GEOGRAFIA Online™ Malaysian Journal of Society and Space **12** issue 6 (34 - 42)
Themed issue on current social, economic, cultural and spatial dynamics of Malaysia's transformation

© 2016, ISSN 2180-2491

Improving the tool for analyzing Malaysia's demographic change: Data standardization analysis to form geo-demographics classification profiles using k-means algorithms

Kamarul Ismail[1], Nasir Nayan[1], Siti Naielah Ibrahim[1]

[1]Jabatan Geografi dan Alam Sekitar, Fakulti Sains Kemanusiaan, Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak

Correspondence: Kamarul Ismail (email: kamarul.ismail@fsk.upsi.edu.my)

Abstract

Clustering is one of the important methods in data exploratory in this era because it is widely applied in data mining.Clustering of data is necessary to produce geo-demographic classification where k-means algorithm is used as cluster algorithm. K-means is one of the methods commonly used in cluster algorithm because it is more significant. However, before any data are executed on cluster analysis it is necessary to conduct some analysis to ensure the variable used in the cluster analysis is appropriate and does not have a recurring information. One analysis that needs to be done is the standardization data analysis. This study observed which standardization method was more effective in the analysis process of Malaysia's population and housing census data for the Perak state. The rationale was that standardized data would simplify the execution of k-means algorithm. The standardized methods chosen to test the data accuracy were the z-score and range standardization method. From the analysis conducted it was found that the range standardization method was more suitable to be used for the data examined.

Keywords: algorithm, data mining, geo-demographics, k-means, standardization, z-score

Introduction

Geodemographic classification involves the classification of small areas according to their inhabitants (Rothman, 1989) and geodemographic systems are built to identify common neighborhood characteristics between their communities. In developing geodemographic systems, several methods are executed to analyze the raw data to ensure that the data used in the geodemographic system is relevant and not repeating the same information. Classification is one of the methods that collects and compiles things according to certain group, certain method. Data mining analysis is one of the classification methods which are often chosen by the researchers. K-mean method is one of the simplest techniques in forming group and clustering by means of optimizing the function of the selected criteria.

K-means clustering algorithms is one of the longest method in the observation of n in d dimensional space (d integer) given and this problem is to determine a point c to minimize the mean squared distance from each data point nearby. K-means algorithm is one of the clustering methods that require few man powers in which this algorithm requires three main steps namely to determine the center coordinate, to determine the distance of each object to the center and to combine objects according to the minimum distance. According to Liu (2007), clustering algorithm is for dividing the data group and building a number of cluster k. In fact, according to Debenham et al (2001), the k-means algorithm is one of the main algorithms in clustering method because the algorithm is driven by using two basic principles which is minimizing the distance between similar objects cluster with the cluster center and maximizing the

© 2016, ISSN 2180-2491

distance of the cluster. In fact, this algorithm requires a minimum computer usage compared to other clustering methods.

By using the k-means algorithm, clustering process is done by selecting the cluster center randomly, changing the point to corresponding cluster, the mean value is updated to the new cluster and repeatedly dividing the cluster again (Ramzah Dambul & Jones, 2007). According to Dennett and Stillwell (2009), the distance between the cluster centers should be calculated because this process affects the entire cluster production process.

Data standardization method is also one of the important methods in developing the classification system. According to Dennett and Stillwell (2009), the standardization method is important especially when there is a unit difference in a group of variables. This is because the format of variables that are stored in the database may not be suitable for processing. Therefore, standardization data method that is conducted using several different standardization methods before production of clusters in order to avoid the cluster analysis results is influenced by variables with great value. However, according to Vickers (2006) before any data being standardized, the data needs to be changed to ensure that data is homogenous when clustering analysis is performed.

Standardized data has been used for a long time from the mid-18th century (Keiding, 1987) in which at that time the calculation method of data homogenization was calculated manually. Standardization needs to be done to ensure that each variable has equal weight in the process of developing a classification system. According to Urdan (2005), changing the original value of variable to standard deviation unit and z-score is one of the common standardization method used in statistical analysis. This standardization method needs to be done to the overall variable and it is not focused on variable in a certain cluster group only (Milligan and Cooper, 1988).

According to Vaishali & Rupa (2001), preprocessing data technique that is applied to the raw data is to ensure the data is consistent, not extreme and for cleaning the data. Homogenized data will standardize the raw data by means of changing the data to a certain distance using linear variation of which it produces a good cluster and improve the accuracy of the algorithm clusters.

Methodology

Area classification

Classification area is area classified to those same groups is based on characteristics that are similar. According to Vickers (2006), output area (OA) is smallest unit to develope geodemography classification based on the research in United Kingdom. However, to research carried out in Perak, smallest area unit is in zone level census as was specified by Department of Statistics Malaysia (DOSM).

Study area

This study focuses on the overall state of Perak and using the data of Population and Housing Census of Malaysia in 2000 and it involves 1884 census zone. Perak is located at the coordinate point between 5 ° 53 'North to 3 ° 42' North and 101 ° 40 ' East to 100 ° 22' East. The area is about 21,035 square kilometers with a population of about1.97 million people.

Number of households according to the Population and Housing Census in 2000 was 454,100 households with 552,185 residences with average household size of 4.35 people. Perak has 9 districts with Kinta district has the most population of about 622,106 of people, or 31.53 percent from the total population in Perak in 2000. The district that has the least population is Kampar with only 81,387 people. Even so, the district which encounters the lowest population growth rate is Hilir Perak district with a population growth rate of -0.63 percent.

© 2016, ISSN 2180-2491

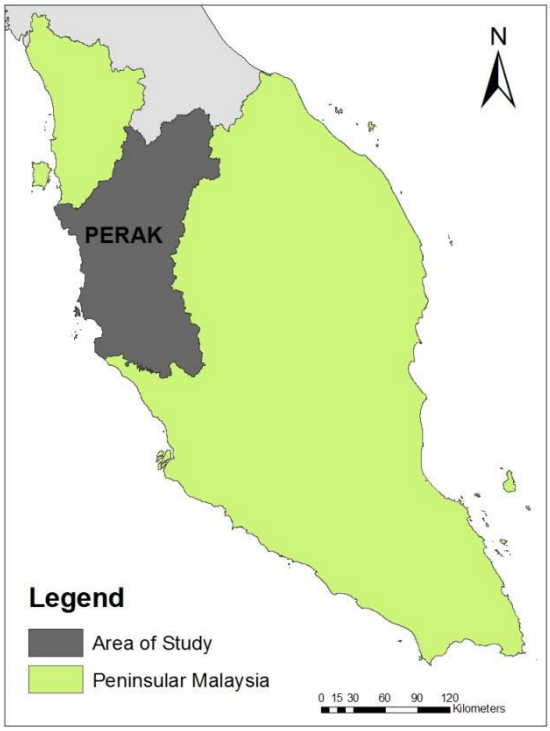

Figure 1. *Map of the study area*

In terms of distribution of land usage in Perak, about 56 percent of the area in the state is forest area, 34 percent is agriculture, 7 percent of the area is water and 3 percent is built up area (Department of Drainage and Irrigation Perak, 2011). There are 11 major river basins in Perak with the wideness of the basin of more than 80 square kilometers. Perak River Basin, which covers 14,908 sq km, is the largest river basin in Perak, which is about 70 percent of the total area of the state. Land use around the state will affect the population economic activities and at the same time can give an overview of the cluster population in an area.

Data analysis

Not all data collected will be used in generating cluster groups of the population. This is because some data have an error and repeating the same information. In fact, if there is information relating to credit reference lifestyle, appropriate classification of the data used for the business purpose. Although the data were achieved from government departments, it does not mean that the data is relevant and accurate hundred percent. Therefore, before the data is standardized, the raw data should be selected first to ensure that they do not repeat the same information and relevant to be used. In fact, unrelated variables that are included in cluster analysis will interrupt in the process of searching the cluster group structure.

a) Highly correlated variables
   Highly correlated variables are not relevant for used because it can repeats the same information. Based on the correlation analysis carried out on the raw data 61 variable, which has a correlation coefficient greater than 0.7071 as proposed by Guilford.

© 2016, ISSN 2180-2491

b) Composite variable
Composite variables can be built when two variables related showed a similar pattern to each other. This method can be used to groups a variable that correlates height and share the same unit of measurement. For example, a divorced group variables and widower or widow groups variable can combined into a separate group variable.

c) Variable that are linear related
To determine either something variable has a linear relationship with each other, principal component analysis (PCA) will be run. Theoretically, component size in correlation matric equivalent to variable total amount used and eigen value calculated through PCA analysis will show variant total that found in certain component. Therefore, the components that have small eigenvalues are not suitable because it was unable to explain the variance that occurs in a component.

Selection of the variable

203 variables have been through the process of selecting variables and only 69 variables are chosen. 134variables were removed because it's not irrelevant and highly correlated.

a) Gender
Male variable and female variable have been removed because they only give a little information about an area. In fact, most areas have a similar population distribution of men and women as shown as in the Table 1.

Table 1. The distribution of the gender by regions in Perak in 2000

City	Sex	
	Male	Female
Batang Padang	76,125	76,076
Manjung (Dinding)	97,066	94,066
Kinta	350,390	353,103
Kerian	76,025	76,886
Kuala Kangsar	71,352	73,066
Larut dan Matang	136,956	136,685
Hilir Perak	96,302	94,566
Ulu Perak	41,778	40,773
Perak Tengah	40,518	41,635

Source: Adapted from the 2000 Census data

b) Ethnic
An ethnic variable is important in the analysis of clusters but not all variables selected. This is because some of ethnic variables are highly correlated with religious variables. For example, Malay variables (v020) correlated with Islam variables (v030) while India variables (v023) correlated with variables Hindu (v032). If one of these variables is not removed, it will repeat the same information.

c) Age
There are 16 variables of age in the Population and Housing Census of Malaysia in 2000. However, there is age variable correlated highly among them and it need to be combined to a few small variable unit like variable v004 (age 00 to 04 years) are combined with variable v005 (age 05 to 09 years) and variable

© 2016, ISSN 2180-2491

v006 (age 10 to 14 years). All three variables merged into composite variable to represent the population aged 00 to 14 years.

d) Industry
There is 18 industry variable in raw data like fishing industry, manufacturing, construction, education, finance, wholesale business and retail and so on. However, only 9 industries preserved because there have industry that correlated with field variable like education field (v079) variable have high correlation with education (v113) industry variable.

Choosing the cluster numbers

To identify cluster number that will be established, algorithm k-means and method ward has been used. K-means algorithm is one of the algorithms are often used because it requires the use of a comprehensive computer than other clustering methods. Vickers (2006) suggested a number of clusters on the first hierarchical cluster is 6 and the second is the hierarchy of the 20 clusters. The k-means algorithm is repeated starting from k = 2 to k = 30 in order to identify the optimum cluster. Ward methods used to ensure optimum accuracy the number of clusters. Ward method need not be carried out repeatedly to find optimum cluster because optimum cluster visible through the elbow in data analysed. Ward method merger with an algorithm k-means to find optimum cluster number has been suggested by Burns (2009) and Burns because it more efficient and the ward method should proceed first to simplify the method of k-means algorithms.

Standardization of the variables

When there is a unit difference in a group of data, the data must be standardized to ensure comparison done between the variables is balanced and to simplify the cluster analysis that will be carried out.

According to Urdan (2005) statistical analysis such as z-score is one of the conversions of original value of variable to standard deviation unit method and it is one of the standardization methods that are often used by researchers. Usage of z-score standardized also suggested by Milligan & Cooper (1988), Lorr (1983) and Vickers (2006). Z-score is a form of homogenization that is used to change the variant value to the standard form. The formula for calculating z-score is shown below.

To obtain the z-score standardization, standard deviation calculation should be done in advance. After that, the average value will be divided with the standard deviation obtained to get the standard z-score value. To ensure that all variables in this cluster analysis are balanced, standardization must be done to all the variables. The mean value will become 0 and the standard deviation will become 1.

Addition of standardization z-score method, method of range standardization can also be used in standardizing variables. This method has been used to generate classification system ONS Local Government Act 1991 (Vickers, 2006) and proposed by Miligan & Cooper (1988). The formulas for these standardization is;

$$R_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

Where X_{max} is the maximum value of X and X_{min} is the minimum value of X. The range standardization method is also known as min-max standardization by some researchers in the field of study or another field. After running this standardization, each variable will have the lowest (min) 0.0 and the highest (max) is 1.0. This method effectively to make comparisons between different scales and units.

© 2016, ISSN 2180-2491

Result and discussion

Number of cluster

From the ward's method analysis conducted, it be found that have an elbow at the zone number 1879 from the total data 1884 zone. The optimum cluster calculated is 5 cluster, which is the number at the elbow were deducted by the total number of the zone cluster. Figure 2 show the finding from ward's method.

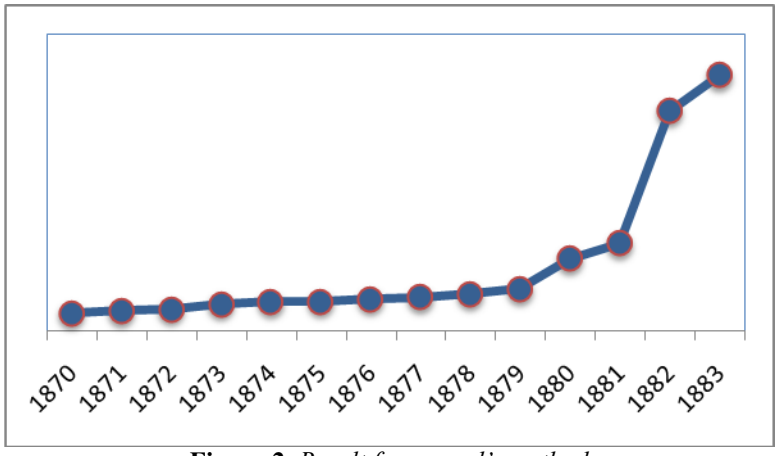

Figure 2. *Result from ward's method*

Meanwhile, the k-means algorithm will repeated the clustering process that is performed starting from k = 2 and k = 30. The optimum number of clusters can be identified by examining the distance of each object to the cluster center. The criteria can be used to determine the optimum number of clusters is to look at the average distance of the object with the lowest possible center of clusters. Figure 3 shows a nearest cluster is on progress k = 5 because the average distance from the center of the cluster object at this stage is the lowest compared to the other clusters. The average distance to the completion of six clusters were higher than 5 cluster solution and is decreased in cluster 7. Therefore, the optimum cluster using the k-means cluster is in progress 5.

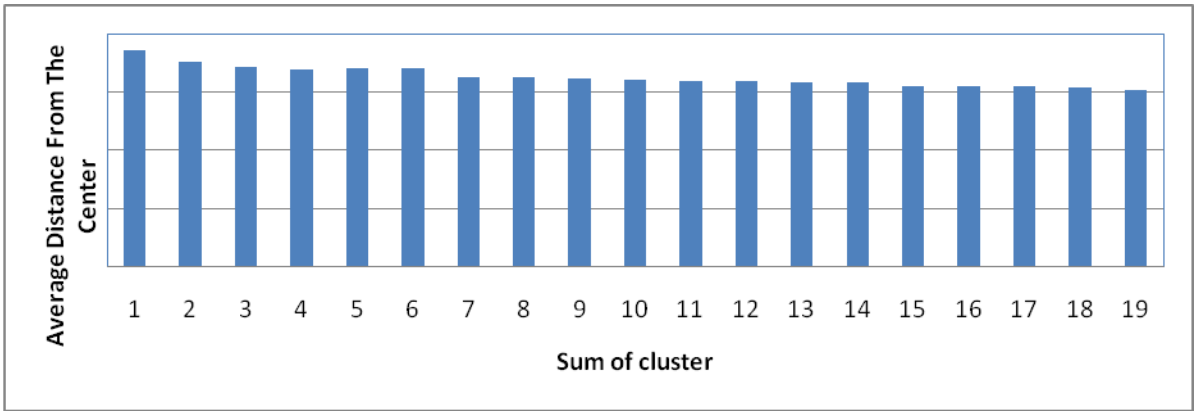

Figure 3. *Nearest cluster based on average distance from the centre*

A total of 69 variables that were selected to carry out this analysis, which are divided into several sectors, namely demographic sector (13), ethnic and religious (7), education (6), employment (10), household composition (6), socioeconomic (4) and housing (23). To solve the problem of the difference

© 2016, ISSN 2180-2491

unit score and the value contained in the set score, there are several methods of standardization of data used. The researcher has chosen to apply the z-score standardization and range standardization using Social Software Package for Social Science (SPSS) to observe which standardization method is more suitable for k-means cluster algorithm analysis. When the data standardized with z-score method and range standardization, the distribution of object clustering will be more balanced.

Table 2. Distribution of data clusters that have not been standardized

Cluster	Frequency	Percent	Cumulative Percent
1	18	1.0	1.0
2	188	10.0	10.9
3	571	30.3	41.2
4	524	27.8	69.1
5	583	30.9	100.0
Total	1884	100.00	

Table 2 shows the distribution of data that has not been run with any standardization analysis. The data that are too focused on specific cluster and not uniform should be analyzed to ensure that it is suitable for cluster analysis.

Range standardization

Table 3. Distribution of data clusters that carried out by range standardization

Cluster	Frequency	Percent	Cumulative Percent
1	80	4.2	4.2
2	103	5.5	9.7
3	891	47.3	57.0
4	622	33.0	90.0
5	188	10.0	100.0
Total	1884	100.00	

Cluster analysis using standardized variables using the range standardization method obtains unbalanced distribution of census zones, especially in cluster 1 and cluster 3. Cluster that has a minimum number of census zones is in cluster 1 which is 80 census zones or 4.2 percent from the total number of census zones. While clusters with the maximum number of census zones is in cluster 3, accounting for 47.3 percent of the entire census zones. Number of census zones in cluster 2, 4 and 5 generally has an unbalanced distribution of census zones namely between 5.5 per cent to 33.0 per cent.

Z-score

Table 4. Distribution of data clusters that carried out by z-score standardization

Cluster	Frequency	Percent	Cumulative Percent
1	59	3.1	3.1
2	163	8.7	11.8
3	597	31.7	43.5
4	437	23.2	66.7
5	628	33.3	100.0
Total	1884	100.00	

© 2016, ISSN 2180-2491

Cluster analysis using standardized data with z-score standardization method acquire the distribution of more balanced clustering objects. As shown in Table 4, the minimum cluster obtained through z-score standardization method is in cluster 1 which is 59 census zones. While the group that has a maximum census zone is in cluster 5, which represent a total of 33.3 percent of the entire census zones. Whereas the number of census zone distribution in cluster 4 is between 2.3 and 8.7 per cent to 31.7 per cent.

The value of sum of squares error (SSE) describes the distance of each variable to the center of the cluster to test which method is more effective. Small value of SSE indicates it is more accurate and more appropriate to be used. Table 5 shows the difference in SSE value for each standardization method.

Table 5. Values of SSE for each method of standardization

Standardization method	Value of SSE
Origin data	1.358
Z-score standardization	1.877
Range standardization	0.930

Conclusion

This study able to identify methods of clustering, which are more efficient and this article also gave an overview that related to process involved before standardization analysis carried out. Area classification and cluster analysis has shown how classification geodemography constructed. In fact, variable selection on the other hand carried out carefully to make sure only data that is really relevant and not repeat the same information. Based on variable analysis, only 69 of 203 variables chosen to perform analysis cluster and from 69 variables that are selected, it forms 5 optimum cluster in first hierarchy.

Based on the standardization analysis that has been done, it was found that the range standardization method is more suitable to be used for data of Population and Housing Census of Malaysia in 2000 for the state of Perak. This is because the distribution of cluster zone conducted with range standardization analysis is more balanced compared with the distribution of cluster zone analyzed using the z-score standardization. Distribution of cluster zone during the standardized process is essential to avoid the data too focused on specific clusters. Standardization analysis can help to reduce the distance between variables as could be seen in the SSE.

For future research, the researcher can compare a few methods of data standardization as proposed by Vickers (2006) and Milligan & Cooper (1988) to find the best method of data standardization. The result of the comparison maybe different from each other based on used data.

Acknowledgement

This research project is under a research grant from the Ministry of Education (MOE) with research code 2013-0015-106-72 RAGS.

References

Debenham J, Clarke G, Stillwell J (2001) Deriving Supply-side Variables to Extend Geodemographic Classification. [Cited 17 November 2015]. Available from: http://eprints.whiterase.ac.uk/5017/1/01-5.pdf.

© 2016, ISSN 2180-2491

Dennet A, Stillwell JCH (2009) A new migration classification for local authority districts in Britain, Working Paper 09/2, School of Geography , University of Leeds, Leeds. [Cited 17 October 2015]. Available from: http://www.geog.leeds.ac.uk/wpapers/index.html.

Department of Irrigation and Drainage Perak (2011) Compendium of Data and Information Basic JPS. [Cited 28 June 2014]. Available from: http://www.jpsperak.gov.my/userfiles/files/PDF/Publication/Kompendium%20Data%20dan%20Maklumat%20Asas%20JPS%20Negeri%20Perak.pdf.

Keiding N (1987) The Method of Expected Number of Deaths, 1786–1886–1986. *International Statistical Review* **55**, 1-20.

Liu B (2007) Web Data Mining –Exploring Hyperlinks, Contents and Usage Data, Springer Series on Data-Centric Systems and Applications.

Lorr M (1983) Cluster Analysis for Social Scientists. Jossey-Bass.

Malaysia (2001) General Report of the Population and Housing Census of Malaysia 2000. Department of Statistics Malaysia, Putrajaya.

Milligan GW, Cooper MC (1988) A Study of Standardization of Variables in Cluster Analysis. *Journal of Classification* **5**, 181-204.

Ramzah Dambul, Jones P (2007) Regional and temporal climatic classification for Borneo. *Geografia-Malaysian Journal of Society and Space* **3** (1). 84-105.

Rothman J (1989) Editorial. *Journal of the Market Research Society* **31** (1).

Urdan TC (2005) Statistics in Plain English. Lawrence Erlbaum Associates, New Jersey.

Vaishali RP, Rupa GM (2011) Impact of outlier removal and normalization approach in modified k-means clustering algorithm. *Int. J. Comput Sci.* **8**(5), 331-336.

Vickers D (2006) Multi-level Integrated Classifications Based on the 2001 Census (Unpublished thesis). Department of Geography. Leeds, University of Leeds. [Cited 14 September 2015]. Available from: http://etheses.whiterose.ac.uk/15/1/d.vickers_thesis_complete_text.pdf.