CLUSTERING OF HOUSING AND HOUSEHOLD PATTERNS USING 2011 POPULATION CENSUS

Nontembeko Dudeni-Tlhone
e-mail: *ndudenitlhone@csir.co.za*
and
Jenny P. Holloway
and
Sibusisiwe Khuluse
and
Renée Koen
Decision Support and Systems Analysis, Spatial Planning and Systems, CSIR Built Environment, P O Box 395, Pretoria, 0001

Key words: *k*-means clustering, housing and household patterns, Ekurhuleni Metro, urban planning, census 2011 data.

Summary: This study looked at a specific application of cluster analysis using the recently released population census 2011 data for the Ekurhuleni Metro in the Gauteng Province of South Africa. The main focus of the clustering was to distinguish housing and household patterns in order to create homogenous groups with similar demands for infrastructure, facilities and services. The *k*-means algorithm was specifically applied to groups of variables (factors) such as the dwelling types, conditions and location characteristics, socio-economic profiles, as well as demographic factors. These groups of clusters were later combined in a sequential manner to obtain a final set of meaningful clusters that could be used as inputs into an urban growth simulation tool.

1. Introduction

This paper looks at a specific application of clustering using the recently released population census 2011 data for the Ekurhuleni Metro in the Gauteng Province of South Africa. The focus of the clustering is to distinguish between different types of housing and households in order to create homogenous groups with similar demands for infrastructure, facilities and services. The need for these clusters stems from an urban planning tool developed by the Council for Scientific and Industrial Research (CSIR), known as the Urban Growth Simulation Platform, which is a numerical modelling and simulation platform developed to study urban growth patterns 30 years into the future. The clusters are required as the baseline data assigned to land parcels in the model.

Previously, clusters that were developed for market segmentation purposes, at the suburb level, could be purchased from Knowledge Factory and these were based primarily on data from the Deeds office and Census 2001. Due to the lack of updated clusters and the need for clusters that are focused on the demand for services at a small spatial area, it was necessary to create clusters from the Census 2011 data by selecting only relevant variables and using the small area layer (SAL) as the base unit.

In the literature there are numerous examples internationally of clustering techniques being applied to population census or to land-use data. Many of these studies involve using census, together with house price and other data sources, to classify the population of interest into homogenous housing submarkets in regions such as Australia (Bourassa, Hamelink, Hoesli and MacGregor, 1999), New Zealand (Bourassa, Hoesli and Peng, 2003) and the USA (Hwang and Thill, 2009; Cho, Poudyal and Lambert, 2008), with the latter paper using the clusters to measure the effects of urban growth boundaries on land development patterns. Vickers and Rees (2007) clustered the 2001 census data from the UK National Statistics Office to provide area classifications that could be distributed by the Office of National Statistics via their website. Cluster analysis is also commonly applied internationally to land-use data for many different types of studies, including: to derive landscape typologies (Van Eetvelde and Antrop, 2009), to differentiate between agricultural land-cover patterns and their dynamics across years (Reger, Otte and Waldhardt, 2007) and to assess areas for residential development due to urban sprawl. Cluster analysis has also been used extensively in transport planning where residential locations have been clustered into similar neighbourhood types with similar lifestyle choices and used as input for assessing travel patterns (Manaugh, Miranda-Moreno and El-Geneidy, 2010; Mohammadian and Zhang, 2007; Krizek and Waddell, 2002). However, similar studies have not been found for South Africa, where urban growth differs significantly from other developed countries, and where homogenous groupings of the population are influenced by previous politics of the country.

The *k*-means algorithm is well known and well documented (MacQueen et al., 1967; Jain, Murty and Flynn, 1999; Huang, 1998). Vickers and Rees (2007), citing Harris, Sleight and Webber (2005), states that it is one of the most commonly used methods in geodemographics. The *k*-means clustering method has been used in many household segmentation problems such as (Vickers and Rees, 2007; Bourassa et al., 1999; Bourassa et al., 2003; Cho et al., 2008; Manaugh et al., 2010); with Vickers and Rees (2007) using a 3-stage *k*-means clustering which clusters on all variables at each stage, while Bourassa et al. (1999), Bourassa et al. (2003) and Manaugh et al. (2010) applied *k*-means clustering to the components or factors derived from principal component or factor analyses. Both Vickers and Rees (2007) and Bourassa et al. (1999) additionally applied hierarchical clustering with Ward's algorithm but found that they obtained better results from the *k*-means method. Hwang and Thill (2009) provide an example of fuzzy clustering to derive housing submarkets while Bação, Lobo and Painho (2004) compared the *k*-means algorithm to self-organising maps for the clustering of census data.

In this paper, we propose clustering the housing and household patterns of Ekurhuleni via the *k*-means algorithm but using a unique approach in which 5 separate sets of clusters are created for all SALs, using different groupings of variables to describe different aspects of land-use type and associated households. These groups of clusters are later combined to create a final set of clusters, as described under the results section. Therefore the paper can be described as the first iteration towards defining an appropriate set of clusters to be used for urban planning. More importantly, it focuses on the description of a methodology which can be used with historical or future data, thereby enabling studies of change in the spatial profiles of housing and households. However, it is envisaged that further refinements will be required before the final set of clusters can be obtained.

2. Materials and methods

In this section, information about the study area and the data used in the cluster analysis of household and housing patterns is given.

2.1. Study area

This study focused on segmenting parts of Ekurhuleni Metropolitan Municipality with respect to housing and household characteristics using population and housing statistics. Ekurhuleni is one of the three metropolitan municipalities situated in the Gauteng province, South Africa. It is the second largest municipality of Gauteng in terms of population size (3.178 million residents) after the City of Johannesburg Metro, which is home to approximately 4.434 million people (Statistics South Africa, 2012). The decision for selecting Ekurhuleni as a study area was made on the basis of synchronising this work with other projects and planning activities already taking place within the CSIR's research and development space.

2.2. Data used

The data used to cluster the households within Ekurhuleni Metro were extracted from the recent (2011) South African census gathered and disseminated by Statistics South Africa. The census data were used as they are (1) believed to largely contain useful information for profiling households within various geographical locations and (2) the recent statistics were disseminated at a finer spatial scale in what is known as the Small Area Layer (SAL). A SAL is a single spatial layer made up of parcels defined as small areas. However, for our own internal communication purposes we used "SALs" to refer to the small areas. It is therefore expected that this level of data will provide insight into current housing and household patterns. A total of 4610 SALs were extracted from the census data and used for cluster analysis of households and housing characteristics. Given that the analysis was done with the SALs as basic units, the results should be interpreted at this area level rather than at the level of individual households and or the members thereof.

For this study, it was important to determine a range of household characteristics which are useful for grouping SALs according to factors influencing housing preferences and locations (particularly in the urban context). Key variables were therefore identified and classified according to factors associated with socio-economic profiles of the households within the SALs, as well as the demographic, life-stage, location features and dwelling types and conditions. These key variables and groupings were determined via consultation with the urban planning experts, who were to be the end-users, taking into account the characteristics that they require in the clusters. The final list of variables used in the cluster analysis of the SALs is shown in Table 1. It is important to note that all of these variables were described in terms of a number of categories (shown in the results), with the exception of the household income and density of dwellings. For instance, in the census data the population group is categorised with respect to the four main population groups in South Africa, which includes Black or African, Coloured, White and Indian or Asian population groups. For use within the clustering, these categories split into separate variables containing counts of people classified within each specific group. Even though the data appear to be either about the households or the members thereof, the analysis was done on summaries at the level of SALs and not at household-level.

Table 1: Variables used in the analysis

Broad classification of variables (factors)	**Key variables**
Dwelling location characteristics and density	Enumeration area type
	Density of dwellings within small areas
Dwelling type and conditions	Type of dwelling
	Household size
	Number of rooms
Socio-economic	Weighted average annual household income
	Employment status of the head of household
	Highest education level of the head of household
Life cycle stage and household structure	Marital status
	Age group
	Relationship structure within households
Demographics	Population group
	Gender of household head
	Gender of persons in the households
	Property ownership of the households

2.3. Method used

As indicated in the introductory section, the k-means procedure was used to detect groups of SALs in Ekurhuleni Metro which exhibit similar characteristics with respect to housing and population statistics. The k-means algorithm is one of the popular unsupervised clustering methods used for partitioning n-dimensional data sets into k clusters. It is known to be reasonably efficient in minimising the within-cluster variance in the resulting segments (MacQueen et al., 1967). The k-means is also known for its efficiency in processing large numeric data sets (MacQueen et al., 1967; Jain et al., 1999; Huang, 1998). Its use is, however, limited in applications using categorical data, and extensions of this method have been developed to specifically provide an analysis tool for categorical values (Huang, 1998). The k-means method assigns observations to clusters through a variety of algorithms, and these algorithms basically differ in the manner by which the starting points representing the cluster centroids are selected. For instance, some k-means algorithms use observations from the dataset as starting cluster centroids while others generate random seeds and use them to represent the initial clusters. The k-means, therefore, requires that k points (which represent the initial group of centroids or the required number of clusters) be determined beforehand. Defining an appropriate number of clusters ensures that the natural structure of the data is well represented in the k sets of clusters to be formed (Berry and Linoff, 1999). Each observation is first assigned to the initial cluster centroid to which it is the closest. Once all observations are linked to the initial centroids, the centroids are recalculated. The process is repeated until there is no change in the centroids.

The k-means method therefore minimises an objective function which consists of the squared error function or an average squared distance between the observations and their cluster centroids. The objective function is therefore a measure of how well the centroids represent the observations in the relevant clusters. The objective function is given by,

$$O = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||x_i^{(j)} - m_j||^2 \tag{1}$$

where k is the number of clusters and n_j denotes the number of observation in cluster j in the data space of n observations. $||x_i^{(j)} - m_j||^2$ is a squared distance between observation $x_i{}^{(j)}$ and its cluster centroid m_j, within cluster j (Jain et al., 1999).

2.4. Data processing and analysis

This section provides a summary of the steps used for pre-processing the data prior to clustering. Firstly, the correlation structure of the dataset was analysed to identify between-variable correlations for variable reduction and to avoid redundancies, as well as to determine whether some categories within variables could be merged. For instance, some age groups were combined if they were highly linearly correlated, according to Pearson's correlation coefficients, such as groups '55 to 59' and '60 to 64'.

As mentioned earlier in Section 2.2, the census parameters (except for income and density) were split into separate variables containing counts per category within each parameter. These counts were converted into proportions summing to 1 across all categories within each census parameter. Income and density were converted to continuous variables. All variables were standardised to have a zero mean and a unit standard deviation to ensure that all variables have equal weights in the clustering. The k-means algorithm was then applied to each group of variables, where groups are defined in column 1 of Table 1, rather than to all variables at once. In applying the algorithm, an appropriate number of clusters for each group was determined using the Elbow criterion (Dimitriadou, Dolničar and Weingessel, 2002), which evaluates the within-cluster variation for different numbers of clusters. Having finalised the number of clusters per group, the resulting sets of clusters for the different groups were combined, as described in Section 3.1.

3. Results

This section provides an overview of the cluster analysis results. Clusters may be described in terms of the number of SALs in a cluster as well as the cluster averages. Since data were standardised, variables contributing more to the cluster profile tend to have cluster means that are further from zero. Figure 1 shows how the variables included in the socio-economic factors contributed to the

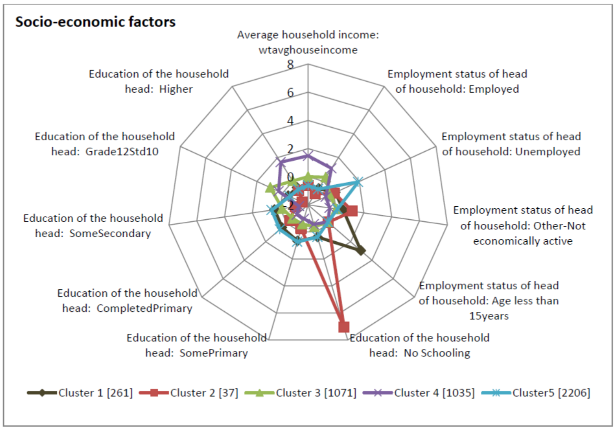

Figure 1: Resulting clusters from the socio-economic factors

clustering of the SALs of the Ekurhuleni metro. Cluster 1, which contained 261 of the total of 4610 SALs in Ekurhuleni, is predominantly characterised by SALs with a large number (compared to the average) of child-headed households. Cluster 2 represents the SALs that have a larger number of households which are headed by persons without any form of schooling, and as a result, are economically inactive. Table 2 gives a summary description of a set of clusters obtained from the dwelling location characteristics and density factors including the density of dwellings and the type of the enumeration area. This table gives the average values for each variable within a cluster. Since the variables were standardised across the dataset, a value of zero indicates the average for Ekurhuleni, while positive and negative values represent population characteristics above and below the Ekurhuleni averages, respectively. Most (72%) of the SALs fall within relatively less dense formal residences while cluster 1 is primarily characterised by SALs with highly dense informal residences as well as collective living quarters. Cluster 5 contains a large proportion of formal residences with high density of dwellings.

Table 2: Cluster summary of the dwelling locations and density factors

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Cluster size (% of SALs)	687(14.90%)	17(0.37%)	3333(72.30%)	37(0.80%)	389(8.44%)	73(1.58%)	68(1.48%)	6(0.13%)
Density of dwellings	0.91	-0.87	-0.31	-0.69	1.53	-0.85	-0.85	-0.64
Formal residential	-2.04	-2.04	0.49	-2.03	0.44	-2.04	-2.04	0.18
Informal residential	2.26	-0.40	-0.39	-0.40	-0.40	-0.40	-0.40	-0.40
Traditional residential	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Collective Living Quarters	0.67	-0.12	-0.12	-0.12	-0.12	-0.10	-0.09	-0.12
Small holdings	-0.12	-0.12	-0.12	-0.12	-0.12	-0.12	8.15	-0.12
Farms/Parks/Vacant	-0.06	41.62	-0.03	-0.03	-0.03	-0.02	0.00	0.93
Industrial	-0.10	-0.10	-0.10	-0.10	-0.10	7.80	-0.10	-0.10
Commercial	-0.10	-0.10	-0.10	11.00	-0.10	-0.10	-0.10	-0.10

3.1. Creating final clusters

Due to the specific urban planning problem for which the clusters are required, the 5 factors (each containing their own set of clusters) were ordered in terms of their level of importance for differentiating between household and land-use types. This prioritisation was decided in consultation with the urban planning experts. Our process for obtaining the final clusters therefore involved a stepped approach in which SALs were first split according to the clusters of the Dwelling location factor, which was defined as the first stage group. One of the characteristics of *k*-means clustering is its sensitivity to outliers and can be used in clustering out outliers (Yoon, Kwon and Bae, 2007). In this case, that characteristic was beneficial to the process as it helped to cluster out the small groups which were different to the predominantly residential population, such as clusters dominated by farms. In the second stage of the clustering, only the 3 residential clusters (low density formal, high density formal and informal) were split further by the clusters from the dwelling type and conditions factor. This group helped to separate the SALs further according to dwelling type, size of dwelling and household size. Some of the clusters that were created at the end of the second stage were then split at the third stage via the socio-economic group of clusters (which define income, employment status and education) and, where appropriate, a fourth stage and fifth stage were applied using the life stage and demographic groups. Figure 2 describes the process up to the third stage. An example of the resulting clusters A and B produced using the proposed approach is shown in the third stage of Figure 2 and highlights the key differences detected. Cluster A represents a group of SALs with

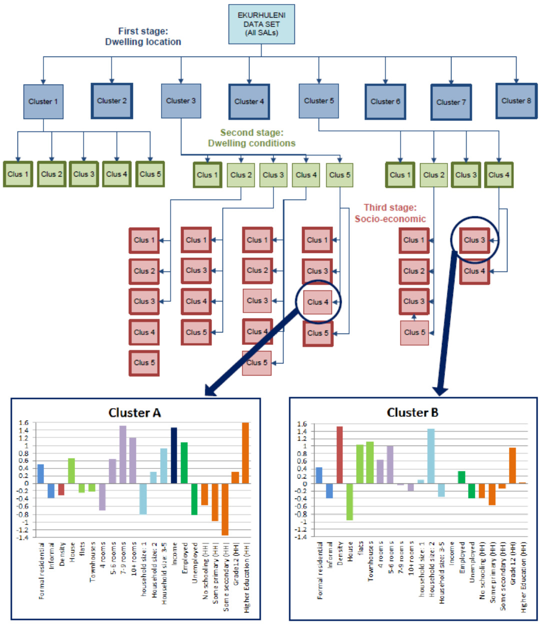

Figure 2: Creating final cluster from the clusters determined in the five groups of variables

households living in large properties (in terms of the number of rooms which mainly range between 7 and 9), having high average household income, and being mainly headed by employed and educated (in possession of tertiary qualifications) individuals. On the other hand, cluster B is characterised by SALs with medium-sized residences (4 to 6 rooms), where heads of households are mainly in possession of matric, and with types of dwellings being mainly flats and townhouses in high density areas.

4. Conclusions and further work

The aim of the study was to detect homogenous groups out of the different household characteristics and housing types in Ekurhuleni Metro using the recent population and housing census for which data were provided at a finer spatial resolution compared to the previous census data. Meaningful clusters were obtained from the clustering performed and it is envisaged that further work would focus on the validation of these clusters using other sources of data. For instance, property values and age of property of small areas are important aspects that can be incorporated in understanding the housing patterns relevant to urban planning as well as identifying areas with specific characteristics or service delivery needs.

References

Bação, F., Lobo, V., and Painho, M. (2004). Clustering census data: comparing the performance of self-organising maps and k-means algorithms. *In KDNet Symposium: Knowledge-Based Services for the Public Sector, 3rd-4th June, Bonn, Germany.*

BERRY, M. AND LINOFF, G. (1999). *Mastering data mining: The art and science of customer relationship management*. John Wiley & Sons, Inc.

BOURASSA, S. C., HAMELINK, F., HOESLI, M., AND MACGREGOR, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics*, **8** (2), 160–183.

BOURASSA, S. C., HOESLI, M., AND PENG, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics*, **12** (1), 12–28.

CHO, S.-H., POUDYAL, N., AND LAMBERT, D. M. (2008). Estimating spatially varying effects of urban growth boundaries on land development and land value. *Land Use Policy*, **25** (3), 320–329.

DIMITRIADOU, E., DOLNIČAR, S., AND WEINGESSEL, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, **67** (1), 137–159.

HARRIS, R., SLEIGHT, P., AND WEBBER, R. (2005). *Geodemographics, GIS and neighbourhood targeting volume 7*. John Wiley and Sons.

HUANG, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, **2** (3), 283–304.

HWANG, S. AND THILL, J.-C. (2009). Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning B: Planning and Design*, **36** (5), 865–882.

JAIN, A. K., MURTY, M. N., AND FLYNN, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, **31** (3), 264–323.

KRIZEK, K. J. AND WADDELL, P. (2002). Analysis of lifestyle choices: Neighborhood type, travel patterns, and activity participation. *Transportation Research Record: Journal of the Transportation Research Board*, **1807** (1), 119–128.

MACQUEEN, J. ET AL. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability volume 1*, 281-297. California, USA, p. 14.

MANAUGH, K., MIRANDA-MORENO, L. F., AND EL-GENEIDY, A. M. (2010). The effect of neighbourhood characteristics, accessibility, home–work location, and demographics on commuting distances. *Transportation*, **37** (4), 627–646.

MOHAMMADIAN, A. AND ZHANG, Y. (2007). Investigating transferability of national household travel survey data. *Transportation Research Record: Journal of the Transportation Research Board*, **1993** (1), 67–79.

REGER, B., OTTE, A., AND WALDHARDT, R. (2007). Identifying patterns of land-cover change and their physical attributes in a marginal European landscape. *Landscape and urban planning*, **81** (1), 104–113.

VAN EETVELDE, V. AND ANTROP, M. (2009). A stepwise multi-scaled landscape typology and characterisation for trans-regional integration, applied on the federal state of Belgium. *Landscape and Urban Planning*, **91** (3), 160–170.

VICKERS, D. AND REES, P. (2007). Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170** (2), 379–403.

YOON, K.-A., KWON, O.-S., AND BAE, D.-H. (2007). An approach to outlier Detection of Software Measurement Data using the K-means Clustering Method. *In First International Symposium on Empirical Software Engineering and Measurement, 2007(ESEM 2007). IEEE, pp. 443–445.*