai pecahan-pecahan batu kecil di dasarnya. Plot yang dihasilkan pada permulaannya mempunyai kecerunan yang tinggi dan seterusnya mendatar pada bahagian lain. Hal ini kerana, hanya sebilangan komponen sahaja yang mempunyai nilai eigen yang tinggi manakala lain-lain komponen mempunyai nilai eigen yang rendah.

Kaedah Kaiser telah diperkenalkan oleh Kaiser (1960) yang mana ianya menganggap komponen yang mempunyai nilai eigen yang lebih kecil daripada satu menunjukkan bahawa maklumat yang terdapat dalam pemboleh ubah tersebut adalah lebih rendah berbanding dengan pemboleh ubah yang lain. Oleh itu, komponen yang mempunyai nilai eigen yang lebih besar daripada satu akan dikekalkan manakala komponen yang mempunyai nilai eigen yang lebih kecil daripada satu akan dikeluarkan daripada analisis. Kaedah yang diperkenalkan oleh Kaiser tidak berapa dipersetujui oleh Jollifie (1972) kerana beliau berpendapat bahawa kaedah ini merupakan proses pemilihan yang terbatas kerana kaedah Kaiser mengehadkan penentuan jumlah komponen dengan hanya menggunakan nilai eigen yang lebih daripada satu. Oleh itu, beliau mencadangkan supaya komponen yang mempunyai nilai eigen lebih daripada 0.7 turut dikekalkan bagi mendapatkan maklumat maksimum dalam sesuatu pangkalan data. Walaubagaimanapun, menurut Field (2009) kaedah yang diperkenalkan oleh Jolliffe ini jarang digunakan dalam kajian kerana kaedah ini akan menyebabkan penghasilan komponen yang berlebihan.

Penggunaan PCA dalam proses pemilihan pemboleh ubah juga boleh dirujuk sebagai analisis data secara penerokaan yang melibatkan penentuan saiz varian yang terdapat dalam suatu komponen (Dennett dan Stillwell, 2009). Pemboleh ubah yang memiliki varian yang tinggi merupakan pemboleh ubah yang penting dalam komponen tersebut. Malahan menurut Everitt *et al.* (2001) pemboleh ubah yang mempunyai nilai varian yang tinggi merupakan pemboleh ubah penting yang perlu dimasukkan dalam analisis kluster. Jumlah varian yang terdapat dalam setiap komponen yang dihasilkan melalui analisis PCA ditunjukkan oleh nilai eigen iaitu nilai eigen yang lebih besar akan menunjukkan lebih banyak varian yang diterangkan oleh setiap komponen. Oleh itu, dalam menentukan jumlah komponen bagi pemboleh ubah banci, kaedah Kaiser dan paparan melalui *scree plot* sebagaimana yang dicadangkan oleh Field (2009) digunakan.

c) Analisis varian

Terdapat beberapa kaedah bagi mengira sisihan data antaranya adalah analisis varian. Daya tarikan (*interestingness*) bagi sisihan data akan menunjukkan jarak di antara min dan varian sama ada terletak berdekatan ataupun terserak. Nilai varian yang tinggi bermaksud daya tarikan pemboleh ubah dengan

nilai min adalah tinggi dan pemboleh ubah itu perlu dikekalkan (Vickers, 2006). Analisis varian digunakan untuk mengenalpasti keserakan skor-skor dalam satu taburan. Varian merupakan kuasa dua bagi nilai sisihan piawai dengan formula

$$\sigma^2 = \sum \frac{(x-\mu)^2}{N}$$

Persamaan 1.2

di mana,

μ - nilai purata
N - jumlah keseluruhan zon

Data primer iaitu data banci penduduk dan perumahan bagi negeri Perak akan dianalisis dengan menggunakan perisian SPSS. Data penduduk akan diselaraskan mengikut jantina, bangsa, etnik, pendidikan, status perkahwinan dan lain-lain. Analisis varian dijalankan pada peringkat awal dalam menjalankan analisis kluster bagi memastikan pemboleh ubah yang dipilih hanyalah pemboleh ubah yang relevan bagi memastikan data yang digunakan untuk pengkelasan geodemografi penduduk mengikut pemboleh ubah adalah tepat dan tidak bertindan. Pengkelasan ini akan dikelaskan mengikut zon banci yang dikeluarkan oleh Jabatan Perangkaan Malaysia.

Analisis Data

Vickers (2006) mencadangkan agar dijalankan analisis korelasi bagi memeriksa matrik korelasi setiap pemboleh ubah bagi mengeluarkan pemboleh ubah yang berkorelasi tinggi dalam kumpulan data bagi mengatasi masalah hubungan berbagai antara pemboleh ubah (*multicolinearity*). Pemboleh ubah yang berkorelasi akan mampu menunjukkan maklumat yang sama di antara satu sama lain dan ini akan menyebakan kesukaran dalam mencari corak sebenar dalam kumpulan data. Jadual 2 menunjukkan pemboleh ubah yang berkorelasi tinggi dan perlu dibuang salah satu pemboleh ubah tersebut iaitu melebihi nilai sebagaimana yang dicadangkan oleh Guilford (1959) iaitu 0.7071.

Jadual 2 Pemboleh ubah yang berkorelasi

	Pemboleh ubah		**Pemboleh ubah**	**Nilai korelasi**
v020	Melayu	v030	Agama Islam	0.99
v023	India	v032	Agama Hindu	0.99
v188	Bekalan air paip yang dirawat	v190	Elektrik dibekalkan 24 jam sehari	0.98
v025	Bukan warganegara	v044	Warga pekerja asing	0.98
v026	Belum pernah berkahwin	v039	Warganegara Malaysia	0.97
v132	Pertalian ketua isi rumah	v027	Berkahwin	0.97
v027	Berkahwin	v181	Kediaman didiami	0.97
v132	Pertalian ketua isi rumah	v181	Kediaman didiami	0.96
v027	Berkahwin	v039	Warganegara Malaysia	0.95
v132	Pertalian ketua isi rumah	v039	Warganegara Malaysia	0.94
v061	Sekolah Men Rendah (Peralihan / Ting 1-3)	v069	Sijil PMR / SRP / LCE	0.94
v027	Berkahwin	v152	Televisyen	0.94
v060	Sekolah Rendah (Darjah / Tahun 1-6)	v068	Tiada sijil	0.94

Selain daripada membuang salah satu pasangan pemboleh ubah yang berkorelasi tinggi, kaedah penggabungan pemboleh ubah juga boleh digunakan untuk menyelesaikan masalah pengurangan pemboleh ubah. Sebagai contoh, sejumlah 16 pemboleh ubah umur dalam kumpulan data asal boleh digabungkan kepada beberapa unit pemboleh ubah kecil dengan menggunakan analisis pekali korelasi.

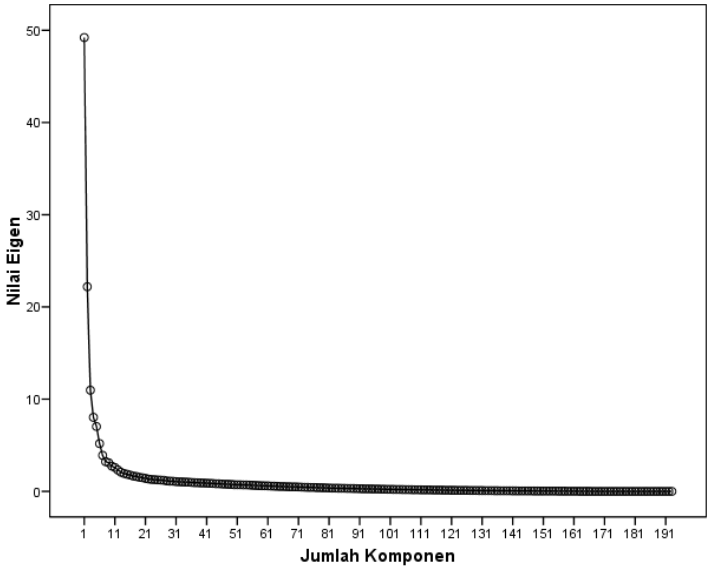
Jadual 3 Jadual korelasi pemboleh ubah umur

Umur	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74
00-04	1.00														
05-09	**0.91**	1.00													
10-14	**0.76**	**0.91**	1.00												
15-19	0.48	0.59	0.66	1.00											
20-24	0.37	0.34	0.32	0.66	1.00										
25-29	0.71	0.58	0.44	0.40	0.63	1.00									
30-34	**0.79**	**0.73**	0.56	0.41	0.47	**0.84**	1.00								
35-39	**0.74**	**0.80**	0.69	0.47	0.40	0.67	**0.86**	1.00							
40-44	0.60	**0.74**	**0.76**	0.58	0.38	0.54	0.68	**0.84**	1.00						
45-49	0.42	0.58	0.70	0.62	0.37	0.43	0.48	0.61	**0.82**	1.00					
50-54	0.27	0.41	0.54	0.51	0.30	0.35	0.34	0.40	0.58	**0.81**	1.00				
55-59	0.18	0.28	0.38	0.37	0.21	0.27	0.25	0.28	0.38	0.61	**0.82**	1.00			
60-64	0.13	0.21	0.31	0.28	0.13	0.18	0.17	0.18	0.27	0.45	0.65	**0.82**	1.00		
65-69	0.05	0.13	0.22	0.21	0.08	0.11	0.10	0.11	0.20	0.37	0.55	**0.73**	**0.85**	1.00	
70-74	0.01	0.10	0.20	0.18	0.05	0.04	0.02	0.05	0.15	0.30	0.47	0.62	**0.79**	**0.83**	1.00
> 75	-0.08	-0.01	0.10	0.11	0.02	0.00	-0.03	-0.01	0.10	0.28	0.47	0.62	**0.74**	**0.79**	**0.81**

Matrik korelasi pemboleh ubah umur seperti yang ditunjukkan dalam jadual 3 mengandungi sejumlah 16 pemboleh ubah yang mewakili struktur umur penduduk dalam kawasan kajian. Analisis korelasi yang dijalankan terhadap semua pemboleh ubah struktur umur tersebut menunjukkan terdapat hubungan korelasi yang kuat antara satu sama lain. Sebagai contoh pemboleh ubah v004 (umur 00 hingga 04 tahun) mempunyai hubungan yang sangat kuat dengan pemboleh ubah v005 (umur 05 hingga 09 tahun). Manakala pemboleh ubah v006 (umur 10 hingga 14 tahun) juga mempunyai hubungan korelasi yang tinggi dengan pemboleh ubah v004 (umur 00 hingga 04 tahun). Ketiga-tiga pemboleh ubah ini digabungkan menjadi pemboleh ubah komposit bagi mewakili penduduk yang berumur 00 hingga 14 tahun. Pemboleh ubah v009 (umur 25 hingga 29 tahun) turut digabungkan bersama-sama dengan pemboleh ubah v010 (umur 30 hingga 34 tahun) dan pemboleh ubah v011(umur 35 hingga 39 tahun) dan membentuk satu pemboleh ubah umur yang lain iaitu pemboleh ubah komposit yang mewakili penduduk berumur 25 hingga 39 tahun. Pemboleh ubah v007 (umur 15 hingga 19 tahun) tidak digabungkan dengan mana-mana pemboleh ubah kerana ianya tidak berkorelasi antara satu sama lain.

Analisis data dengan menggunakan kaedah *scree plot* adalah seperti yang ditunjukkan dalam rajah 2 yang merupakan hasilan daripada analisis PCA dengan menggunakan perisian SPSS. Dalam rajah tersebut, nilai eigen tertinggi ditunjukkan oleh komponen 1 dan diikuti dengan nilai eigen bagi komponen kedua dan ketiga. Sekiranya kaedah penentuan titik peralihan sebagai asas penetapan jumlah komponen sebagaimana yang dikemukakan oleh Cattell (1996), maka hanya sejumlah 7 komponen sahaja yang digunakan. Oleh itu, pengkaji telah menggunakan kaedah penentuan dengan menggunakan kaedah Kaiser dan paparan melalui *scree plot* sebagaimana yang dicadangkan oleh Field (2009).

Scree plot yang ditunjukkan seperti rajah 2 adalah merupakan hasil yang diperolehi daripada analisis PCA dengan menggunakan perisian SPSS. Dalam *scree plot* tersebut, nilai eigen tertinggi ditunjukkan oleh komponen satu dan diikuti oleh nilai eigen dua dan tiga.



Rajah 2 *Scree plot* bagi keseluruhan komponen

Menurut Field (2009) walaupun kaedah *scree plot* sangat berguna dalam proses pemilihan pemboleh ubah, tetapi ianya tidak boleh dijadikan sebagai kaedah utama dalam proses penentuan jumlah komponen kerana kesukarannya untuk menetapkan titik peralihan (*point of inflexion*). Malahan, kaedah ini juga mengekalkan sebahagian kecil varian yang terdapat dalam kumpulan data asal. Hal ini bersesuaian dengan pandangan Jolliffe (1972) yang menyatakan kaedah yang diperkenalkan oleh Kaiser untuk mengehadkan penentuan jumlah komponen dengan hanya menggunakan nilai eigen lebih daripada satu merupakan proses pemilihan yang terbatas.

Sepertimana yang ditunjukkan di dalam jadual 4, hanya 51.26 peratus daripada jumlah varian dikekalkan berbanding dengan jumlah keseluruhan varian dalam kumpulan data sebenar. Peratus jumlah varian yang besar ini menunjukkan bahawa sejumlah 51.26 peratus maklumat yang terdapat dalam kumpulan data adalah secara automatik. Hal ini bertepatan dengan kajian Field (2009) yang menyatakan bahawa walaupun kaedah pemerhatian secara grafik ini sangat berguna dalam proses pemilihan pemboleh ubah, namun ia tidak boleh dijadikan sebagai kaedah utama dalam proses penentuan jumlah komponen disebabkan oleh kesukaran untuk menetapkan titik peralihan. Selain itu, kaedah ini hanya mengekalkan sebahagian kecil varian yang terdapat dalam kumpulan data asal.

Jadual 4 Penentuan saiz komponen dengan menggunakan *scree plot*

Komponen	Nilai Eigen Awal			Jumlah Putaran Memusatkan Kuasa Dua		
	Jumlah	Varian (%)	Kumulatif (%)	Jumlah	Varian (%)	Kumulatif (%)
1	49.21	25.50	25.50	41.63	21.57	21.57
2	22.19	11.50	37.00	19.77	10.24	31.81
3	10.97	5.68	42.68	11.69	6.06	37.87
4	8.03	4.16	46.84	8.98	4.65	42.52
5	7.03	3.64	50.49	8.49	4.40	46.92
6	5.18	2.68	53.17	4.54	2.35	49.27

7	3.91	2.02	55.20	3.84	1.99	51.26

Jollfie (1972) mencadangkan agar komponen yang mempunyai nilai lebih daripada 0.7 dikekalkan bagi mendapatkan maklumat yang maksimum dalam pangkalan data. Namun begitu, kaedah ini tidak dipersetujui oleh Field (2009) kerana beliau berpendapat bahawa kaedah ini akan menyebabkan penghasilan komponen yang berlebihan. Namun, untuk kajian ini, kaedah Kaiser sebagaimana yang dicadangkan oleh Jolfie (1972) digunakan agar komponen yang mempunyai nilai eigen lebih daripada 0.7 dapat dikekalkan bagi mendapatkan maklumat yang maksimum dalam sesuatu pangkalan data

Jadual 5 Saiz komponen dengan menggunakan kaedah kriteria Kaiser

Komponen	**Nilai Eigen Awal**			**Jumlah Putaran Memusatkan Kuasa Dua**		
	Jumlah	**Varian (%)**	**Kumulatif (%)**	**Jumlah**	**Varian (%)**	**Kumulatif (%)**
1	49.21	25.50	25.50	41.63	21.57	21.57
2	22.19	11.50	37.00	19.77	10.24	31.81
3	10.97	5.68	42.68	11.69	6.06	37.87
4	8.03	4.16	46.84	8.98	4.65	42.52
5	7.03	3.64	50.49	8.49	4.40	46.92
6	5.18	2.68	53.17	4.54	2.35	49.27
7	3.91	2.02	55.20	3.84	1.99	51.26
8	3.22	1.67	56.86	3.74	1.94	53.20
9	3.11	1.61	58.47	3.34	1.73	54.93
10	2.74	1.42	59.89	3.11	1.61	56.54
11	2.59	1.34	61.24	3.01	1.56	58.10
12	2.32	1.20	62.44	2.85	1.48	59.58
13	2.07	1.07	63.51	2.54	1.31	60.89
14	1.94	1.01	64.52	2.22	1.15	62.04
15	1.86	0.96	65.48	2.11	1.09	63.13
16	1.77	0.92	66.40	2.08	1.08	64.21
17	1.66	0.86	67.25	1.93	1.00	65.21
18	1.62	0.84	68.09	1.73	0.90	66.11
19	1.53	0.79	68.89	1.73	0.90	67.01
20	1.49	0.77	69.66	1.72	0.89	67.90
21	1.41	0.73	70.39	1.70	0.88	68.78
22	1.34	0.69	71.08	1.57	0.82	69.60
23	1.29	0.67	71.75	1.55	0.80	70.40
24	1.27	0.66	72.41	1.53	0.79	71.19
25	1.24	0.64	73.05	1.53	0.79	71.98
26	1.22	0.63	73.69	1.52	0.79	72.77
27	1.20	0.62	74.31	1.45	0.75	73.52
28	1.13	0.59	74.90	1.43	0.74	74.26
29	1.11	0.57	75.47	1.35	0.70	74.96
30	1.09	0.57	76.04	1.27	0.66	75.62

Berdasarkan analisis penentuan jumlah komponen dengan menggunakan kaedah kriteria Kaiser sepertimana yang ditunjukkan dalam jadual 5, secara umumnya dapat disimpulkan bahawa setiap komponen mempunyai jumlah varian berbeza dan hanya beberapa komponen utama sahaja yang mempunyai majoriti jumlah varian. Sebagai contoh komponen satu mempunyai jumlah varian tertinggi dan diikuti oleh komponen dua dan tiga. Everitt *et al.* (2001) menyatakan proses pemilihan varian ini sebagai mengukur daya tarikan pemboleh ubah. Secara asasnya, semakin tinggi daya tarikan

sesuatu pemboleh ubah, maka lebih sesuai pemboleh ubah tersebut dimasukkan ke dalam analisis kluster. Beliau turut menegaskan bahawa pemboleh ubah yang mempunyai nilai varian yang lebih tinggi dalam setiap komponen boleh dianggap sebagai pemboleh ubah yang penting dalam sesuatu kumpulan data.

Dapatan Kajian & Perbincangan

Berdasarkan kepada analisis yang dilakukan terhadap sejumlah 179 pemboleh ubah banci yang berpotensi, hanya 69 pemboleh ubah yang dipilih untuk dilakukan proses pengklusteran bagi membentuk kluster penduduk. Pemboleh ubah tersebut dibahagikan kepada 7 sektor iaitu demografi (14). Etnik dan kepercayaan (7), pendidikan (6), pekerjaan (10), komposisi isi rumah (6), sosioekonomi (4) dan perumahan (24). Keseluruhan 69 pemboleh ubah yang telah dipilih ini digunakan sebagai input dalam proses pembangunan sistem geodemografi penduduk di Perak. Jika dibandingkan dengan dapatan daripada Vickers (2006) berdasarkan daripada hasil analisis pemilihan pemboleh ubah, hasil dapatan beliau turut merangkumi aspek kesihatan. Daripada hasil analisis kajian yang dihasilkan oleh Vicker (2006) dan pengkaji, sektor pekerjaan dan sektor demografi merupakan sektor yang dominan.

Jadual 6 Senarai pemboleh ubah akhir untuk analisis kluster

No	Pemboleh ubah	Sektor
v001	Jumlah penduduk	Demografi
v002	Penduduk berumur 00 hingga 14 tahun	Demografi
v003	Penduduk berumur 15 hingga 19 tahun	Demografi
v004	Penduduk berumur 20 hingga 24 tahun	Demografi
v005	Penduduk berumur 25 hingga 39 tahun	Demografi
v006	Penduduk berumur 40 hingga 64 tahun	Demografi
v007	Penduduk berumur 65 tahun dan ke atas	Demografi
v008	Warganegara Malaysia	Demografi
v009	Warga Penduduk Tetap	Demografi
v010	Warga Pelajar Asing	Demografi
v011	Warga Pekerja Asing	Demografi
v012	Lain-lain warga	Demografi
v013	Bukan warganegara	Demografi
v014	Berpisah	Demografi
v015	Agama Islam	Etnik dan kepercayaan
v016	Agama Kristian	Etnik dan kepercayaan
v017	Tiada agama	Etnik dan kepercayaan
v018	Bumiputera lain	Etnik dan kepercayaan
v019	Cina	Etnik dan kepercayaan
v020	India	Etnik dan kepercayaan
v021	Lain-lain etnik	Etnik dan kepercayaan
v022	Institusi Kemahiran teknikal dan Perdagangan	Pendidikan
v023	Sijil SPVM / SPM (V) / MCVE	Pendidikan
v024	Bidang Kesenian dan Kemanusiaan	Pendidikan
v025	Bidang Kejuruteraan, Pembinaan dan Latihan Kemahiran	Pendidikan
v026	Bidang Kesihatan dan Kebajikan	Pendidikan
v027	Kerja Pekerjaan Asas (Elementary)	Pekerjaan
v028	Industri Pertanian, Pemburuan dan Perhutanan	Pekerjaan

v029	Industri Perikanan	Pekerjaan
v030	Industri Perlombongan dan kuari	Pekerjaan
v031	Industri Pembinaan	Pekerjaan
v032	Industri Hotel dan Restoran	Pekerjaan
v033	Industri Pengangkutan, Penyimpanan dan Perhubungan	Pekerjaan
v034	Industri Harta Tanah, Sewa dan Aktiviti Perniagaan	Pekerjaan
v035	Industri Kesihatan dan Kerja Sosial	Pekerjaan
v036	Industri Isi Rumah Persendirian dengan Pekerja Bergaji	Pekerjaan
v037	Jumlah isi rumah 1 orang	Komposisi Isi Rumah
v038	Jumlah isi rumah 11-15 orang	Komposisi Isi Rumah
v039	Pertalian Menantu	Komposisi Isi Rumah
v040	Pertalian Cucu	Komposisi Isi Rumah
v041	Pertalian Ibu / bapa	Komposisi Isi Rumah
v042	Pertalian Lain-lain orang bersaudara	Komposisi Isi Rumah
v043	3+ Motokar	Sosioekonomi
v044	2 Motosikal	Sosioekonomi
v045	Komputer Peribadi (PC)	Sosioekonomi
v046	Tiada Peralatan yang tersebut diatas	Sosioekonomi
v047	Tempat kediaman Sesebuah	Perumahan
v048	Tempat kediaman Berkembar	Perumahan
v049	Tempat kediaman Rumah pangsa / apartment / kondominium	Perumahan
v050	Tempat kediaman Rumah kedai, pejabat	Perumahan
v051	Tempat kediaman Bilik (mempunyai jalan masuk terus)	Perumahan
v052	Tempat kediaman Khemah Buruh	Perumahan
v053	Bahan binaan Papan	Perumahan
V054	Bahan binaan Batu dan papan	Perumahan
v055	Bahan binaan Lain-lain	Perumahan
v056	Kediaman kosong	Perumahan
v057	Pemilikan Kerajaan / Badan Berkanun	Perumahan
v058	Pemilikan Sektor swasta	Perumahan
v059	Pemilikan Lain-lain	Perumahan
v060	Bekalan air sumber lain	Perumahan
v061	Elektrik dibekalkan kurang dari 24 jam sehari	Perumahan
v062	Penjana Kuasa (Generator) Sendiri	Perumahan
v063	Tiada bekalan elektrik	Perumahan
v064	Tandas Curah / Siram	Perumahan
v065	Lubang	Perumahan
v066	Ruang tertutup di permukaan air	Perumahan
v067	Tiada tandas	Perumahan
v068	Pengutipan sampah ke kawasan ini	Perumahan
v069	Tiada pengutipan sampah	Perumahan

KESIMPULAN

Analisis pemilihan pemboleh ubah memainkan peranan penting dalam menghasilkan kluster bagi sistem geodemografi. Setiap analisis yang dilakukan dalam proses pemilihan pemboleh ubah adalah penting bagi mengelakkan pemboleh ubah yang berkorelasi, pemboleh ubah yang mempunyai nilai varian rendah dan pemboleh ubah yang mempunyai variabiliti rendah manakala pemboleh ubah yang berpotensi dan relevan akan dikekalkan. Analisis pemilihan pemboleh ubah mampu mengurangkan

bilangan pemboleh ubah yang tidak relevan dan pemboleh ubah yang akan mengganggu proses analisis kluster. Daripada keseluruhan 179 pemboleh ubah banci, hanya 69 pemboleh ubah yang berpotensi telah dipilih untuk analisis kluster.

RUJUKAN

Agarwal, R & Rao, A.R. (2006). *Data reduction techniques, Indian agricultural statistics research institute Ebook Series.* URL : http://iasri.res.in/ebook, akses 20 April 2013.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-76.

Dennett, A. & Stillwell, J.C.H. (2009). *A new migration classification for local authority districts in Britain*, Working Paper 09/2, School of Geography, University of Leeds, Leeds, p.118. Available at: http://www.geog.leeds.ac.uk/wpapers/index.html

Everitt , B. S., Landau, S. & Leese, M. (2001). *Cluster Analysis*. 4th Ed. London: Arnold.

Field, A. (2009). Discovering statistics using SPSS (and sex and drugs and rock 'n' roll). New Delhi: Sage Publications.

Friendly, M. (2008). *Exploratory and Confirmatory Factor Analysis*. http://www.datavis.ca/courses/ factor/efacfa-handout1-2x2.pdf. capaian pada 3 Nov 2013.

Gopalan, N.P. dan Sivaselvan, B. (2009). *Data Mining: Techniques and Trends*, New Delhi: PHI Learning Private Limited.

Goss, J. (1995). We know who you are and we know where you live: The instrumental rationality of geodemographic systems. *Economic Geography*, 71(2): 171-198.

Harris, R., Sleight, P. & Webber, R. (2005). *Geodemographics, GIS and Neighbourhood Targeting*. London: Wiley.

Jolliffe, I.T. (1972). Discarding Variables in a Principal Component Analysis. I; Artificial Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 21 (2): 160-173.

Kaiser, H.F.(1960). The application of electronic computers to factor analysis. *Educational and Psyhological Measurement*, 20: 141-151.

Kamus Dewan Edisi Keempat. Dewan Bahasa dan Pustaka. http://prpm.dbp.gov.my/

Shepherd, P.J. (2006). Neighbourhood profiling and classification for community safety. School of Geography. University of Leeds. Leeds. http://etheses.whiterose.ac.uk/374/1/uk_bl_ ethos_436430.pdf. capaian 14 Feb 2014.

Toru Ueda. (2006). Pembinaan dan perkembangan konsep *native* di Borneo Utara pada zaman kolonial. *Akademika,* 68: 65-89.

Vickers, D. (2006). *Multi-level integrated classification based on the 2001 Census* (unpublished thesis). Department of Geography, Leeds, University of Leeds. http://etheses.whiterose.ac.uk/15/1/d.vickers_thesis _ complete_text.pdf, capaian 14 Mei 2014.