doi:10.1111/j.1757-7802.2012.01076.x

Creating a small scale area classification for understanding the economic, social and housing characteristics of small geographical areas in the Philippines

Adegbola Ojo, Daniel Vickers, Dimitris Ballas

University of Sheffield, Department of Geography, Winter Street Sheffield S10 2TN, UK
(e-mail: A.ojo@adegbolaojo.co.uk; D.vickers@sheffield.ac.uk; D.ballas@sheffield.ac.uk)

Received: 27 September 2011 / Accepted: 13 July 2012

Abstract. The Philippines is one of the most populous countries in the world. In terms of population, it ranks twelfth globally and seventh in Asia behind China, India, Indonesia, Pakistan, Bangladesh and Japan. The estimated population of the country in 2010 was 94 million people. Using data from the Philippines 2000 Census, this paper presents a discussion of the creation of a 3-tier hierarchical geodemographic system for the country at Barangay scale. Barangays are the smallest spatial entities in the structure of the administrative geography of the country. Most popular geodemographic systems are typically developed from continuous datasets. In this paper, we discuss how a geodemographic classification system can be created by combining categorical and continuous datasets. The first level of the Philippines geodemographic hierarchy ensures the population can be profiled broadly at Barangay level into seven super-groups. The super-groups are further subdivided into 24 groups and finally into 66 subgroups.

JEL classification: J11, J2, I3, C1, C15

Key words: Philippines, geodemographics, area classifications, Barangays

1 Introduction

The Philippines National Statistics Office (NSO) is the government agency charged with the conduct and publication of national Censuses in the country. On 1 May 2000, the first census of the current millennium was conducted in the country. In most countries of the world including the Philippines, the census is traditionally the most comprehensive body of information on population socio-demographics and housing.

The sheer volume of data collected during censuses makes data processing a daunting task. As such, in most countries, it takes two to three years to manipulate the data and make results available in formats suitable for further end-user analysis.

Data from the Philippines 2000 Census were released in phases. From the 2000 Census, the estimated population of the country was about 76.5 million people (NSO 2010). The Census

showed a relative balance in the ratio of men to women. For every 100 women, there were 101.43 men. This is an internationally high sex ratio which stems from the dominance of women in a large emigration flow from the Philippines (Omelaniuk 2005; Rodriguez 2005). However, half of the national population was under the age of 21 years and there were a greater number of males in the younger age categories (NSO 2002). Apart from two age categories – 0 to 19 years and 25 to 54 years, all other age groups were dominated by females (NSO 2002).

It is common practice in many developing countries that the national or regional level of geography is typically the unit of analysis and reporting of national statistics. A good example of this practice is the popular and useful Demographic and Health Survey (DHS) conducted in over 70 countries by Macro International and supported by the United States Agency for International Development (USAID) and partner countries. Analysis conducted at higher levels of geographical aggregation often conceal inequalities at sub-national geographies (Schelzig 2005; Dorling and Ballas 2008).

One of the reasons some analysts and policy-makers seem disinterested with manipulating multivariate data at fine spatial scales is due to the volume of information entailed. For instance, it may be relatively easier for the human mind to assimilate information on a number of indicators spread across the current 17 regions of the Philippines. Digesting the similar information for the over 40,000 Barangays[1] in the country may however be complex.

One way around this problem is to condense complex multidimensional attribute data such that the resultant output reflects the prevailing characteristics of the residential population at this fine scale of geography. This process of simplifying multidimensional attribute data often results in the classifying of areas into groups. Geodemographics, 'the analysis of people by where they live' (Sleight 2004, p. 18) is one of the most commonly used of methods for producing classifications of small areas. There remains paucity in the proliferation and use of these systems within much of the developing world (Ojo and Ezepue 2011; Ojo et al. 2012) despite recent wide use of geodemographic methods and systems across much of the developed world (Longley and Singleton 2009; Singleton 2009; Singleton and Longley 2009; Harris et al. 2010).

The aim of this paper is to provide a robust description of the procedures and decisions made during the development of the Philippines geodemographic classification system; and subsequently to describe key features of the social groups identified at the Barangay spatial scale. We commence by providing a rationale for creating a small area segmentation system for the Philippines. Section 3 details the sources of data for the analysis and provides details on how the data inputs were selected. In Section 4, we provide an account of the clustering analytics while Section 5 details how the resulting solutions were evaluated. Section 6 provides some insight into the neighbourhood types identified and their spatial distributions while conclusive comments are given in the final section.

2 Rationale for area segmentation system for the Philippines

We argue that developing countries can benefit tremendously from small area segmentation systems. In Nigeria for instance, a geodemographic system has been recently developed and is being used to understand spatial inequalities in the country (Ojo 2010; Ojo et al. 2012; Ojo and Ezepue 2012).[2] In spite of the benefits of small area segmentations, their proliferation to countries in the developing world has been relatively slow.

[1] Barangays are the smallest spatial units in the structure of the administrative geography of the Philippines.
[2] For more details see: http://www.nigerianlgaclassification.com/

Like many other developing countries, spatial analytical techniques for decision-making in the Philippines is still evolving. As such researchers still face some challenges when conducting research studies aimed at understanding local level spatial inequalities. For instance there remains a misconception of the importance of spatial data infrastructure. The emphasis placed by the government on developing physical infrastructure like good roads, hospitals, bridges and providing clean water supply sometimes obscure the significance of developing efficient data infrastructure. Ironically, policy decisions on the provision of some of these physical infrastructures should be based on evidence derived from timely and informative datasets. Another impediment faced by researchers is the problem of gaining access to available datasets. This is sometimes due to bureaucratic legal frameworks and the fact that some of the relevant data are not available in digital formats.

The Philippines can benefit tremendously from the development of an area segmentation system particularly in the policy-making arena. Geodemographic segmentations offer researchers and policy-makers an alternative option for investigating local level inequalities especially within a data-scarce country. By creating a small area classification with national or near national coverage, it is possible to analyse and evaluate populations and their characteristics by area types. National surveys can be plugged into the classification to generate insight into the fundamental characteristics of respondents to such surveys. Results can also be extrapolated nationally (Harris et al. 2005).

The segmentation system can also help the government and other important stakeholders make informed decisions, target interventions and allocate development resources more judiciously when trying to target small areas. The targeting interventions require reliable identification of special or vulnerable populations. Identifying such groups requires unveiling their attributes and locating their distribution across geographical space. A useful feature of geodemographic systems is the textual and graphical explanations that accompany the results of analysis from which they are developed (Harris et al. 2005). These descriptions help summarize the predominant attributes of the population groups and further elucidate (in qualitative terms) information inherent in complex quantitative analysis. It therefore means that if ancillary datasets are linked with geodemographic typologies, decision-makers are provided with a wealth of social and spatial correlates which may not have been immediately obvious without the geodemographic link.

This is a strong call that reinforces the view that addressing social and spatial problems of poverty requires a multi-faceted approach. The problem of data supply, which may not be peculiar to the Philippines, is also an issue. Of paramount importance however is how to improve targeting efficiently and effectively so that the most vulnerable population groups can be reached. Essentially, the process of tackling the problem of within-country social and spatial inequalities needs to be made a local issue (Ojo 2010).

3 Data sources and selection of input variables

Cluster analysis is an important method for developing geodemographic systems. It entails recognizing identical groups of objects also called clusters from a pool of several objects. Objects that fall into the same cluster share common characteristics and are generally dissimilar to objects assigned to a different cluster.

According to Milligan (1996), the key steps in the cluster analysis process include:

Step 1: Clustering elements (Objects to cluster, also known as operational taxonomic units)
Step 2: Clustering variables (Attributes of objects to be used)
Step 3: Variable standardization

Step 4: Measure of association (Proximity measure)
Step 5: Clustering method
Step 6: Number of clusters
Step 7: Interpretation, testing and replication.

What these steps suggest is that the process of developing a social area segmentation system is not merely running data through a clustering algorithm but involves a number of important steps which require both quantitative and qualitative analysis.

For the purpose of the research presented here, the clustering elements are the Philippines Barangays. The focus of this section is to describe the techniques used to determine the appropriate variables chosen for inclusion in the clustering algorithm.

The process of policy-making is the means by which governments convert their political vision into actions (Lindblom and Woodhouse 1992). Of particular significance especially to developing countries is its potential application towards making better policies. While the lack of political will has often been blamed in many developing countries including the Philippines for slow growth (Montinola 1999) some policy-makers may be well intentioned but may be misguided.

The choice of variables considered for inclusion in the Philippines classification took cognizance of current policy debates especially as they relate to the targets of the eight MDGs. An example is drawn from the Medium Term Philippine Development Plan (MTPDP), which seeks to reduce poverty. It has been argued that policies focused on the dynamics of the population will be useful in achieving this feat (Schelzig 2005).

The datasets for the census were derived from two sources. The National Statistics Office (NSO) in Manila provided geographic information systems (GIS) digital boundaries as well as census data for multiple geographical levels. Most of the data provided at the finest administrative geography (Barangays) were categorical datasets. The second source of data was derived from the Integrated Public Use Microdata Series-International (IPUMS). IPUMS is based at the University of Minnesota in Minneapolis. The data from IPUMS comprised a large database of almost 7.5 million individual records – 10 per cent of the Philippines 2000 Census and covering a variety of topics and indicators.

Data was not made available for 14 Barangays. Further correspondence with staff from the NSO revealed that these Barangays either were evacuated during the conduct of the census due to armed clashes between rebel forces and government troops; or had been evacuated due to volcanic eruption at Mount Pinatubo. These 14 Barangays were excluded from the analysis.

3.1 Appraisal of categorical and continuous datasets

The statistical evaluation of the suitability of input variables is important when choosing variables. However, when attempting a classification system for any developing country, it is imperative for such a system to encapsulate variables that reflect key policy debates. This is likely to increase the usability and usefulness of the system because it will capture population patterns that will improve the targeting of such policies.

A total of 435 variables were available for initial consideration. For easy understanding, all the variables were aligned along 10 domains listed below:

1 Demographic
2 Housing
3 Education
4 Religion

5 Employment
6 Services
7 Health
8 Socio-economics
9 Household Infrastructure
10 Transportation

These domains reflect the principal dimensions of the dataset which comprised categorical and continuous variables. For categorical variables, the numerical values assigned to specific geographic areas are placed only into categories or classes while continuous variables represent numerical measurements on a continuous scale and can take any numerical value within the continuum (Crawshaw and Chambers 2001). It is of course possible for some other researchers to come up with a set of different domains.

The reduction of initial variables and the selection of the final list were done in three phases. First, it was decided that all 23 categorical variables should be analysed together as the statistical tests deployed on categorical data differ from continuous data. Second, all continuous datasets would be examined on a domain-by-domain basis and thirdly all continuous data surviving the domain-by-domain reduction would be evaluated together.

3.2 The analysis of principal components

Principal components analysis (PCA) was used to investigate which variables are likely to have the greatest influence on the classification. Both categorical and continuous variables were analysed. The objective of the analysis was to determine the variables that are likely to have the greatest influence on the intended classification. The first principal component produced in the course of the analysis accounts for the greatest possible proportion of the variance of the variable set while the second accounts for the maximum remaining variance. Essentially later components explain less of the variation within the data (Dunteman 1989; OECD 2008). The first component was therefore used to determine which variables have the greatest influence on the dataset.

Table 1 shows the results of the analysis conducted for the categorical variables. The presence or absences of a postal and telephone service are the two categorical indicators that are likely to have the greatest influence on the classification system.

Among continuous variables, we determined that the variable representing population without access to piped water has the highest loading for component 1. Other highly important

Table 1. Top 10 loadings of the first principal component of categorical variables

Categorical Variable	Loading
Is there a postal service in the Barangay?	0.65
Is there a telephone in the Barangay?	0.65
Is there a telegraph in the Barangay?	0.62
Is there a newspaper circulation in the Barangay?	0.58
Does the Barangay have a street pattern of at least 3 street roads?	0.54
Is there a high school in the Barangay?	0.49
Is there a town/city hall or provincial capital in the Barangay?	0.48
Is there a market or building with trading activities in the Barangay?	0.44
Is there a college/university in the Barangay?	0.43
Is there electric power in the Barangay?	0.43

variables include population employed in public administration, free occupancy take-up, houses built with bamboo and retail trade establishments.

3.3 Spatial dispersion

For adequate distinctions between areas, input variables need to demonstrate reasonably high levels of variation. The standard deviation has been suggested by (Bulmer 1979; Urdan 2005) as a useful statistic for measuring geographic dispersion of data. Standard deviation (σ) provides a measure of the magnitude by which values tend to depart from the mean. This makes the derivation of the mean a focal point in calculating the standard deviation. This statistic was used to assess continuous datasets.

For the categorical datasets, we adapted a method to first calculate the mean and convert the variables into numbers. This was done by deriving the weighted average of all probable values that each variable can take on. The weights used in calculating the average correspond to probabilities. Crawshaw and Chambers (2001) provide detailed explanation on how an expectation or expected mean can be calculated for discrete categorical distributions using probability theory. Table 2 shows the categorical variables arranged in ascending order of their variation across geographical space.

Table 2. Standard deviation values for categorical variables

variables	Expectation E(X) or μ	μ^2	$E(X^2)$	Var(X)	SD (σ)
Does the Barangay have a street pattern of at least 3 street roads?	1.50	2.24	2.49	0.26	0.51
Is Barangay part of town city proper or poblacion?	1.55	2.41	2.66	0.25	0.50
Is there a community waterworks system in the Barangay?	1.52	2.32	2.56	0.24	0.49
Is there a health centre in the Barangay?	1.39	1.93	2.16	0.23	0.48
Is there a public plaza or park in the Barangay?	1.64	2.70	2.92	0.22	0.47
Is there an elementary school in the Barangay?	1.27	1.60	1.79	0.19	0.43
Is there a telephone in the Barangay?	1.73	3.00	3.18	0.19	0.43
Is there electric power in the Barangay?	1.24	1.54	1.71	0.17	0.41
Is the Barangay accessible to the national highway?	1.22	1.48	1.65	0.17	0.41
Is there a cemetery in the Barangay?	1.79	3.19	3.35	0.16	0.40
Is there a postal service in the Barangay?	1.81	3.28	3.42	0.14	0.38
Is there a market or building with trading activities in the Barangay?	1.82	3.33	3.47	0.14	0.38
Is there a church/chapel or mosque in the Barangay?	1.18	1.38	1.52	0.14	0.37
Is there a high school in the Barangay?	1.83	3.35	3.48	0.13	0.37
Is there a newspaper circulation in the Barangay?	1.84	3.37	3.50	0.13	0.36
Is there a Barangay hall in the Barangay?	1.13	1.27	1.37	0.10	0.32
Is there a housing project in the Barangay?	1.91	3.63	3.71	0.08	0.27
Is there a town/city hall or provincial capital in the Barangay?	1.93	3.73	3.79	0.06	0.25
Is there a telegraph in the Barangay?	1.93	3.74	3.79	0.05	0.22
Is there a hospital in the Barangay?	1.95	3.81	3.85	0.03	0.19
Is there a college/university in the Barangay?	1.97	3.87	3.89	0.03	0.16
Is there a public library in the Barangay?	1.97	3.87	3.89	0.02	0.14

Table 3. Continuous variables with the lowest standard deviations

Variables	SD (σ)
Insurance and pension funding establishments	0.02
Age 91	0.02
Establishments manufacturing office, accounting and computing machinery	0.02
Establishments manufacturing coke, refined petroleum and other fuel products	0.02
Five married couples in household	0.02
Establishments Manufacturing of tobacco products	0.01
Age 93	0.01
Age 95	0.01
Age 94	0.01
Age 96	0.01
Age 98	0.01
Age 97	0.01
Air transport establishments	0.01
Age 100	0.01
Six married couples in household	0.01

Measures of dispersion were calculated for continuous variables using the more common method of deriving the standard deviation. Table 3 shows some continuous variables with very low distribution.

While age is important especially in explaining life-stage activities, Table 3 shows that single year age groups are less likely to provide good discrimination across geographical space. A solution to this problem was to create composites of the various age groups. The one year age groups were combined into six age groups as shown below.

1 Age 0 to 4
2 Age 5 to 9
3 Age 10 to 19
4 Age 20 to 44
5 Age 45 to 64
6 Age 65+

Apart from increasing the standard deviations of the variables, creating composite age groups also helped minimize the positive skew in age variables.

3.4 Measuring association

When two or more highly correlated variables are included in a clustering algorithm, there is the tendency that the outcome will consist of redundant information (Everitt et al. 2001; Vickers and Rees 2006). This can have a negative influence on the clusters produced. Although multicollinearity can be problematic in clustering, there are alternate views on how this can be handled. Many commercial classification builders allow these relationships and manage their effects by weighting (Harris et al. 2005).

In order to eliminate multicollinearity as much as possible, strength of association tests were deployed on the variables. A useful measure of the strength of association in categorical data is

to calculate the proportional difference between concordant (**P**) and discordant (**Q**) pairs (Agresti and Finlay 2009) and derive a measure called the gamma statistic using Equation 1.

$$\gamma = \frac{P - Q}{P + Q} \tag{1}$$

Gamma (γ) can take values ranging from −1 to +1. A value closer to +1 indicates a positive (concordant) association, while a value closer to −1 indicates a negative (discordant) relationship. To derive the gamma, a crosstab query was run on all categorical variables to derive two-by-two matrices for each pair. Each of the two-by-two matrices was ordered by yes and no responses. The strength and direction of the association between the different pairs of variables is shown Table 4.

Generally, there is weak positive association across the dataset. The shaded cells in Table 4 indicate variables demonstrating high levels of positive correlation. Gamma values of +0.7 and above have been defined as high.

3.5 Handling skewed indicators

Prior to choosing a variable for inclusion in a clustering algorithm, it is important to consider the extent to which it is skewed. Skewness defines the extent to which a variable distribution is symmetrical about its mean (Crawshaw and Chambers 2001). A variable is said to exhibit positive skew if its asymmetric tail (when charted) extends towards the positive values of the distribution while for a negative skew, the tail extends towards the negative values.

Skew can result from a number of factors. If a variable represents a small proportion of the population, most values will be concentrated around the lower end of a 0 to 100 per cent scale. Skew can also result from extreme values or outliers. An example of a variable that may demonstrate this feature in most developing countries is population density. It is common in most developing countries to find urban centres with disproportionately higher population concentrations than their rural counterparts as a result of mass rural-urban drift (George 1999).

The problem with introducing highly skewed variables into the analysis is that they may obscure the rest of the dataset and create artificial or outlier clusters. It is therefore important to test variables for their skewness and avoid where possible the inclusion of highly skewed variables in the analysis. It is practically impossible to completely eliminate the presence of skew when creating classifications; however their effects can be minimized by transforming the data (Walker and Maddan 2008).

Variables within the employment domain showed some of the largest positive skews. With the exception of Philippines benevolent missionaries (religion domain) every other indicator in Table 5 is within the employment domain. One reason for this is the number of categories of employment. Clearly the more the categories, the more the data is split within these categories and the more the chance of zero values for some Barangays.

3.6 List of final variables

Outside the framework of the statistical consideration, some decisions are subjective while others are based on experience and knowledge of the study area. The judgments made in the selection process were carefully considered and are by no means finite. A different classification developer for the Philippines may derive an entirely different list of indicators. Table 6 shows the list of the 69 variables selected for further analysis and inclusion in the subsequent clustering

Table 4. Gamma Statistic Matrix for Categorical Datasets

	D1	E1	E2	E3	H1	H2	HO1	R1	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	SE1	T1	T2
D1																						
E1	–0.23																					
E2	0.32	0.53																				
E3	0.63	0.22	0.91																			
H1	0.54	0.03	0.67	0.85																		
H2	0.03	0.52	0.43	0.30	0.48																	
HO1	0.25	0.06	0.48	0.68	0.55	0.20																
R1	–0.06	0.64	0.36	0.20	0.19	0.52	0.08															
S1	0.78	0.01	0.67	0.72	0.75	0.32	0.36	0.21														
S2	0.10	0.24	0.37	0.39	0.34	0.28	0.24	0.40	0.58													
S3	0.06	0.31	0.46	0.29	0.30	0.14	0.08	0.49	0.47	0.17												
S4	0.51	0.21	0.67	0.77	0.76	0.56	0.49	0.27	0.82	0.63	0.30											
S5	–0.02	0.46	0.32	0.12	–0.36	0.61	–0.04	0.36	–0.09	0.38	–0.17	0.19										
S6	0.42	–0.23	0.38	0.66	0.59	0.22	0.55	–0.16	0.45	0.07	–0.16	0.61	0.20									
S7	0.54	–0.25	0.46	0.79	0.66	0.26	0.59	–0.12	0.58	0.19	–0.16	0.66	0.35	0.86								
S8	0.74	–0.14	0.65	0.80	0.78	0.27	0.59	0.00	0.87	0.44	0.22	0.80	–0.01	0.83	0.86							
S9	0.43	0.00	0.54	0.67	0.65	0.39	0.46	0.06	0.76	0.26	0.09	0.72	0.25	0.87	0.77	0.94						
S10	0.32	0.00	0.34	0.54	0.44	0.24	0.36	0.16	0.42	0.18	0.10	0.53	0.26	0.55	0.55	0.63	0.51					
S11	0.30	0.02	0.48	0.69	0.55	0.43	0.42	0.08	0.38	0.21	–0.23	0.67	0.45	0.85	0.90	0.81	0.81	0.52				
SE1	0.38	0.20	0.63	0.62	0.63	0.42	0.34	0.35	0.73	0.39	0.48	0.64	0.14	0.33	0.39	0.66	0.50	0.41	0.19			
T1	0.58	0.01	0.49	0.74	0.61	0.22	0.45	0.09	0.67	0.38	0.19	0.62	0.22	0.61	0.67	0.78	0.60	0.37	0.52	0.49		
T2	0.26	0.01	0.31	0.58	0.46	0.33	0.35	0.14	0.32	0.12	–0.14	0.61	0.31	0.68	0.68	0.57	0.59	0.32	0.70	0.30	0.17	

Notes: D1 = Is Barangay part of town city proper or poblacion? E1 = Is there an elementary school in the Barangay? E2 = Is there a high school in the Barangay? E3 = Is there a college/university in the Barangay? H1 = Is there a hospital in the Barangay? H2 = Is there a health centre in the Barangay? HO1 = Is there a housing project in the Barangay? R1 = Is there a church/chapel or mosque in the Barangay? S1 = Is there a town/city hall or provincial capital in the Barangay? S2 = Is there a public plaza or park in the Barangay? S3 = Is there a cemetery in the Barangay? S4 = Is there a public library in the Barangay? S5 = Is there a Barangay hall in the Barangay? S6 = Is there a newspaper circulation in the Barangay? S7 = Is there a telephone in the Barangay? S8 = Is there a telegraph in the Barangay? S9 = Is there a postal service in the Barangay? S10 = Is there a community waterworks system in the Barangay? S11 = Is there electric power in the Barangay? SE1 = Is there a market or building with trading activities in the Barangay? T1 = Does the Barangay have a street pattern of at least 3 street roads? T2 = Is the Barangay accessible to the national highway?

Table 5. Variables with the largest positive skew

Variable	Skew
Ancillary unit establishments	58.33
Financial intermediation establishments	50.1
Integrated paper and paper products establishments	45.59
Other manufacturing establishments	42.42
Recycling establishments	38.62
Philippines benevolent missionaries	36.54
Construction establishments	33.49
Forestry, logging and related service activities establishments	31.58
Electricity, gas, steam and hot water supply establishments	30.57
Electricity, gas and water supply establishments	29.51

algorithm. The process of variable selection entailed making complex decisions. It is almost impossible to narrate the entirety of the issues taken into consideration in the process. The initial list comprised 435 variables with the demographic and employment domains recording the highest proportions of 31 per cent and 25 per cent respectively. The other eight domains shared the rest of the 44 per cent with the transportation and socio-economic domains accounting for low values of 1 per cent and 2 per cent respectively.

At the end of the first phase (intra-domain variable reduction/selection) of the selection process 81 per cent of the initial variables were rejected reducing the number to 81 variables. At this stage, the proportion of all variables accounted for by the demographic and employment domains had reduced to 44 per cent. The final phase of reductions took cognizance of the merged variables and inter-domain analysis. Another 15 per cent of the variables were rejected. The final list comprises 69 variables. The employment and demographic domains account for the largest proportions of 22 per cent and 19 per cent respectively. Three domains – household infrastructure, housing and services each account for 12 per cent. The education, religion and socio-economic domains account for 6 per cent each. The health domain has a representation of 4 per cent while the transportation domain accounts for 3 per cent of all variables.

4 Developing the Barangay classification system

Prior to proceeding with clustering any dataset, it is important to consider the different scales of measurement used for the input variables. This is because clustering variables measured in different units and between different scales will not yield the true picture of the areas (Vickers 2006). To re-scale the dataset, the variables were converted into standard normal variate scores.

A logarithmic transformation was applied to reduce the effect of skew within the dataset. The method was used because of its ability to cope well with positive skew. Vickers (2006) also found out that this method worked well with UK output area census data.

Once the dataset had been prepared, a choice of clustering algorithm had to be made. The size of the data required an algorithm that could handle the volume of data. The other issue to consider was the fact that the database was a combination of categorical and continuous datasets. Most clustering algorithms work well with continuous datasets. An algorithm specifically designed to handle combinations of categorical and continuous datasets is the two-step cluster analysis procedure (Banerjee et al. 2004). The procedure gives the best results if categorical datasets appear to have a multinomial distribution and continuous variables display a normal distribution.

Table 6. The final list of variables

Variable	Domain	Variable	Domain
Part of city or poblacion	Demographic	Electricity supplied	Household Infrastructure
Population density	Demographic	No piped water	Household Infrastructure
No married couples	Demographic	Cooking energy – liquid fuels	Household Infrastructure
Two or more married couples	Demographic	Cooking energy – wood	Household Infrastructure
Stepchild of head	Demographic	No telephone	Household Infrastructure
Age 0 to 5	Demographic	Television	Household Infrastructure
Age 6 to 10	Demographic	Water closet toilet	Household Infrastructure
Age 11 to 20	Demographic	Latrine	Household Infrastructure
Age 21 to 44	Demographic	Dwelling owned	Housing
Age 45 to 64	Demographic	Members squatting	Housing
Age 65+	Demographic	Free occupancy	Housing
Widowed	Demographic	Land is occupied with consent	Housing
Speaks Filipino	Demographic	Built with bamboo	Housing
Elementary school	Education	Built with bricks, stone or concrete	Housing
High school	Education	Built with mixed materials	Housing
Less than primary education	Education	Iron and concrete roofing	Housing
Post secondary technical education	Education	Buddhist	Religion
Professionals and Senior Officials	Employment	Muslim	Religion
Agricultural and fishery workers	Employment	Roman Catholic	Religion
Transportation and communications employment	Employment	Other Christians	Religion
Public administration and defence employment	Employment	Town hall	Services
Education employment	Employment	Public plaza or park	Services
Health and social work employment	Employment	Cemetery	Services
Overseas worker	Employment	Barangay hall	Services
Real estate and business establishments	Employment	Telegraph service	Services
Private establishments	Employment	Postal service	Services
Co-operative establishments	Employment	Waterworks system	Services
Small establishments	Employment	Method of waste water disposal	Services
Media establishments	Employment	Market	Socio-economic
Retail trade establishments	Employment	Manufacturing establishments	Socio-economic
Banking institutions	Employment	Auto repair shops	Socio-economic
Recreational establishments	Employment	Restaurants and personal services	Socio-economic
Hospital	Health	At least 3 street roads	Transportation
Health centre	Health	Accessibility to the highway	Transportation
Disabled	Health		

During the first step of this clustering procedure, preclusters are formed using a distance measure. The preclustering process does not require a pre-specification of the desired number of clusters. Every object is considered in relation to already formed clusters and based on the distance measure, it is decided if an object should start a new precluster or be assigned to an already formed cluster (Hellerstein and Stonebraker 2005).

Once the preclusters have been formed, each of the clusters is considered as a single object. The second step of the procedure deploys a hierarchical algorithm on the preclusters (Hellerstein and Stonebraker 2005). During this step, cognizance is taken of the number of preclusters formed. Large number of preclusters results in better solutions (Hellerstein and Stonebraker 2005) but demand more computational power which in turn slows down the algorithm.

All clustering algorithms group cases based on similarity or dissimilarity. Similarity of cases within taxonomic space is measured by deriving a statistical quantification of distance (Everitt

et al. 2001). Generally, similar cases have a closer distance. A likelihood ratio test has been recommended by Agresti (2007) as a reliable test for significance when dealing with categorical datasets. The log-likelihood ratio test was used to evaluate similarity within the datasets. It is based on probabilities and compares the maximized likelihoods of a null hypothesis to an alternative one. The larger the value of the statistic, the less the within-cluster variation and the more compact cases are.

4.1 Problems with Initial Clustering

The analysis was deployed with the hope of creating a hierarchy similar to the Nigerian local government area (LGA) classification (Ojo et al. 2012). The top-down methodology employed by Vickers (2006) in creating the hierarchy of the UK output area classification was adopted because it ensures cluster groups in lower hierarchies maintain as much qualities of their parent clusters.

When the algorithm was run on the database, the first observation was a massive imbalance in the distribution of cases within clusters. Too few cases were concentrated in some clusters. Additionally, all Barangays in the National Capital Region (NCR) were put into a single cluster. This problem was caused because of an initial use of range standardization. When dealing with mixed variables, range standardization can present problems for the clustering algorithm. Indeed Bacher et al. (2004) seemed to suggest this when they wrote as follows:

> Simulation studies suggest, that z-standardization is ineffective (for a summary of simulation results; see Everitt et al. 2001: 51). Better results are reported for standardization to unit range (ibidem). However, standardization to unit range is problematic for mixed type attribute (Bacher et al. 2004, p. 5).

The problems arising from range standardization necessitated the consideration of z-scores, which performed better after deploying the Two-Step clustering algorithm.

5 Evaluating the solution

In order to create a three-tier hierarchy with the two-step clustering method, an acceptable number of clusters had to be derived at the topmost hierarchy. This again can be very daunting. The software used to run the clustering algorithms was the Statistical Package for the Social Sciences (SPSS). One of the limitations of SPSS is that it has no options for replication and the resulting classifications are dependent on the ordering of the objects to be classified. Usually, SPSS selects the starting seed centroids from the top row. To minimize the effect of this problem, the clustering algorithm was run several times with different seed centroids and the solution with the smallest mean distance to cluster centre was used.

When deciding on the choice of the number of clusters, one of the approaches suggested by Everitt et al. (2001) is to assess the clustering criterion against the number of clusters. Other issues to consider include the need for a balanced distribution of cases within clusters; a relatively balanced population distribution; an acceptable national distribution of clusters suitable for further analysis.

The clustering criterion selected for the exercise is the Schwarz Bayesian Information Criterion (BIC). The change in BIC is basically the difference between the log likelihood ratio statistic and the clustering parameters (Akaike et al. 1998). A perceived ideal solution would be the point at which there is an abrupt increase in the BIC (Banerjee et al. 2004; Larose 2006). From Figure 1, this point is located at the seven cluster solution after running the analysis for up to 50 clusters. After carefully deciding on the number of clusters at the topmost level of the

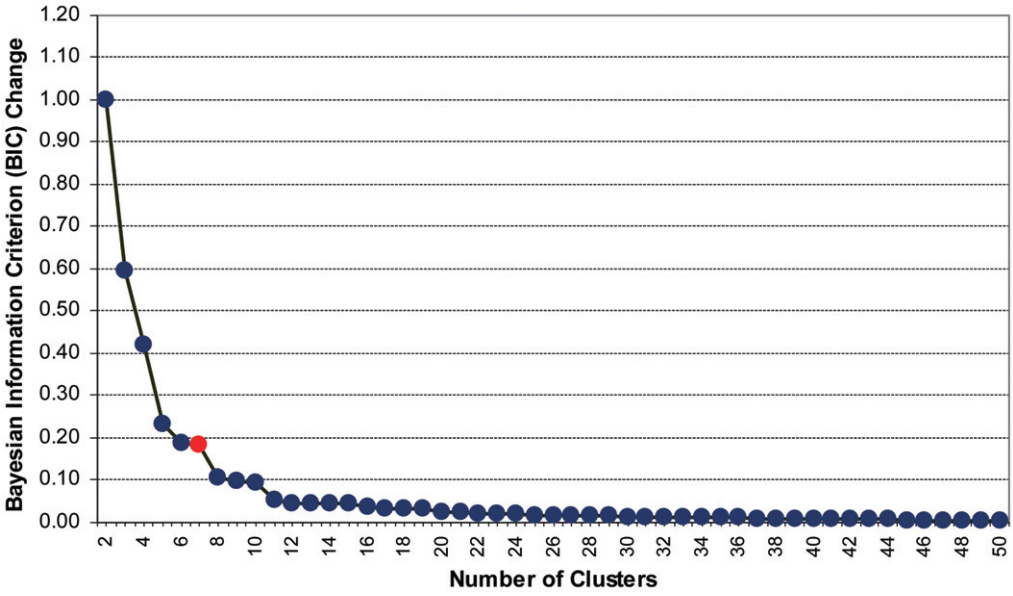

Fig. 1. Change in the average distance of cases to cluster centre

hierarchy, other levels of the hierarchy were created. The unclassified cases were excluded from further analysis and the clustering algorithm was deployed on each of the seven clusters at the top level.

Issues described above were taken into consideration all the way and a second hierarchy of 24 clusters was created. The 24 clusters were also analysed further to create a third hierarchy of 66 clusters. The top hierarchy of seven is called super-groups; the middle hierarchy of 24 is called groups; while the finest level of aggregation comprising 66 clusters are called subgroups.

5.1 Naming of clusters

The process of naming clusters brings the classification system to life. It is both an art and a science requiring the fusion of knowledge from different disciplines and a high level of creativity. It also demands a good grasp of the meanings of different words. The naming process takes the general socio-demographic characteristics and geographical location of each cluster into consideration.

It can be very contentious to assign a label to a group of people. Different views will occur on what groups should be called. In multi-ethnic and multi-religious developing countries like the Philippines, labelling groups of people can often give rise to debate. It is important to try as much as possible to avoid stigmatizing a group of people. Figure 2 shows the names assigned to the seven super-groups; the 24 groups and the number of subgroups into which each group has been split.

Apart from the broad geodemographic characteristics of clusters and their geographic locations, the labels for each super-group were restricted to a maximum of two words to allow for easy memorization. Groups were restricted to a maximum of three words.

While the labelling of clusters are a very important part of the exercise because this is often what captures the interest of users. It is important to state that a two-word label is not enough to summarize the multivariate characteristics of clusters. Labels should therefore be used with

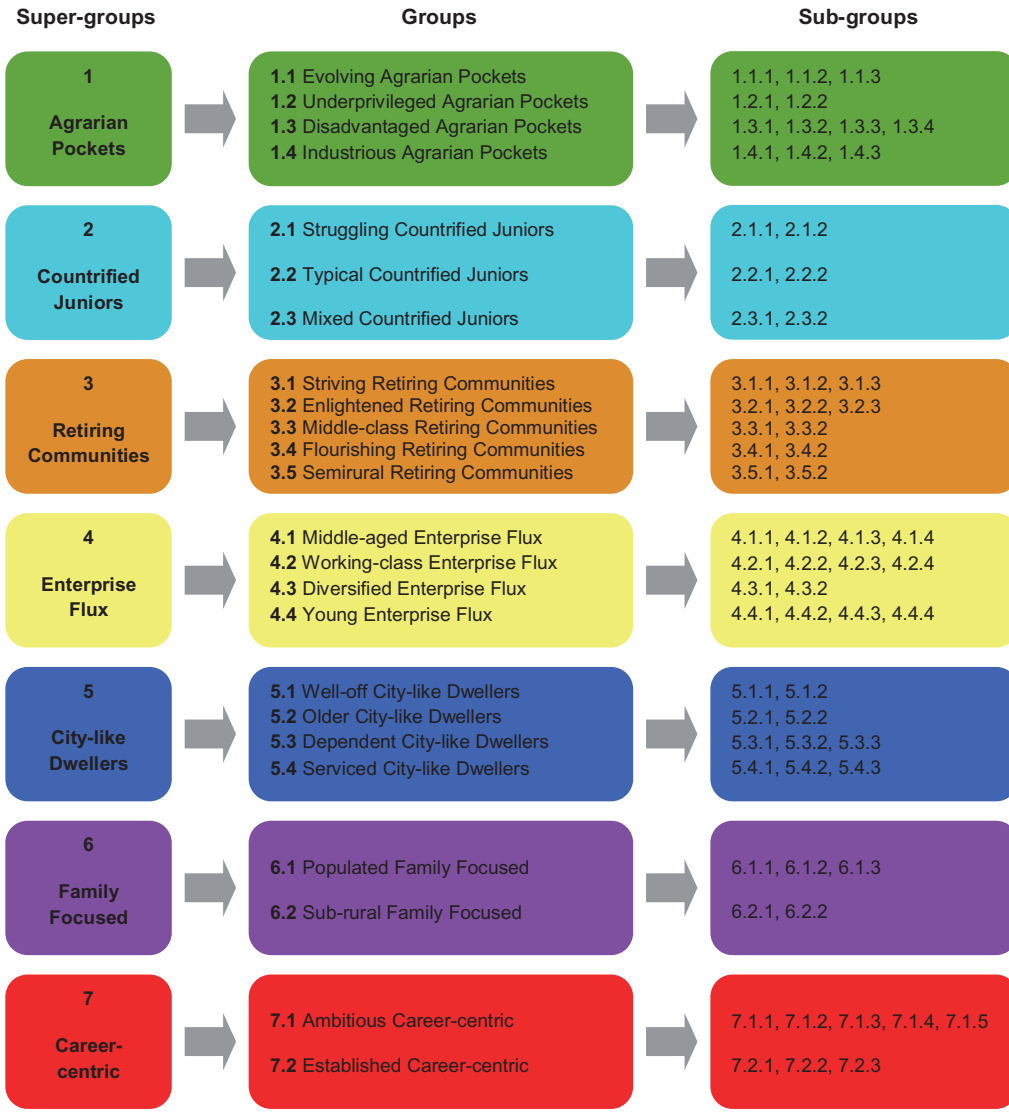

Fig. 2. Hierarchical structure of the Philippines geodemographic classification system

utmost care and greater importance should be paid on the evidence upon which the labels have been assigned. This evidence is defined by the variations in the importance and behaviours of the input variables as discussed in the next section.

5.2 The distribution of cases and population within clusters

It is important for the number of Barangays and population assigned to each cluster to be reasonably balanced. This will ensure that the solution is robust and suitable for use for further analysis (Vickers 2006). Usability of the classification will be increased if surveys are representative in their distribution across the different clusters. Table 7 shows the extent to which clusters are sized in terms of Barangays and population.

Table 7. Distribution of Barangays and Population in Each Cluster

Cluster	Barangays (count)	Barangays (%)	Population (count)	Population (%)
Agrarian pockets	9,862	23.51	12,494,336	16.35
Countrified juniors	5,233	12.48	8,802,716	11.52
Retiring communities	6,919	16.50	10,365,522	13.57
Enterprise flux	7,955	18.97	12,143,974	15.89
City-like dwellers	5,751	13.71	20,648,202	27.02
Family focused	2,258	5.38	4,036,362	5.28
Career-centric	3,564	8.50	6,953,227	9.10
Unclassified	398	0.95	969,318	1.27

Table 8. Percentage Distribution of Barangays in Each Super-group Cluster by Regions

Administrative Region	Agrarian Pockets	Countrified Juniors	Retiring Communities	Enterprise Flux	City-like Dwellers	Family Focused	Career-centric	Unclassified	Total
Autonomous region in Muslim Mindan	26.38	13.94	17.77	21.61	6.69	4.49	8.37	0.75	100
Cordillera Administrative Region	26.71	12.20	17.66	19.45	11.95	4.69	7.25	0.09	100
Ilocos Region	24.99	13.05	17.86	19.94	10.38	4.93	8.18	0.67	100
Cagayan Valley	25.57	12.98	17.57	18.52	9.69	4.85	9.56	1.25	100
Central Luzon	18.83	13.20	13.64	19.61	19.20	4.99	10.01	0.54	100
Southern Tagalog	21.05	12.02	15.53	18.45	17.24	5.61	8.41	1.67	100
Western Mindanao	25.74	12.92	17.29	21.04	8.64	4.32	9.02	1.03	100
National Capital Region	13.05	10.04	14.05	6.55	40.50	8.85	5.14	1.83	100
Bicol Region	23.23	13.22	15.71	19.91	12.50	5.67	8.78	0.97	100
Western Visayas	25.04	12.91	17.41	19.58	10.46	5.18	8.53	0.90	100
Central Visayas	23.86	12.03	18.22	18.55	13.24	5.18	8.59	0.33	100
Eastern Visayas	26.84	12.74	17.69	19.02	9.98	5.47	7.88	0.38	100
Northern Mindanao	22.24	11.94	15.81	18.37	17.45	5.45	8.27	0.46	100
Southern Mindanao	21.32	12.03	15.44	18.32	15.91	6.15	9.22	1.60	100
Central Mindanao	26.85	10.21	15.03	21.61	10.14	5.87	7.97	2.31	100
Caraga	24.10	11.25	15.53	20.96	12.17	5.28	10.25	0.46	100

A total of 384 Barangays are included in the unclassified clusters. These are areas with characteristics that do not readily fit in any of the existing clusters – primarily those with a large non-household population. These 384 Barangays were added to the 14 initially excluded due to lack of data to give 398 which constitute the unclassified group of cases.

Among the seven clusters, the largest comprises 9,862 Barangays while the least is made up of 2,258 Barangays giving a range of 7,604 Barangays. Population-wise, the largest cluster comprises 20,650,941 people while the smallest has a population of 4,036,687 people.

It is also interesting to observe that the cluster with the largest number of Barangays is not the cluster with largest number of people. One factor accounting for that is the level of urbanization of the cluster with the largest population share.

The National Capital Region, Southern Tagalog and Central Luzon have the highest levels of urbanization. Most Barangays within these three regions fall into the city-like dwellers cluster as shown in Table 8. However, urbanization alone may not be the only factor accounting for the

large concentration of Barangays. It will be interesting to explore the use of family planning measures and the influence of religious beliefs within these areas. Some faith groups in the Philippines have been known to oppose some of these health measures (Ballweg 1972). Table 8 shows row percentages and illustrate how robust the clusters are across all the regions. This will make it easier to extrapolate ancillary data linked to the classification system for further analysis and visualize results nationally.

6 Social profiles and spatial distribution of clusters

To gain an understanding of the geodemographic characteristics of each cluster, certain techniques were used to evaluate the outputs from the cluster analysis. It is important to name the clusters such that they reflect the general characteristics of the people and the areas in which they live (Vickers 2006). It is also important to get an understanding of the variables which drive the

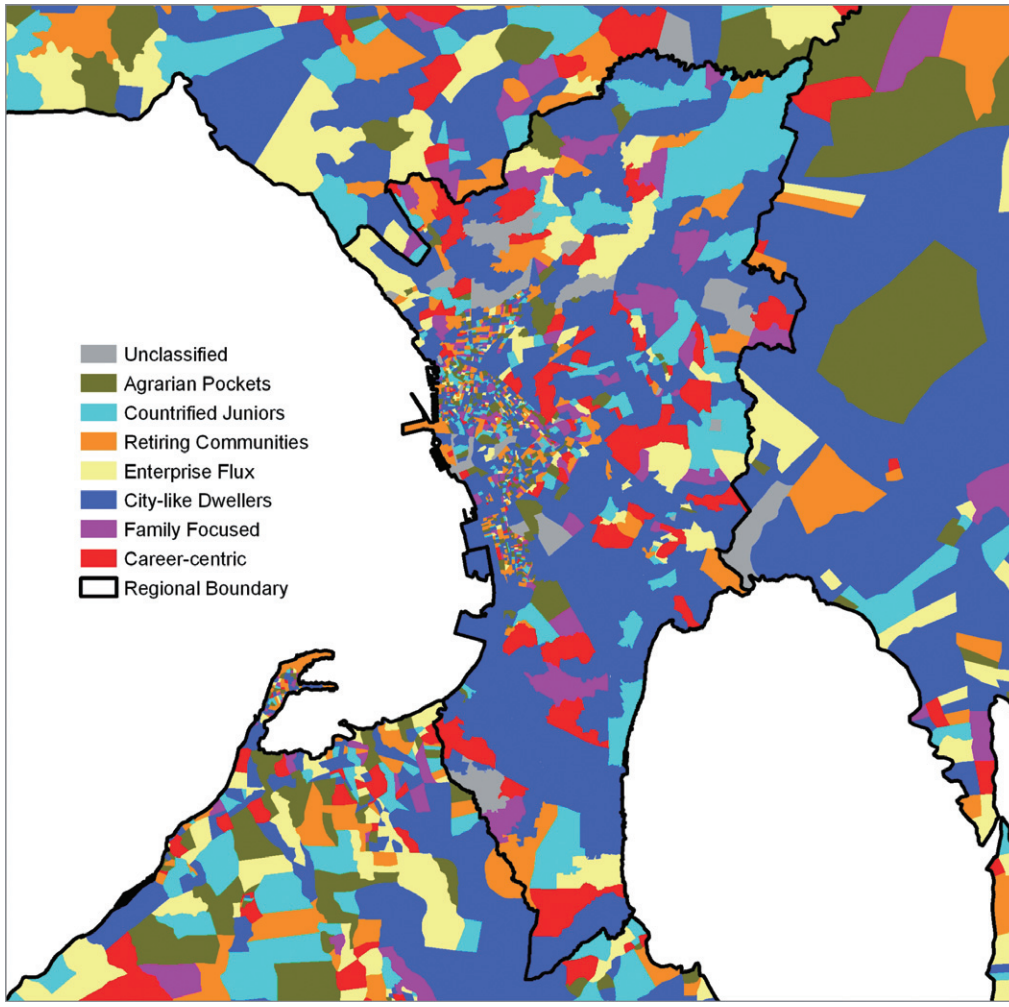

Fig. 3. Spatial distribution of super-groups in the National Capital Region

classification and in particular which are of greater importance for each cluster. This will ensure textual profiles are more meaningful and ultimately the uniqueness of different clusters can be elucidated.

Cluster profiles refer to detailed representations of the characteristics of clusters. To generate such profiles, it is important to investigate the within and between cluster variations of variables used to create the segmentation system. Radar charts and other visuals come in handy when trying to summarize such multivariate information. They are concise and relatively easy to interpret. Each of the super-groups, groups and subgroups has been profiled using the input variables contained in Table 6.

In Figure 3, we show the spatial distribution of neighbourhood types in the National Capital Region at the super-group level. As expected, there is a disproportionate concentration of city-like dwellers within the region.

In the remainder of this section, we provide textual details of the key geodemographic characteristics of all the Super-group clusters.

6.1 Agrarian pockets

Cluster 1 was named agrarian pockets because of the dominance of people who engage in agricultural and fisheries occupations and the conspicuous absence of core establishments like manufacturing. Figure 4 shows the profile of Agrarian Pockets. Among all regions, Southern

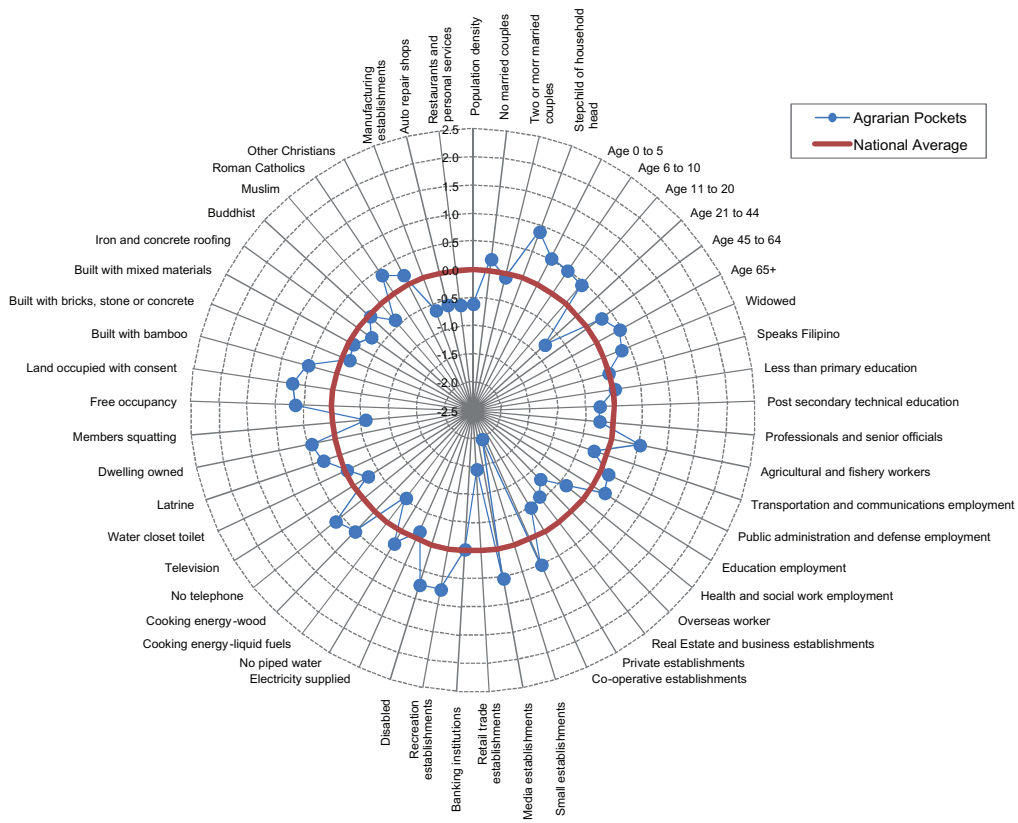

Fig. 4. Profile of agrarian pockets

Tagalog has the largest proportion of people within Agrarian Pockets. It is followed by Western Visayas.

6.2 Countrified juniors

Cluster 2 was named countrified juniors due to the rather large concentration of young children and teenagers. The super-group has the largest national proportion of children aged 0 to 5 years. The agricultural sector is the major employment sector. This also includes fishing. Figure 5 shows the profile of countrified juniors. Just like the agrarian pockets, the region with the largest population share of countrified juniors is Southern Tagalog Region. Next to Southern Tagalog in population proportion is the National capital Region.

6.3 Retiring communities

Households this cluster are dominated by people in older age categories, over 45 years and nearing the end of their working lives. Like agrarian pockets, retiring communities have a high

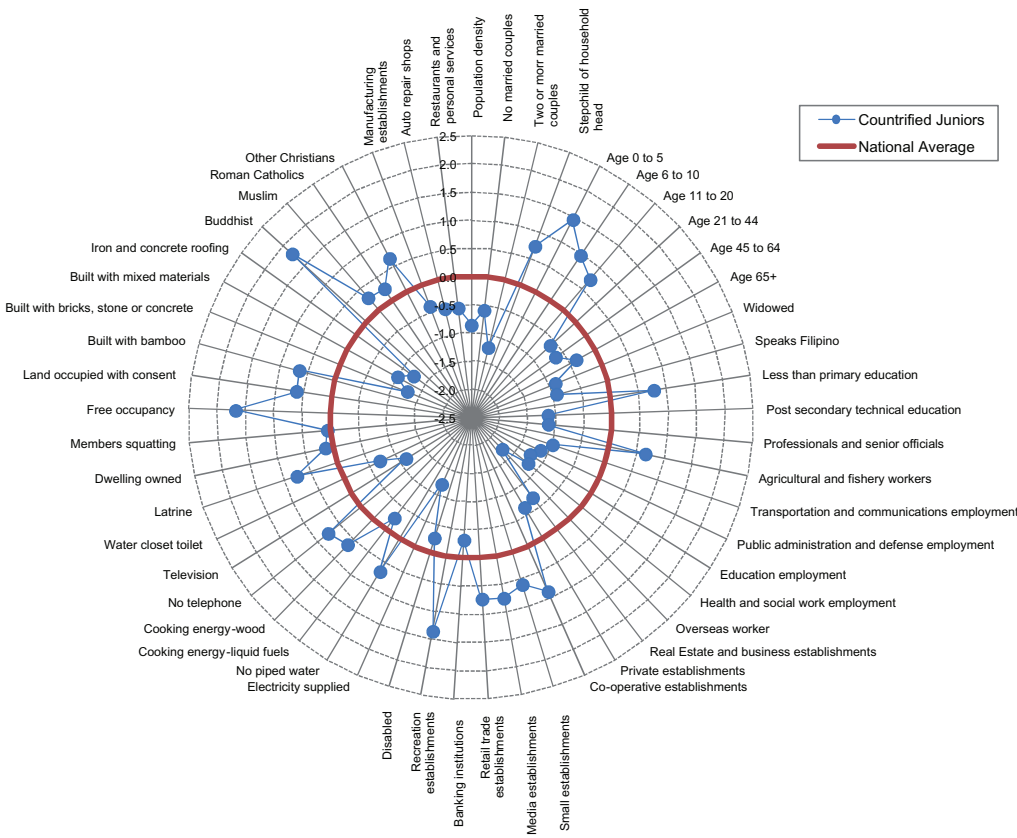

Fig. 5. Profile of Countrified Juniors

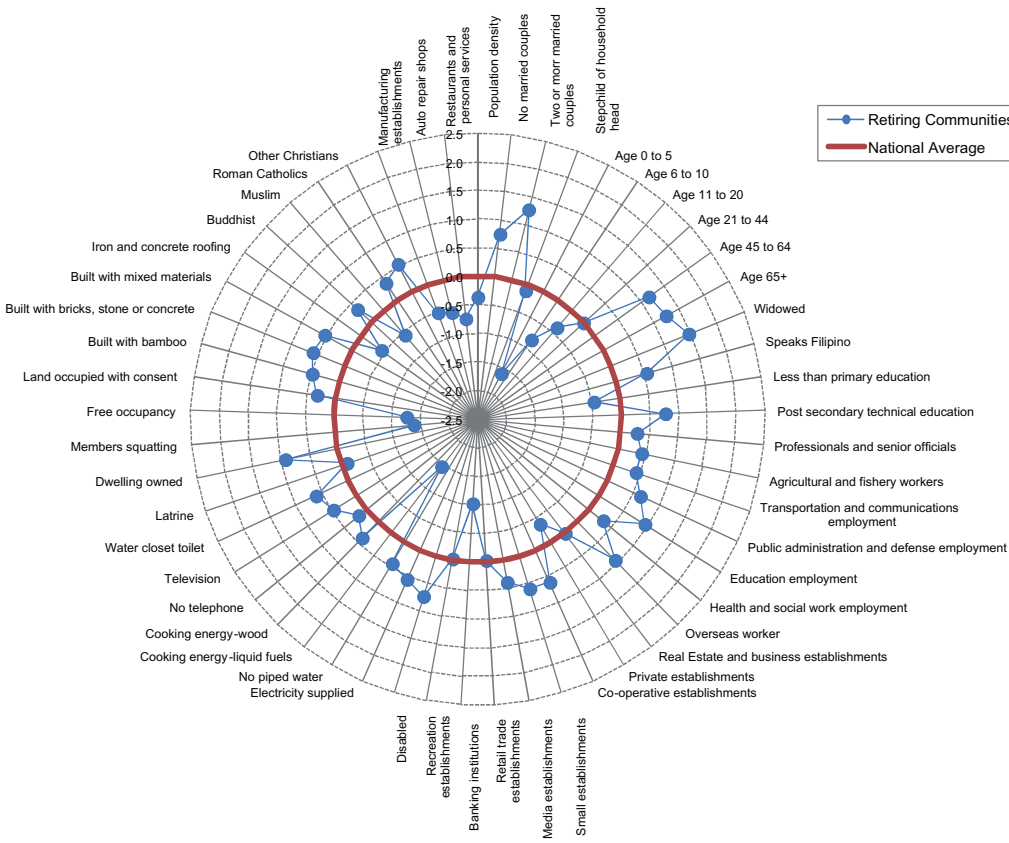

Fig. 6. Profile of retiring communities

incidence of people requiring health care for different forms of disabilities. Figure 6 shows the profile of retiring communities. The largest regional population shares for retiring communities can be found in Southern Tagalog, Central Luzon, the National Capital Region and Western Visayas.

6.4 Enterprise flux

As shown in Figure 7, this cluster is characterized by massive presence of business activities. Professionals and senior officials are present in high proportions. Transport and communications employment, the educational sector, health and social care, real estate and business, manufacturing and car repair shops employ a substantial proportion of residents. Small establishments are also quite common. Southern Tagalog is again the region with largest share of enterprise flux population. It is followed by Central Luzon and Western Visayas respectively.

6.5 City-like dwellers

Due to the level of urbanization within this cluster, city-like dwellers enjoy relatively high levels of public service facilities. However the pressure on such facilities appears to be quite high due

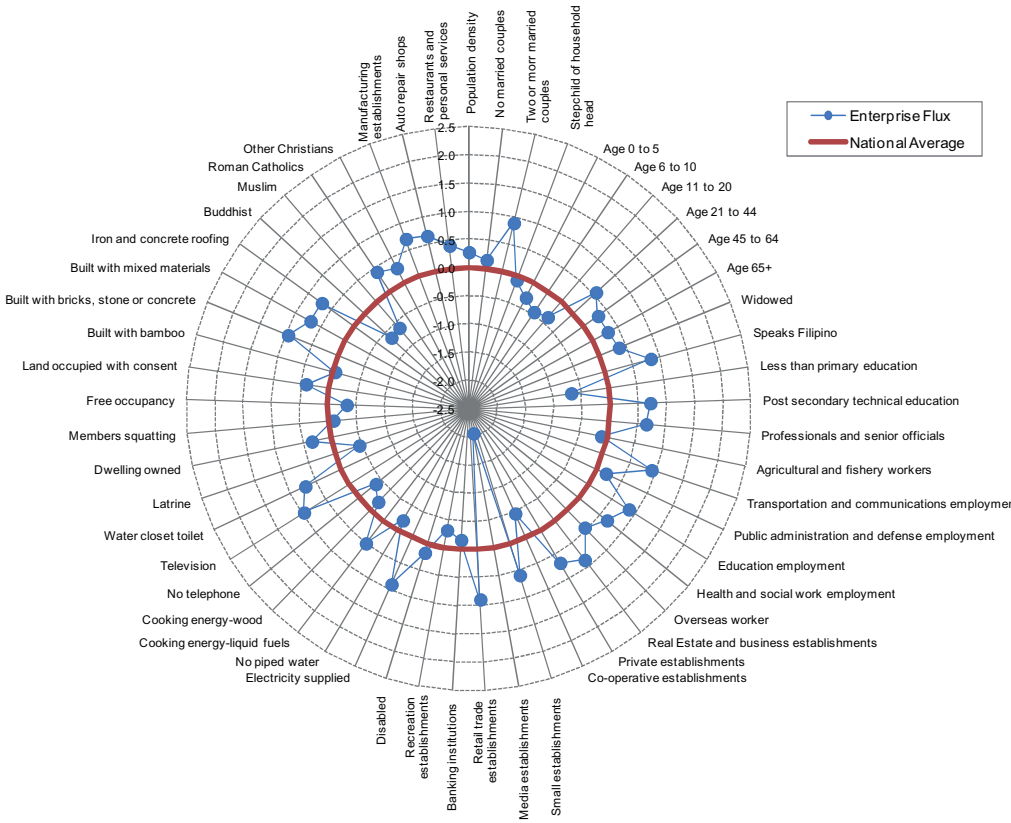

Fig. 7. Profile of enterprise flux

to very high population density. Figure 8 shows the profile of city-like dwellers. The National Capital Region has the largest population share defined as city-like dwellers. Other regions with large city-like dwellers' population concentration are Southern Tagalog, Central Luzon and Southern Mindanao.

6.6 Family focused

As shown in Figure 9, family focused is dominated by children and teenagers. The cluster has the largest incidence of people aged between 6 and 20 years. A large proportion of residents have less than primary education. To some extent, this is connected with the disproportionate concentration of young children. The largest population share of family focused can be found in the National Capital Region. It is closely followed by Southern Tagalong and Central Luzon.

6.7 Career-centric

Figure 10 shows the profile for the career-centric cluster. People are generally well educated, many up to post secondary level receiving training in technical and professional fields. It is very

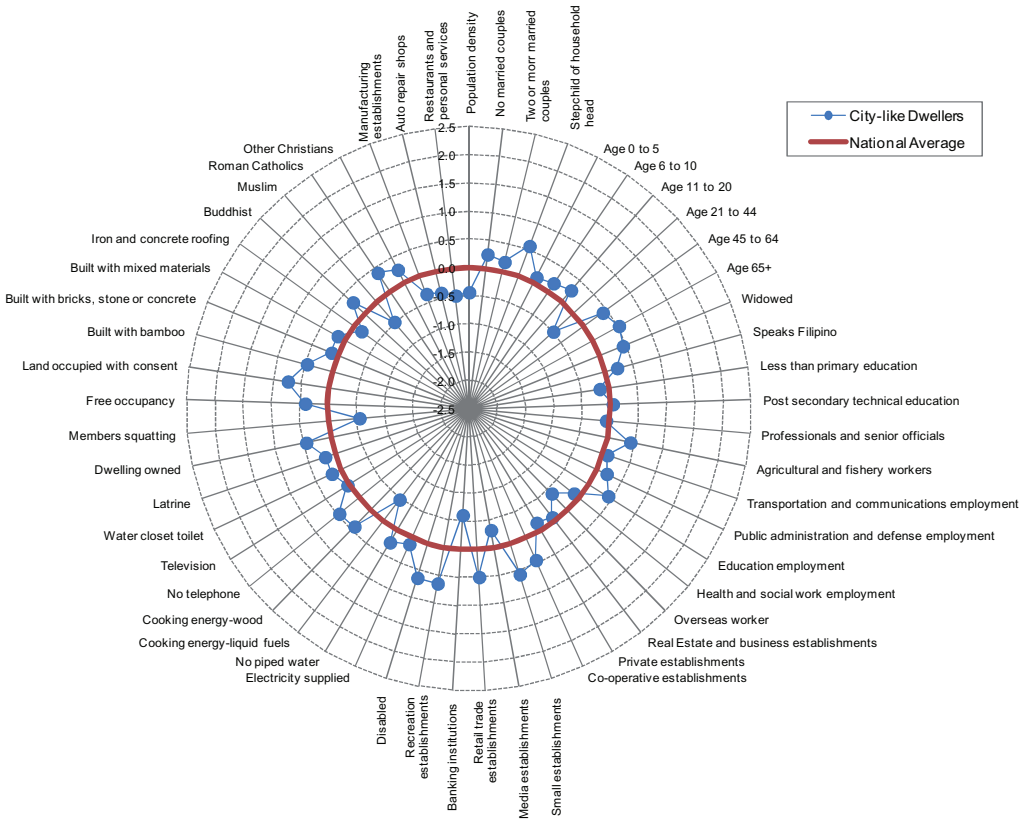

Fig. 8. Profile of city-like dwellers

common to find people employed in senior professional positions, public administration, defence, health and social work and real estate. The most dominant age group within this cluster ranges from 21 to 44 years. Career-centric areas are also very densely populated. The regions with the largest national share of people defined as career-centric are Southern Tagalog, the National Capital Region and Central Luzon.

7 Summary and conclusions

The analysis discussed in this paper has shown how a geodemographic classification system can be created by combining categorical and continuous datasets. We have also shown how such a system can be used to comprehend the spatial distribution of social groups in a developing country.

Barangays are the smallest administrative units of population aggregation in the Philippines. They therefore serve as an important basis for understanding local level inequalities when trying to address key relating to national policy debates.

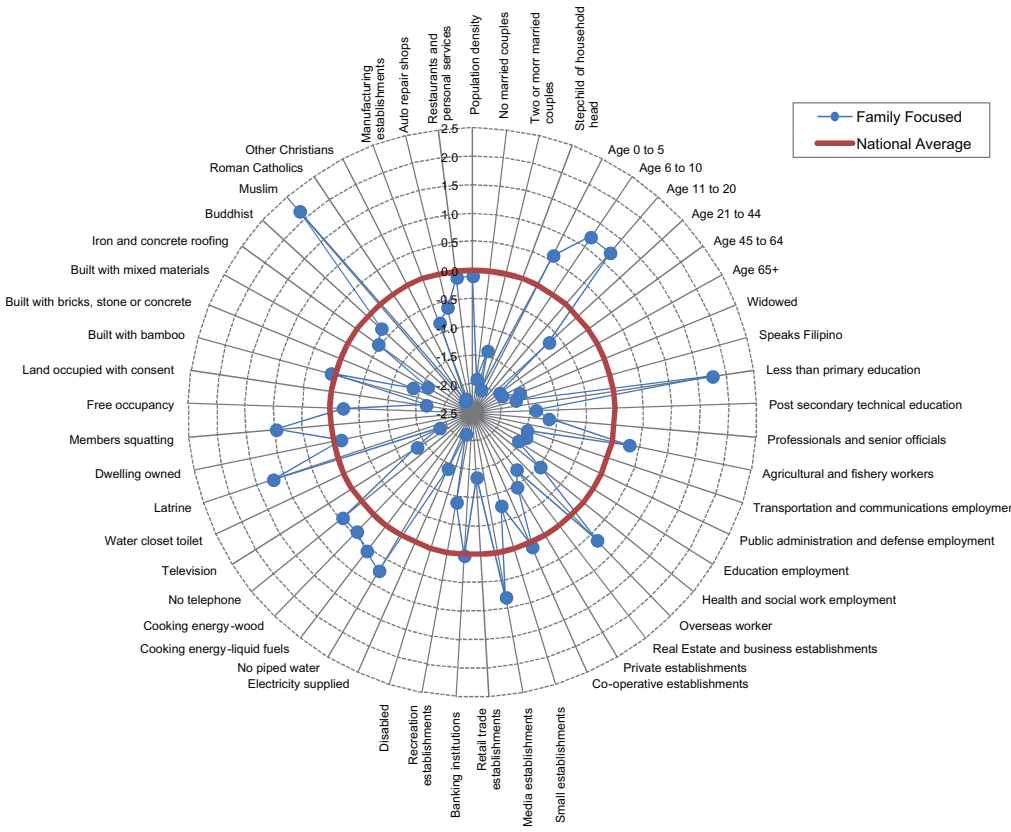

Fig. 9. Profile of family focused

Data challenges presented in the course of the analysis required the combination of categorical and continuous datasets. An important finding of the analysis is that when dealing with mixed variables, range standardization can present problems for a clustering algorithm. This finding supports the hypothesis postulated by Bacher et al. (2004).

For the first time in the Philippines, census related geospatial datasets with national coverage has been used to create a national geodemographic classification system with an open methodology. A total of 69 variables have been used to create the three tier hierarchical classification based on a two-step clustering algorithm.

It is fair to say that the results from the analysis confirm that amidst challenges surrounding availability and access to relevant datasets developing countries like the Philippines can also benefit from geodemographic methods. Apart from contributing to the literature on open geodemographics, the creation of the Philippines system will be useful in solving numerous unanswered policy related questions.

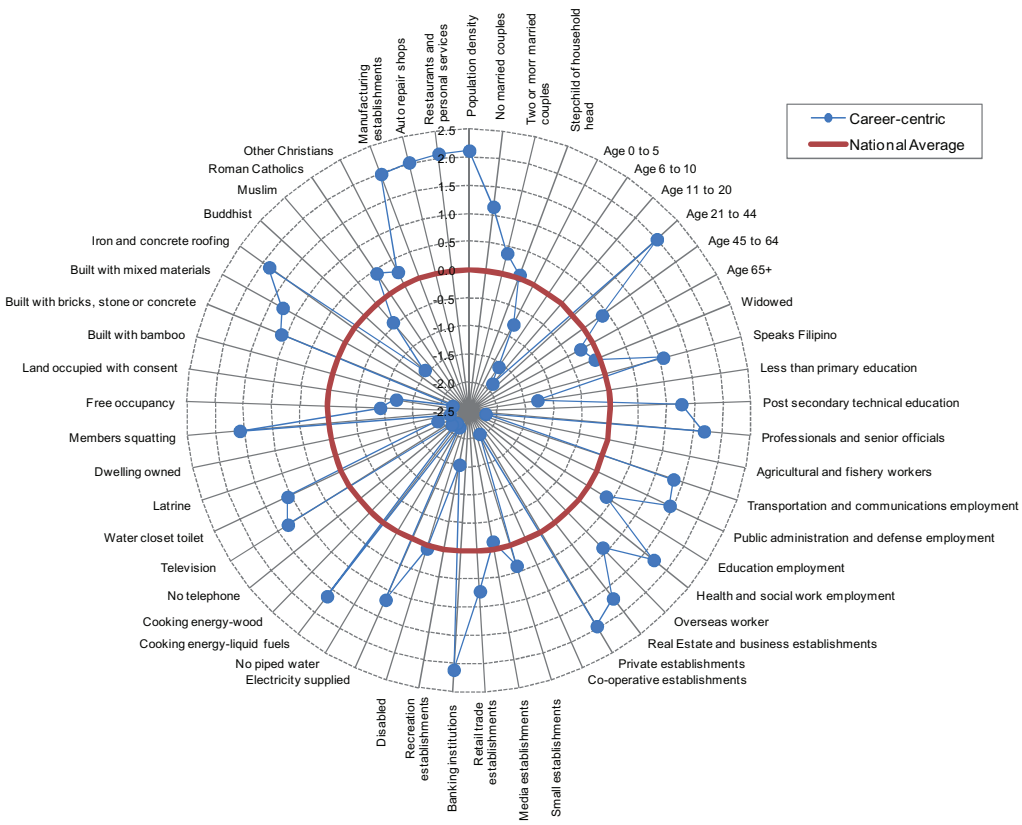

Fig. 10. Profile of career-centric

References

Agresti A (2007) *An introduction to categorical data analysis*. 2nd edn, John Wiley and Sons, New Jersey

Agresti A, Finlay B (2009) *Statistical methods for the social sciences*. 4th edn, Pearson Prentice Hall, New Jersey

Akaike H, Parzen E, Tanabe K, Kitagawa G (1998) *Selected Papers of Hirotugu Akaike*. Springer, New York

Bacher J, Wenzig K, Vogler M (2004) *SPSS twostep cluster: a first evaluation*, Arbeits- und Diskussion papiere, University of Friedrich – Alexander, Bavaria. URL: http://www.statisticalinnovations.com/products/twostep.pdf

Ballweg JA (1972) Sinfulness of Family Limitation: Interpretation vs. Practice in the Philippines. *Sociological Analysis* 33: 110–116

Banerjee S, Carlin BP, Gelfand AE (2004) *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC Press, London

Bulmer MG (1979) *Principles of statistics*. Dover Publications, New York

Crawshaw J, Chambers J (2001) *A concise course in advanced level statistics with worked examples*. 4th edn. Nelson Thornes, Cheltenham

Everitt BS, Landau S, Leese M (2001) *Cluster analysis*. 4th edn. Arnold, London

Dorling D, Ballas D (2008) Spatial divisions of poverty and wealth. In: Ridge T, Wright S (eds) *Understanding poverty, wealth and inequality: policies and prospects*. Policy Press, Bristol

Dunteman GH (1989) *Principal components analysis*. Sage Publications, Thousand Oaks, CA

George CK (1999) *Basic principles and methods of urban and regional planning*. 1st edn, Libro-Gem, Lagos

Harris R, Singleton AD, Grose D, Brunsdon C, Longley PA (2010) Grid-Enabling Geographically Weighted Regression: A Case Study of Participation in Higher Education in England. *Transactions in GIS* 14: 43–61

Harris R, Sleight P, Webber R (2005) *Geodemographics, GIS and neighbourhood targeting*. Wiley, London

Hellerstein JM, Stonebraker M (2005) *Readings in database systems*. 4th edn, Massachusetts Institute of Technology Press, Cambridge

Larose DT (2006) *Data mining methods and models*. Wiley, New Jersey
Lindblom CE, Woodhouse EJ (1992) *The policy making process*, 3rd edn, Prentice Hall, New Jersey
Longley P, Singleton AD (2009) Classification through Consultation: Public Views of the Geography of the e-Society. *International Journal of Geographical Information Science* 23: 737–763
Milligan GW (1996) Clustering validation: results and implications for applied analyses. In Arabie P, Hubert LJ, De Soete G (eds) *Clustering and classification*. World Scientific, Singapore
Montinola GR (1999) Parties and Accountability in the Philippines. *Journal of Democracy* 10: 126–140
NSO (2002) *Population expected to reach 100 million Filipinos in 14 years, 2000 census of population and housing – national reports*. National Statistics Office, available at http://www.census.gov.ph/data/pressrelease/2002/pr02178tx.html
NSO (2010) *Philippines in figures 2010*. National Statistics Office, available at http://www.census.gov.ph/data/publications/2010PIF.pdf
OECD (2008) *Handbook on constructing composite indicators: methodology and user guide*. Organisation for Economic Co-operation and Development Publishing, Paris
Ojo A (2010) *Geodemographic classification systems for the developing world: the case of Nigeria and the Philippines*. PhD dissertation. Department of Geography, University of Sheffield, Sheffield
Ojo A, Ezepue PO (2011) How Developing Countries can Derive Value from the Principles and Practice of Geodemographics, and Provide Fresh Solutions to Millennium Development Challenges. *Journal of Geography and Regional Planning* 4: 505–512
Ojo A, Ezepue PO (2012) Modeling and Visualising the Geodemography of Poverty and Wealth across Nigerian Local Government Areas. *The Social Sciences* 7: 145–158
Ojo A, Vickers D, Ballas D (2012) The Segmentation of Local Government Areas: Creating a New Geography of Nigeria. *Applied Spatial Analysis and Policy* 5: 25–49
Omelaniuk I (2005) *Gender, Poverty Reduction and Migration*. The World Bank, Washington, DC, available at http://siteresources.worldbank.org/EXTABOUTUS/Resources/Gender.pdf
Rodriguez RM (2005) *Domestic insecurities: female migration from the Philippines, development and national subject-status*. Working Paper 114. The Centre for Comparative Immigration Studies, University of California, San Diego
Schelzig K (2005) *Poverty in the Philippines: income, assets, and access*. Asian Development Bank, Manila
Singleton AD (2009) Course Choice Behaviour and Target Marketing of Higher Education. *Journal of Targeting, Measurement and Analysis for Marketing* 17: 157–170
Singleton A, Longley PA (2009) Creating Open Source Geodemographics: Refining a National Classification of Census Output Areas for Applications in Higher Education. *Papers in Regional Science* 88: 643–666
Sleight P (2004) *Targeting customers: how to use geodemographic and lifestyle data in your business*. World Advertising Research Centre, London
Urdan TC (2005) *Statistics in plain English*. 2nd edn, Lawrence Erlbaum, New Jersey
Vickers DW (2006) *Multi-level integrated classifications based on the 2001 census*. PhD dissertation. School of Geography, University of Leeds, Leeds
Vickers D, Rees P (2006) Introducing the Area Classification of Output Areas. *Population Trends* 125: 15–29
Walker JT, Maddan S (2008) *Statistics in criminology and criminal justice: analysis and interpretation*. Jones and Bartlett, London

doi:10.1111/j.1757-7802.2013.01076.x

Resumen. Filipinas es uno de los países más poblados del planeta. En términos de población, ocupa el duodécimo lugar a nivel mundial y el séptimo en Asia por detrás de China, India, Indonesia, Pakistán, Bangladesh y Japón. La población estimada del país en 2010 era de 94 millones de personas. Utilizando datos del Censo de Filipinas de 2000, este artículo presenta una discusión sobre la creación de un sistema geodemográfico jerárquico de 3 niveles para el país a escala de barangay. Los barangays son las entidades espaciales más pequeñas en la estructura de la geografía administrativa del país. Los sistemas geodemográficos más populares suelen ser desarrollados a partir de conjuntos de datos continuos. En este artículo se discute cómo se puede crear un sistema de clasificación geodemográfica mediante la combinación de conjuntos de datos categóricos y continuos. El primer nivel de la jerarquía geodemográfica de Filipinas asegura que se puede obtener un perfil general de la población, a nivel de barangay, compuesto por siete súper-grupos. Los super-grupos se subdividen primero en 24 grupos y finalmente en 66 subgrupos.

要約　フィリピンは、世界で人口の多い国の一つである。人口では世界第12位、アジアでは、中国、インド、インドネシア、パキスタン、バングラディシュ、日本についで第7位となっている。2010年の同国の推計人口は9, 400万人である。本論文は、フィリピンの2000年の国勢調査のデータを用いて、バランガイレベルでの同国の3階層の地理人口学的システムの構築について考察する。バランガイは同国の地方自治体の最小単位である。もっとも一般的な地理人口学的システムは、通常は連続データから構築される。本論文では、カテゴリーデータと連続データを結びつけることで地理人口学的分類システムをどのように構築できるかを考察する。フィリピンの地理人口学的階層の第1レベルで、人口はバランガイレベルで大きく7つのグループに分類することができる。このグループを24のサブグループに分類し、さらに66のサブグループに分類する。