Environment and Planning B: Planning and Design 2013, volume 40, pages 95 – 116

doi:10.1068/b37137

Fuzzy clustering analysis in geomarketing research

George Grekousis
Department of Environmental Resources and Engineering, College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210, USA;
e-mail: geograik@gmail.com, ggraikou@esf.edu

Thomas Hatzichristos
Department of Geography and Regional Planning, National Technical University of Athens, I. Polytechniou 9, 15786 Zografou, Greece; e-mail: thomasx@survey.ntua.gr

Received 29 October 2010; in revised form 16 June 2011

Abstract. In this study we use geographic information systems (GIS) and computational intelligence for geomarketing analysis. GIS technology offers a powerful set of tools for the input, management, and output of data, whereas computational intelligence is used for the analysis and the clustering of data by the use of unsupervised fuzzy clustering and the Gustafson–Kessel algorithm. The advantage of fuzzy geomarketing segmentation is that a customer is not assigned exclusively to one segment only, but rather with a membership value to each cluster. The proposed methodology is applied to the metropolitan area of Athens, Greece. A dataset describes 130 demographic, lifestyle, and economy variables, and the results are analysed and discussed.

Keywords: geomarketing analysis, fuzzy clustering, Gustafson–Kessel algorithm, geographical information systems

1 Introduction

Recent years have witnessed fundamental changes in the financial and retail service sector, with competition among companies intensifying as a result of changing markets. The need to cater to the specific requirements of target customer groups has become the guiding principle behind the business strategies adopted by companies. These customer needs and, subsequently, the provision of the appropriate products and services to the customer are dependent on where they live, their personal characteristics (eg, age, education, income, and assets), and their habits. The study of the values and lifestyles of consumers is an aspect of consumer behaviour research that has taken on considerable importance since the 1990s (Cliquet, 2006).

Geomarketing systems are computer systems that analyse these types of data for a better understanding of the consumer behaviour and better performing marketing strategies (Lefebure and Venturi, 2000). Many authors define geomarketing as a "specific application of the spatial economy" (Latour and Le Floch, 2001). Geomarketing systems were designed to incorporate space (geo) into the statistical analysis of demographic, lifestyle, and economic data (marketing) and are defined as a collection of tools enabling the manipulation of geocoded data, the analysis of data, the conception of strategies and decision making (Cliquet, 2006). These tools can be divided into four categories (Latour and Le Floch, 2001):

- data-analysis tools (clustering techniques, data mining);
- geographic information systems (GIS), which provide all the necessary tools for analysing and mapping data and implement space in marketing. GIS have been applied extensively in local governance, business and service planning, logistics, environmental management, and modelling (Longley et al, 2001) and more recently in marketing applications (McLuhan, 2003);

- statistical tools for predicting and modelling;
- decision-analysis tools for evaluating alternative solutions.

Clustering techniques are the core of geomarketing systems, and the clustering methods they use are of great importance for their final success. On the basis of the theory that people form clusters over space and that conclusions may be deduced for them according to each cluster, geomarketing systems allow the content of an advertising message or the product of a marketing strategy to be targeted directly on the basis of the local specificities of an intended target (Gallopel, 2006). The characterisation of clusters formed in geomarketing analysis attributes a ‘label’ to each territory based on common behaviours and habits.

Geomarketing analyses the demographic data and lifestyle data of individuals, combining them with geographic concepts such as location, distance, and contiguity, to decode the specific characteristics of each geographical area and create demographic and lifestyle profiles of the people who live in these areas (Latour and Le Floch, 2001). The strength of these systems lies in the fact that the analysis and management of a large volume of data goes beyond the classic statistical approach, which provides centralised and tabulated results. In addition, geomarketing allows the targeting of prospects with strong potential in a given trade area (Douard, 2006). On the basis of the principle that the location of residence is in part responsible for purchasing behaviour, the profile of the optimal customers can be drawn by focusing on the residents residing in specific zones. Geomarketing also allows the determination of new location sites for distributors by identifying geographic zones corresponding to the distributor’s needs. Finally, geomarketing, using GIS techniques, can serve as a tool in mapping the position of competitors and in allowing a company to keep up with competitive innovations (Douard, 2006).

Geomarketing systems are closely related to geodemographics, which is “the analysis of people by where they live” (Sleight, 2004). More specific geodemographics are based on the principle that the place and the population are linked and, as a result, knowing where somebody lives can reveal information about that person (Vickers and Rees, 2007).

Although geomarketing and geodemographics appear similar, it is important to note some slight differences. Geodemographics has its origin in the 1950s in human geography, urban sociology, and social area analysis (Shevky and Bell, 1955). Early geodemographic systems were only based on census data such as age, ethnicity, education, employment, and type of housing. From the 1970s, financial and lifestyle data were embedded on geodemographic systems to produce more-detailed segmentations to describe the consumption habits of the ‘average’ residents of a neighbourhood (Harris et al, 2005). Geodemographic analysis is based primarily on data collected by the national statistical services of each country for a very specific range of details and large geographic regions (Harris and Longley, 2004). Furthermore, the collection of lifestyle data is based mainly on surveys of public opinion. On the other hand, in geomarketing analysis a company’s internal variables are included. For example, geomarketing systems for food retailers store the information from the use of credit card payments, which reveal the products that have been bought by each customer. Geomarketing systems then connect together the location analysis with consumption activity. A consumer address makes it possible to assign a consumer to a set of external information such as residence, neighbourhood, and census description for the wider territory, which qualifies the consumer in a more precise manner. Geomarketing analysis incorporates details on lifestyle and financial data to a large extent, updating them at regular intervals and focusing on small geographic zones. In conclusion, geomarketing analysis is a more commercial-driven analysis, whereas geodemographic analysis is a more social-driven analysis. Even so, in many cases their distinction is not very clear.

An extensive bibliography on clustering techniques in geodemographic analysis mainly based on statistical techniques (Openshaw and Turton, 1996; Punj and Stewart, 1983) can be extended to geomarketing analysis.

There are two basic categories of clustering methods: the hierarchical and the nonhierarchical. Hierarchical approaches (agglomeration and divisive) are generally computationally intense and restricted in use, with small datasets. In contrast, nonhierarchical approaches, such as *k*-means, work better with large datasets and generally capture more of the variance in a dataset (Exter and Mosley, 2004). The major drawbacks to this technique are that the random starting positions can have a dramatic effect on the final outcome and the measures of similarity between the centroids and the observations are not necessarily sensitive to the variance within the data space (Exter and Mosley, 2004). An example of the use of a *k*-means algorithm can be found in Petersen et al (2010), where they examined the potential of a customised health geodemographic system for neighbourhood targeting in London, using local hospital-demand data. Research has also focused on the combination of nonhierarchical and hierarchical techniques, using methods such as autoclustering, two-stage clustering (Parthasarathy, 2003) and hierarchical self-organising maps (Hagenbuchner et al, 2003).

More recently, artificial intelligence techniques were used in spatial analysis (Grekousis and Photis, 2011) and in spatial clustering due to their ability to cluster complex data into their principal structures (Exter and Mosley, 2004). These techniques emerged in the 1990s when research in fuzzy logic and neural networks relating to pattern recognition and clustering gained prominence (Schürmann, 1996). Although a neural network classifier shares many characteristics with *k*-means, such as random starting locations or the usage of a centroid-like classification scheme, there are some distinguishing points as to how similarity measurements are calculated or how fast neural networks perform (Exter and Mosley, 2005). The Kohonen self-organising map (K-SOM) (Kohonen, 1982) is one of the most efficient neural network methods for unsupervised clustering. Spielman and Thill (2008) presented an application combining GIS with a K-SOM neural network. The K-SOM algorithm is used to develop a geodemographic clustering of a dataset describing census tracts in New York City. The K-SOM is used to extend current geodemographic practices by the formalisation of spatial relationships between geographical and social space. K-SOM, according to Openshaw (1994a; 1995), has been proved more capable for census-data clustering and is superior to *k*-means. The K-SOM method has also been used by Openshaw and Wymer (1991) to define clusters that rely on socioeconomic, demographic, and health data.

The fuzzy logic theory (Zadeh, 1965) and fuzzy clustering, especially, present many advantages because they provide the opportunity to combine many hierarchical and nonhierarchical procedures. Although there have been some applications of fuzzy clustering in geodemographic analysis, its application to more consumer-oriented data has been limited (Hatzichristos, 2004). A more extensive description of geomarketing and fuzzy clustering is presented in section 2.

In this paper we suggest a method for geomarketing analysis combining GIS and fuzzy clustering. Our main motives are to present the advantages and disadvantages of fuzzy clustering in customer segmentation analysis, to present the use of the Gustafson–Kessel algorithm (Gustafson and Kessel, 1978) for geomarketing clustering and, finally, to use a dataset with 130 demographic, economic, and lifestyle variables for the delineation of geomarketing regions in the metropolitan area of Athens. This is the first approach in Greece referred to as geomarketing analysis.

The paper is organised as follows. In section 2 we give a detailed description of the advantages and disadvantages of fuzzy clustering in geomarketing, and present research in this area. In section 3 we analyse the topic of variable selection in geomarketing systems and give details about the methods used in this study. In section 4 we present a case study, and section 5 concludes the paper.

2 Fuzziness in geomarketing systems

Geomarketing analysis performs clustering to create homogenous areas or clusters that represent reality in the best way possible. The typical process in a geomarketing analysis involves the selection of variables, the normalising or orthogonalising of the data, the classification of the data, and the cluster labelling according to centre values. According to nonfuzzy geomarketing systems, each area or neighbourhood is assigned to only one cluster (Feng and Flowerdew, 1999). However, the boundaries of classification units (eg, post codes) are administrative and do not divide people into different groups (Morphet, 1993). As a result, two neighbouring areas that belong to different clusters, when performing Boolean logic classification, are likely to share common characteristics. The interaction of neighbouring areas and their overlapping, to some extent, in their census and lifestyle profiles makes it too simplistic to allege that a unit belongs to only one cluster.

To overcome such problems, Openshaw (1989) was the first to incorporate fuzziness into geodemographics. Fuzziness refers to the inherent imprecision and ambiguity of reality (Feng and Flowerdew, 1999) and relates to geographical aspects as the neighbourhood effect (See and Openshaw, 2001). According to this phenomenon, people who live near each other tend to present similar behaviour and habits despite other differences (Openshaw et al, 1995). To achieve better geomarketing segmentations, geomarketing systems should incorporate the fuzzy nature of reality, through fuzzy clustering (Hall et al, 1992). Even though fuzzy clustering does not entirely solve these problems, it is still a very satisfactory choice for approaching geomarketing analysis.

Although fuzzy clustering has been employed in geodemographic systems (Feng and Flowerdew, 1998; 1999; Openshaw, 1989; See and Openshaw, 2001) there is a limited use of fuzzy clustering in consumer segmentation systems. Most studies concerning geodemographic analysis and fuzzy clustering have tended to focus on census data and employ the fuzzy c-means algorithm (Feng and Flowerdew, 1998; 1999; Fritz et al, 2000; Gerla, 2006; Hatzichristos, 2004; Openshaw, 1989; See and Openshaw, 2001).

The advantages of fuzzy clustering have been presented in many applications by various authors (Hatzichristos, 1999; McBratney and de Gruijter, 1992; Odeh, 1990; Openshaw, 1994b; Van Gaans and Burrough, 1993) and can be expanded into geomarketing systems.

In classical methods where there is the exclusive assignment of an object or customer to one cluster, there is the disadvantage of information loss because a crisp assignment does not delineate reality in a natural way. In fuzzy customer segmentation though, a customer is not assigned exclusively to one segment, but with varying membership values to more than one cluster. This gives a better profile of each region, avoiding the incorrect assumption that relationships between variables observed at the aggregated level (eg, a zip code) are the same at the individual level (ecological fallacy). Using fuzzy clustering, we give a more detailed description, and show that a customer of an area in one cluster presents to some extent the same behaviour as another customer in an area that belongs to a different cluster.

Fuzzy clustering produces local optimal geomarketing clusters, which maximise the initial information and minimise the error in later interventions with other statistical procedures (McBratney and de Gruijter, 1992). Furthermore, when the clustering process is repeated regularly with new data, for the same geographical zones (eg, zip codes), tendencies can be traced in zones to one cluster or another. Fuzzy geomarketing segmentation leads to spotting trends in the movements of customers from one cluster to another, so that the marketing directed towards any particular customer can be promptly recognised and reviewed accordingly.

For example, if a zip code has membership values 0.3, 0.6, and 0.1 to three clusters A, B, and C, respectively, and during the second clustering the membership values change to 0.4, 0.55, and 0.05, respectively, this demonstrates a change towards cluster A. However, in crisp clustering it would be classified in cluster B in both cases without giving us any hint of this change. With fuzzy clustering, we trace a movement that triggers questions, such as why has this happened or is it going to change to cluster A in a third clustering. The above temporal analysis may be used when cluster centres do not change dramatically from clustering to clustering, in order for the clusters to be regarded as stable as far as the general description is concerned. Although example cluster A (eg, elite) may have slightly different variable statistics from clustering to clustering (eg, increased income or more business trips), it still refers to the same class of customers. When examined in short time intervals, the temporal study of membership-value variance of an area in the clusters to which it belongs reveals tendencies to shift from one cluster to another or to strengthen its presence in a particular one.

Despite the advantages of fuzzy clustering, there are certain disadvantages. The classical methods of clustering have been used extensively in many geomarketing analysis applications with very good results, thereby gaining the trust of researchers as being based on solid mathematical foundations. In contrast, fuzzy clustering has a more idiosyncratic perspective; it combines mathematical logic with verbal logic, and its supporters have been subject to criticism over the years. Establishing the appropriate membership function and the choice of the appropriate algorithm is a process that requires experience and knowledge. Finally, fuzzy clustering, since it does not produce dual-logic answers (eg, yes or no) may lead to multiple interpretation of results biased by the researchers' beliefs.

3 Data and methods

3.1 Data

There is a plethora of and a great diversity between the variables that may be used in geomarketing systems. The final choice depends to a large degree on the problem at hand (Spielman and Thill, 2008), the system used, and the desired configuration of the socioeconomic classes (Boyer and Burgaud, 2000). The choice of variables must reflect the purpose of the clustering and is considered to be hard (Blake and Openshaw, 1995). There is an absence of suitable theory to guide variable selection (Spielman and Thill, 2008).

The key variables are child and adult age, number of persons in a household, migration, religion, ethnicity, affluence (household income and occupation) and suburban lifestyle (low household density in nonrural areas) (Jennings and Dubitsky, 2003). The combination of these categories gives a clearer picture of the socioeconomic class and consumer behaviour of an area's residents.

There are three main categories of variables used in geomarketing systems (Cliquet, 2006; Gabor, 2010):

- Demographic data, mainly collected by a national census, that refer to variables such as gender, age, ethnicity, family situation, profession, education, health, and transport means.
- Economic data that refer to variables such as income, expenses, and credit card usage.
- Lifestyle data collected from researches conducted by private companies or from analysing the data warehouses of retail companies such as supermarkets or shopping malls. Lifestyle data that refer to the way of life of the residents of an area, such as habits, free time, security, preferred consumer goods, buying frequency, and consumption frequency.

Duard (2006) suggests a more detailed categorisation of variables:

- Objective internal company variables such as customer address, sales, and products consumed.
- Objective variables describing the consumer such as age, profession, and education.
- Subjective variables such as loyal customer and good credit.
- Location variables available at regular intervals of time such as residence and workplace.
- Variables concerning space such as organisation and logistics.

Variable selection should take into account a balance between variables from different domains: for example, age and household characteristics, housing, socioeconomic characteristics, financial data, behavioural data, and variables related to the product or service surveyed (Foxall, 2010). The inclusion of pairs of variables that are highly correlated should be avoided (Vickers and Rees, 2007). From within sets of highly correlated variables, only the most predictive should be selected (Openshaw, 1995). Factor analysis and principal components analysis can be used for tracing highly correlated variables. Variance analysis (univariate or multivariate) and covariance analysis should also be investigated to identify the nature of the relationship between the surveyed variables. Examples of these methods for marketing data can be found in Foxall (2010). Variables should correlate with consumer behaviour (Cliquet, 2006) and show variation over space so that they are not heavily concentrated in a small number of geographical areas (Vickers and Rees, 2007).

It is generally considered that the combination of demographic, economic, and lifestyle data improves the descriptive classes that predict population behaviour. However, in some cases it is preferable to use only demographic or economic data, such as when establishing classes of a social or economic character within the boundaries of a geographic region. On the other hand, there are many marketing researchers who prefer the exclusive analysis of lifestyle data, providing information on an individual's subjective behaviour and bypassing the environmental factor and their position in space.

A critical aspect in both geomarketing and geodemographic systems is the openness of data and methods. The transparency in geodemographic segmentation is very important, especially when dealing with applications that raise issues of social equity in the allocation of public goods and services (see Singleton and Longley, 2008; Vickers and Rees, 2007). These systems are called open geodemographics (Singleton and Longley, 2009). The lack of this openness has led to serious limitations in the use of geodemographic classifications from the public sector because such classification usually brings obvious implications for the allocation of publicly funded resources (Singleton and Longley, 2009). On the other hand, openness in geomarketing segmentation systems is more rare because these are commercial products.

In conclusion, the choice of appropriate variables depends on the goals of the research and the aforementioned data, and is a product of research and constant observation of evolution on a social and economic level.

3.2 Methods

For the delineation of geomarketing segments we combine GIS and fuzzy unsupervised clustering, and we describe a simple three-level process (figure 1) based on the methodological framework suggested by Openshaw (1995).

3.2.1 *Level 1: GIS*

The first level, using GIS techniques and methods, defines:

- the geographic area of the study,
- the date and variables of the system,
- the spatial unit for the analysis of the results (eg, zip code),
- implementation of a spatial database,
- implementation of a descriptive database.

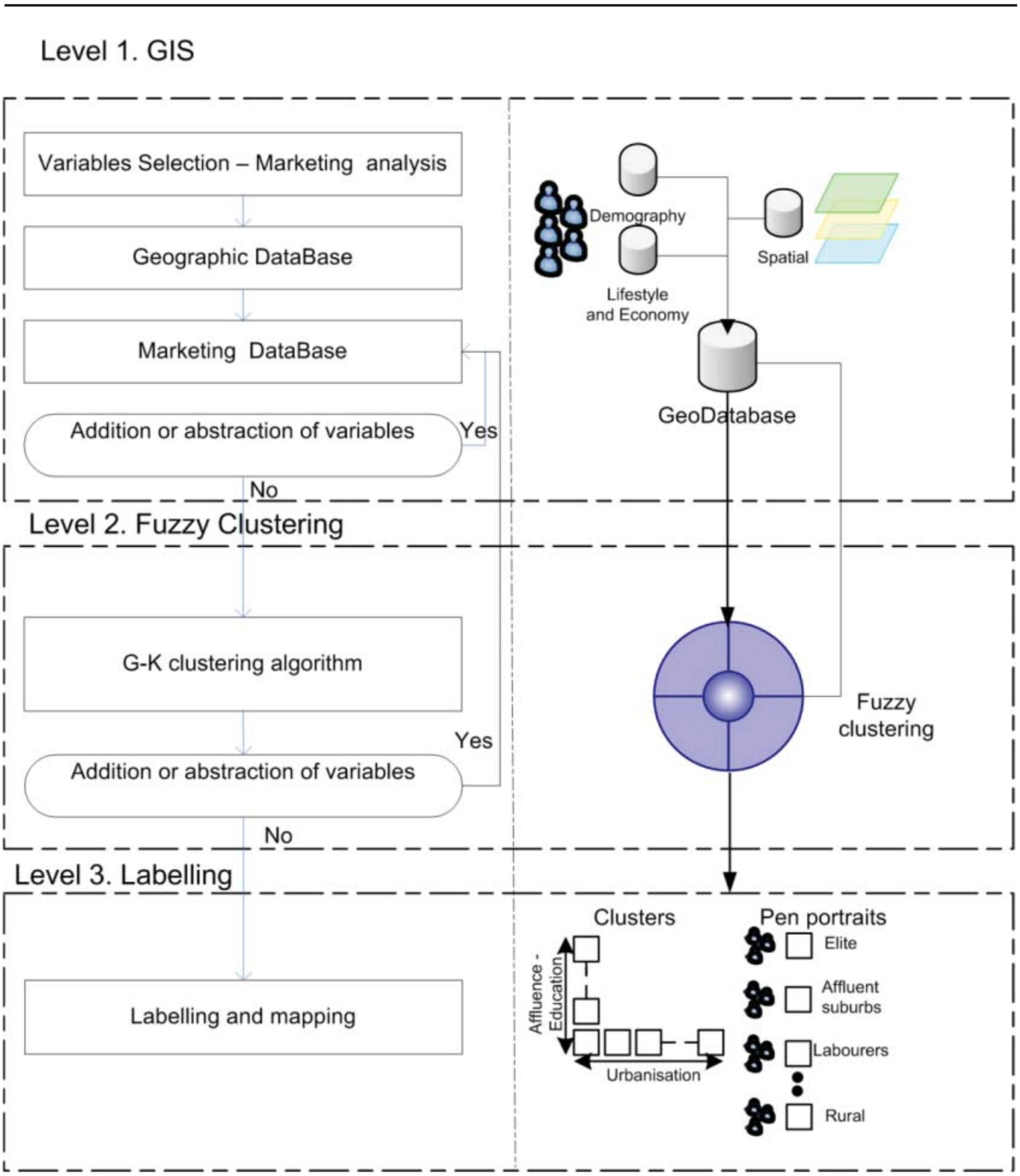

Figure 1. [In colour online.] Three-level process for the delineation of geomarketing regions.

3.2.2 *Level 2: fuzzy clustering*

The second level performs:

- Geographic segmentation of the study area with the use of a fuzzy clustering algorithm.
- Finalisation of the variables participating in the geomarketing system. There may be changes to the spatial and the descriptive database, in which case the processes return to the previous GIS level.

We use the Gustafson–Kessel algorithm (Gustafson and Kessel, 1978) and the Mahalanobis distance for the fuzzy clustering demographic and lifestyle data. The data are, typically, observations. Each observation consists of n measured variables, creating an n-dimensional vector $\boldsymbol{x}$. A set of N observations is denoted by $X = \{x_k \mid k = 1, 2, \ldots, N\}$.

The Gustafson–Kessel algorithm creates fuzzy sets $\{A_i, i = 1, 2, ..., c\}$ on a universe of data points, X. The objective function [equation (1)] that must be minimised is:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^{m} D_{ikA_i}^{2} . \tag{1}$$

This objective function may be regarded as a generalisation of the least squares method, with N as the number of data points, c as the number of clusters, D [equation (2)] as a measure of distance between the vector from the point x_k and the cluster centre v_i [equation (3)] and $\mu_{ik} \in [0, 1]$ [equation (4)] denoting the membership by the vector $\boldsymbol{x}$ in cluster c_i.

Membership values, μ_{ik}, create the fuzzy partition matrix $\mathbf{U}$, which is used for the termination criterion ε. The fuzzy partition matrix contains the membership values for each observation to each fuzzy subset A_i. The algorithm stops iterating when the norm of the difference between $\mathbf{U}$ in two successive iterations is smaller than the termination parameter ε.

$$D_{ikA_i}^{2} = (x_k - v_i)^{\mathrm{T}} A_i (x_k - v_i) . \tag{2}$$

$$v_i = \frac{\sum_{k=1}^{N} \mu_{ik}^{m} x_k}{\sum_{k=1}^{N} \mu_{ik}^{m}} . \tag{3}$$

$$\mu_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d_{ikA_i}}{d_{jkA_i}} \right)^{2/(m-1)} \right]^{-1} . \tag{4}$$

The Gustafson–Kessel algorithm creates ellipsoidal clusters of equal size using covariability arrays in calculating D [equation (2)] (Balasko et al, 2003). The objective function [equation (1)] cannot be minimised with respect to A_i, because it is linear in A_i. As a result, A_i must be constrained using the determinant of A_i (Babuska, 1998):

$$\det(A_i) = \rho_i, \qquad 0 \leqslant \rho \leqslant 1 , \tag{5}$$

where ρ permits clusters to have different sizes and its choice requires a priori knowledge about the clusters. If $\rho = 1$, then the clusters have the same size. According to Gustafson and Kessel (1978), by using the Lagrange multiplier method, A_i is:

$$A_i = [\rho_i \det(\mathbf{F}_i)]^{1/n} \mathbf{F}_i^{-1} , \tag{6}$$

where $\mathbf{F}_i$ is the fuzzy covariance matrix of the ith cluster:

$$\mathbf{F}_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^{m} (x_k - v_i)(x_k - v_i)^{\mathrm{T}}}{\sum_{k=1}^{N} (\mu_{ik})^{m}} . \tag{7}$$

When clusters exist that are considerably extended along the length of the greatest vector, the computed covariance matrix cannot estimate the distribution of the underlying data, which is something that causes arithmetical errors (Balasko et al, 2003). To overcome this problem, a scaled identity matrix can be added to the covariance matrix by changing the values (g) of the matrix from 0 to a scalar: $0 \leqslant g \leqslant 1$ (Balasko et al, 2003).

The exponent $m \in R > 1$ represents the parameter of fuzziness. It controls the weights used in the distance function and defines how fuzzy the results are, that is, the participation percentages for every point in every cluster. No theoretical or mathematical proof distinguishes the best m, but it is generally appropriate to achieve values between 1 and 3. When $m = 1$, there is no fuzziness whatsoever, and each point belongs to a single cluster. When $m \rightarrow \infty$, there is complete fuzziness, and all points display cluster membership to a considerable degree. To estimate the appropriate value for *m*, along with other parameters such as the number of clusters, several experiments have to be carried out and the results must be tested with validation indices, such as the partition coefficient (PC), the clustering entropy (CE) (Bezdek, 1981), the Xie–Beni index (XB) (Xie and Beni, 1991), the partition index (SC) (Bensaid et al, 1996), and the separation index (Bensaid et al, 1996).

Cluster validation aims to find the most natural partition that best fits the underlying data among the many partitions generated by a clustering process. An extensive review on fuzzy cluster validity indices can be found in the work of Wang et al (2007). In the case of large multidimensional datasets, effective visualisation of the dataset is also useful for evaluating the results. The Sammon mapping method (Sammon, 1969) attempts to find a low-dimensional (normally 2-dimensional or 3-dimensional) representation of a set of points distributed in a high-dimensional pattern space so that the Euclidean distance between points in the map are as similar as possible to the Euclidean distance between the corresponding points in the high-dimensional pattern space.

The algorithm is executed in the following steps (Balasko et al, 2003), given the dataset consisting of N objects, the number of clusters $1 < c < N$, the fuzziness $m > 1$, the termination criterion $\varepsilon > 0$ and the cluster volumes ρ_i. Initialise randomly the membership array $\mathbf{U}^{(0)}$.

For $1 \leqslant i \leqslant c, 1 \leqslant k \leqslant N$ and for iterations $t = 1, 2, \ldots$.

Step 1. Cluster centre calculation $v^{(t)}$

$$v_i^{(t)} = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(t-1)})^m x_k}{\sum_{k=1}^{N} (\mu_{ik}^{(t-1)})^m}, \qquad 1 \leqslant i \leqslant c .$$

Step 2. Calculation of fuzzy covariance arrays of the clusters

$$\mathbf{F}_i = \frac{\sum_{k=1}^{N} (\mu_{ik}^{(t-1)})^m (x_k - v_i^{(t)})(x_k - v_i^{(t)})^t}{\sum_{k=1}^{N} (\mu_{ik}^{(t-1)})^m}, \qquad 1 \leqslant i \leqslant c .$$

Step 3. Calculation of distance

$$D_{ikA_i}^2 = (x_k - v_i^{(t)})^t [(\rho_i \det(\mathbf{F}_i)^{1/n} \mathbf{F}_i^{-1}](x_k - v_i^{(t)}), \qquad 1 \leqslant i \leqslant c, 1 \leqslant k \leqslant N .$$

Step 4. Update membership values array $\mathbf{U}^{(t)}$ until

$$\| \mathbf{U}^{(t)} - \mathbf{U}^{(t-1)} \| < \varepsilon, \qquad 0 < \varepsilon < 1 .$$

3.2.3 *Level 3: labelling*

The third level presents results both literally (labelling) and graphically (mapping). By comparing the characteristics of the clusters we determine their key features and build a picture of the nature of the zones that fall into that category. Cluster labels and 'pen portraits' can then be derived. Pen portraits are small descriptive analyses of the clusters that draw upon their main identifiable characteristics. Furthermore, an index table is produced that provides a convenient and simple means of comparing cluster diagnostics (Batey and Brown, 1995). Index tables compare the cluster averages for a given variable against the global average (standardised to 100) across all the clusters.

Mapping fuzzy clusters is not a trivial task, and it is an active research topic that goes beyond the scope of this work. However, we suggest three series of maps for depicting fuzzy geomarketing results. The first depicts the maximum membership value per zone. Areas with high maximum values are typical for their clusters. On the other hand, areas with small maximum values are not that typical for their cluster and share common characteristics with other clusters. This type of map may be applied for the second maximum membership value and so on. Using this depiction, one can focus on areas having high second-maximum membership values and trace which areas share characteristics common to two clusters. The second series of maps depicts the membership value for each cluster separately. A graduated colour map based on the membership function is created to make a scalar depiction of fuzziness for each area in a specific cluster. In such a way, one can trace which areas have high membership values in a cluster or what the membership values of this cluster are for the surrounding areas.

Finally, defuzzification methods may be used to convert the fuzzy results to crisp results (Bezdek, 1981). Defuzzification is part of fuzzy clustering and it is a natural, and in many cases necessary, process (Ross, 2004). To depict graphically the results of the geomarketing segmentation, the simplest way is to create a map of the clusters where each region will be assigned to the cluster with the highest membership value [maximum-membership method, see Ross (2004)]. This will give a quick and easy view of the consumer segmentation, which is convenient for specific tasks such as excluding a trading area or focusing on another one. Furthermore, it is an easy way to compare with other nonfuzzy clustering techniques results such as *k*-mean clustering. To make full use of the advantages of fuzzy clustering one must examine this type of map in combination with the fuzzy clustering results and the maps of each specific region. A good review of defuzzification methods and examples can be found in the work of Ross (2004).

4 Mapping metropolitan Athens

4.1 General

Metropolitan Athens is located in southern Greece and is characterised by a concentration of a variety of economic sectors and activities, high-level public administration business services, and a population of approximately 4 million (40% of the national total) (Petrakos and Economou, 1999). The metropolitan area of Athens is marked by a highly diverse spatial distribution of ages, professions, and housing types. Until 1940 Athens was a city with mainly two-storey or three-storey buildings, and there were only 349 buildings with more than five floors (Maloutas and Karadimitriou, 2001). By 1980 34 000 buildings with more than five floors had been built in the centre of the town and the neighbouring suburbs (Leontidou, 1990). As a result of this big increase in building density, there was a diminution in lifestyle standards and many people, mostly affluent, moved to the northern and eastern suburbs of the metropolitan area (Maloutas and Karadimitriou, 2001). In 1971 only 10% of affluent people were living in these suburbs, while in 1991 30% of affluent people were living there. Between 1971 and 1991 the concentration of affluent people living in the centre of town decreased from 62% to 27% (Maloutas, 2004). On the other hand, people with lower incomes moved to the southern and western suburbs of the metropolitan area (Maloutas, 2004).

By the early 1990s, and after political changes in Eastern Europe and the Balkans, a mass wave of economic immigrants entered Greece, provoking a dramatic change in the racial and ethnic profile of the metropolitan area of Athens. The majority of the immigrants settled in the already deprived centre of Athens and in the far outskirts of the city, where there were many agricultural jobs (Lazaridis and Psimenos, 2000). Although immigrants came from all over the world, 90% were Albanian (Kandylis, 2006). This wave of immigrants pushed more people to the suburbs, making the northern and eastern suburbs more affluent while the southern and western suburbs were inhabited by people with lower incomes and less education (Kandylis, 2006).

4.2 Level 1: GIS

The database is organised on the first (GIS) level. The spatial resolution is that of zip-code zones that include more than 1000 inhabitants and are available only from the National Statistical Service of Greece (NSSG). We used the same demographic dataset from the 2001 Census as was used by Hatzichristos (2004) to delineate the demographic regions of the municipality of Athens, now extended to the metropolitan area.

Table 1. Synoptic view of system.

Demographic variables	Lifestyle and economic variables
Age groups (years) 0–4 (infants) 5–14 (children) 15–24 (young adults) 25–44 (adults) 45–64 (middle-aged) 65–74 (recently retired) 75–84 (elderly) 85+ (the aged) Foreigners Albanians Africans Australians Asians Balkanians (except Albanians) People from the USA People from other American countries Europeans People from former USSR countries Family status married single number of children Education elementary middle school high school college university Household number of members rooms Employment high-flying executives managers service providers workers	Television (hours viewing per day) ½ ½–1 1–3 3+ Newspapers political economic sports gossip Cinema (times per month) 1 2 3 >3 Entertainment cafes restaurants bars clubs Trips business holiday Car ownership none 1 2 >2 Car size (cc) 1000 1000–1200 1200–1600 1600+ Internet use (yes/no) Computers (yes/no) Income per year (€) 10 000 10 000–15 000 15 000–30 000 30 000+ Credit card use (yes/no) Expenses per year for goods (€) 7000 7000–10 000 10 000–12 000 12 000+ Loans (yes/no)

Census data are provided free of charge for research purposes only from the NSSG at the zip-code level. Zip codes present a large variance in their size and variable values. For example, the number of inhabitants of zip codes varies from 1000 to 25 000. Furthermore, the boundaries of some zip codes coincide with those of a municipality. Although this is a serious drawback for the purposes of this study, this is the only dataset available for the metropolitan area of Athens. At this geographical scale, data are generalised to a large extent, and that is why information about immigrant populations is available only for some specific ethnic groups. The variables of each zip code were normalised by dividing each variable by its population. The 285 available zip codes in the metropolitan area of Athens is not a large dataset compared with those used for studies in geomarketing and geodemographic systems in other countries.

Furthermore, economic data coming from the Ministry of Finance for 2008, such as annual income, family expenses, and loans, have also been used at the same spatial resolution. These data come from the analysis of the economic status of all tax-paying citizens living in each zip code and are very accurate.

Finally, lifestyle information was migrated to the database using consumer surveys from MRB Hellas, one of the most experienced companies in market research analysis and surveys in Greece. These data refer to consumer preferences for 2008 and were obtained from a sample of 10 000 people from the metropolitan area of Athens.

Highly correlated variables were excluded, and 130 variables were selected to produce a balance between different domains. A synoptic view of the basic variables used is given in table 1.

4.3 Level 2: fuzzy clustering

Once the GIS level is complete, the fuzzy clustering level follows. We use the Gustafson–Kessel algorithm for the customer segmentation. Several runs must be performed, with a different number of clusters specified for each run, to establish the optimum number of clusters. We calculate validation indices PC, CE, XB, and SC to define the optimal number of clusters. The values of the validity measures over the number of clusters are plotted in figure 2.

PC and CE indices display a slight change in the gradient at 12 clusters, while SC and XB validation indices are minimised at 12 clusters. As a result, the optimal cluster number is 12. After testing different values for m, g, and ρ and validating the results through validation indices, we end up performing the clustering process with values for $c = 12$, $m = 1.5$, $g = 0.5$, and $\rho = 1$. Cluster centres are depicted in figure 3, using the Sammon mapping method. In this map, the dots mark the data points, while the asterisks mark the cluster centres in a projected two-dimensional space. The contour map shows the position of the twelve ellipsoidal clusters.

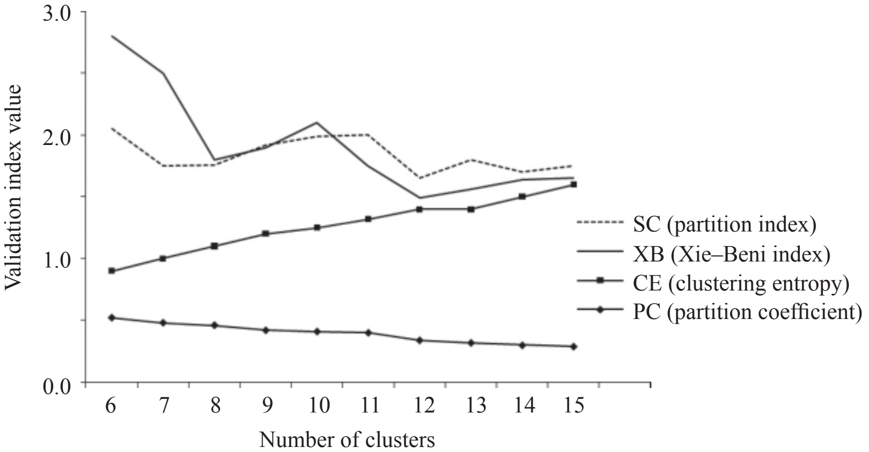

Figure 2. Values of validation indices over different numbers of clusters.

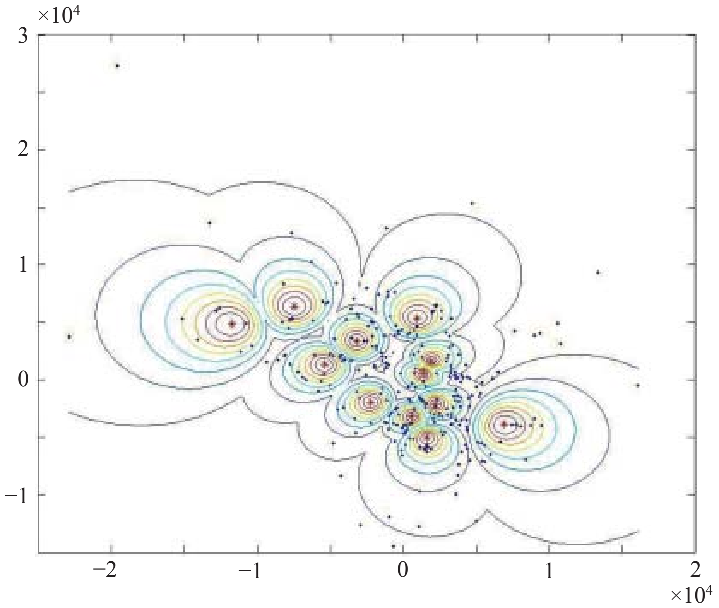

Figure 3. [In colour online.] Result of the Gustafson–Kessel algorithm, $m = 1.5$, $g = 0.5$.

The result of Sammon mapping shows a good distribution of the centres of the clusters and the data points around them.

4.4 Level 3: labelling and geomarketing results

The segmentation of the area by means of fuzzy clustering and GIS is now complete. The customer data are clustered into twelve clusters and can now be labelled on the basis of the interpretation of their cluster centres. This is required to provide a simple description of the segmentation and to help in its usage in marketing initiatives. Each cluster is described according to its dominant characteristics: for example, wealthy and educated (see table 2).

Each cluster is represented spatially in relation to the others in two-dimensional space based on the education and affluence and the urbanisation variables, as shown in figure 4. The horizontal line represents the level of urbanisation. The further to the right a cluster is on the line, the greater its level of urbanisation. Correspondingly, the further to the left a cluster is on the line, the lower its level of urbanisation. The vertical line represents the level of education and affluence. The higher the cluster is on the line, the higher the income and educational level of its residents. If a cluster is located in the centre of the graph, then its population has an average economic and educational level and lives in semiurban areas. The figure was designed to present an image of the relationships between the clusters and not the exact mathematical distance from the centres of the corresponding lines. In this way, a quick correlation can be made between the clusters, and clusters with common or opposite characteristics can be traced graphically. This would otherwise be very time consuming if it had to be done by comparing all the variables from the tables or from general descriptions of the clusters.

Each zip code is not assigned exclusively to one cluster but rather a membership value to each cluster. We depict the maximum membership value per zip code in figure 5. Zip codes coloured in white have a membership value less than 55% in their dominant cluster. These zip codes require further study to avoid assigning to them the general characterisation of the dominant cluster and instead use a combination of the characteristics of the clusters in which they participate at a high level. To gain a better aspect of the area, we depict two more views, the city centre and the southeastern suburbs. It is characteristic that the zip codes with low membership values are located in the centre, inside and around the city of Athens. In contrast, the majority of zip codes with membership values over 95% in the

Table 2. Cluster description.

Cluster	Label	Description
C1	With prospects	Sales people, assistant technicians, machine operators. Average education level. Limited income for entertainment.
C2	Elite	Elderly with large residences. High education level. Executives, academics, and artists. Luxury cars. Travel a lot.
C3	Affluent suburbs	Large houses in the northern suburbs. Families with high income. Two cars.
C4	Up and coming suburbs	Suburbs along the eastern coast. Small home-owned residences. Many family outings. Many Balkans and Asians.
C5	Prosperous	Northern and eastern suburbs. Average to high incomes. High-flying executives, scientists. Highly educated.
C6	Elderly and young workers in the centre	City centre. Small residences. Elderly, students, and workers. Average income. Old, low-capacity car. Many Africans and Balkans.
C7	Economic migrants	Asians, Latin Americans, and eastern Europeans. Young with low educational level. Lowest income.
C8	Affluent centre	Older people, high education level. Visits to theatre and concerts.
C9	Labourers	Western suburbs. Several children per family. Machine operators and unskilled labourers. Above average consumption of sport newspapers and men's magazines.
C10	Rural areas	Large spacious residences, low-income families. Farmers and employees. Many Asians and Africans.
C11	Comfortably off	Several children per family. Average educational level. Travel a lot.
C12	Industrial and manufacturing areas	Industrial areas with low-educated residents. Large owner-occupied residences.

dominant cluster are located on the outskirts. Figure 6 shows the distribution frequency of the maximum membership value per zip code. A total of eighty-two zip codes have a maximum membership value below 65%, and they belong to clusters 1, 5, 6, and 9.

Generally, the zip codes in the outskirts tend to be more concrete and clear in their demographic and lifestyle profile. On the other hand, in the centre, where urbanisation increases, zip codes are harder to cluster and they tend to belong to more than one cluster. This is a result of the large variance in size and the variable values among zip codes as referred to in subsection 4.2. This is one more reason why fuzzy clustering is better than crisp clustering. It provides an in-depth view of every cluster and zip code. Especially for these zip codes, if we assign the characteristics of the prevailing cluster to create their profile then we will lose the socioeconomic diversity of the area and fail to catch its real demography.

A series of graduated colour maps based on the membership function were created to make a scalar depiction of the fuzziness of each zip code. Figure 7 depicts the membership value of the zip codes where their predominant cluster (maximum membership value) is C2 (with prospects). In this way one can trace which zip codes have high or low membership values and identify which zip codes are typical (high membership values) for cluster C2 and which have to be delineated using descriptions from other clusters (low membership values). As we can see from the enlarged area in figure 7, there are zip codes that have high membership values in clusters C5 (prosperous) and C9 (labourers) meaning that they share common characteristics. For these zip codes we must give a more detailed profile that will combine to some extent the characteristics from the three clusters. The relevance of C1 with C5 and C9 can also be seen in figure 4, as these clusters are neighbouring each other.

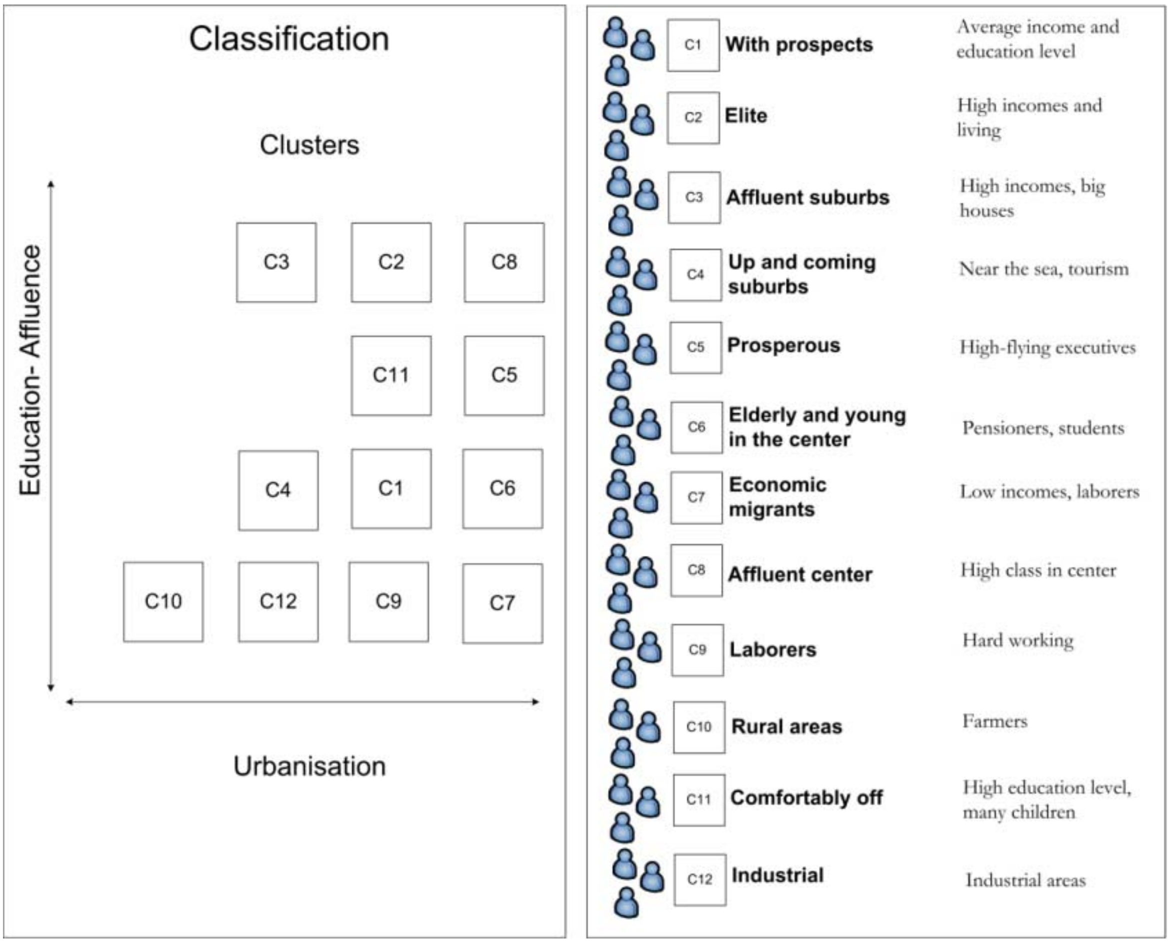

Figure 4. Mapping clustering results for twelve clusters C1–C12 by urbanisation and education and affluence.

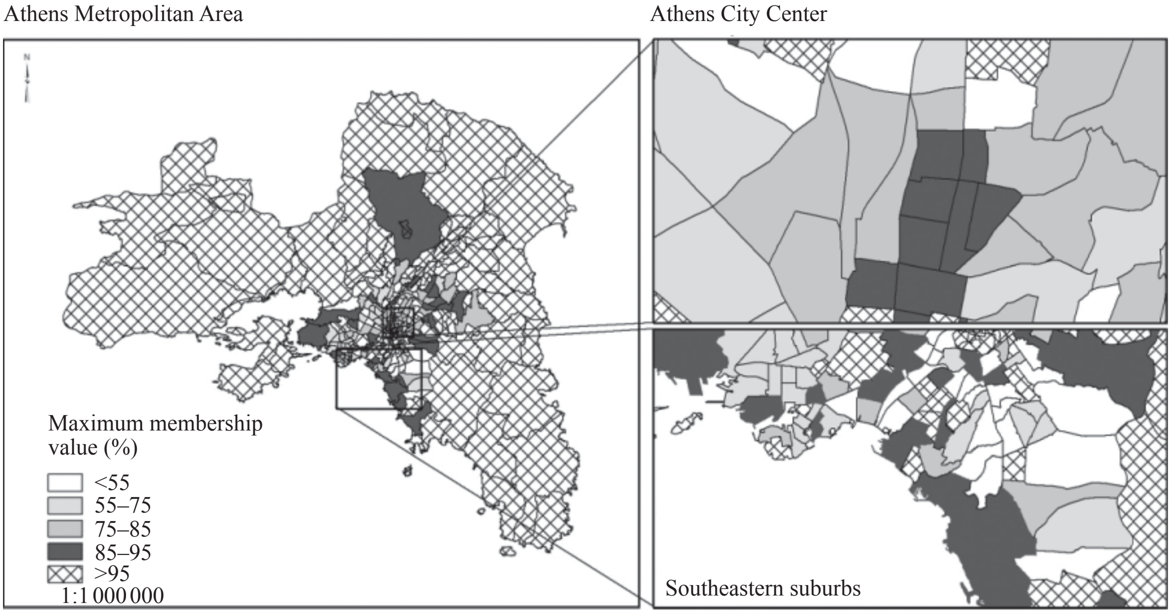

Figure 5. Maximum membership value per zip code.

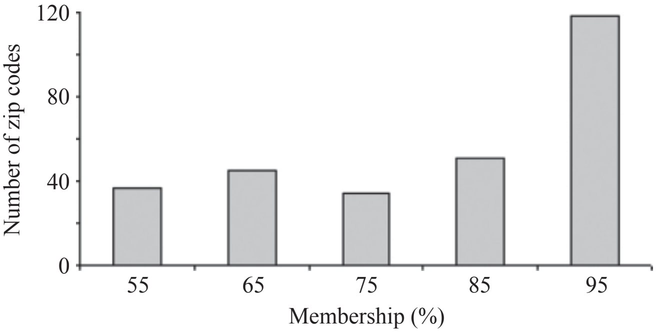

Figure 6. Distribution frequency of maximum membership value per zip code.

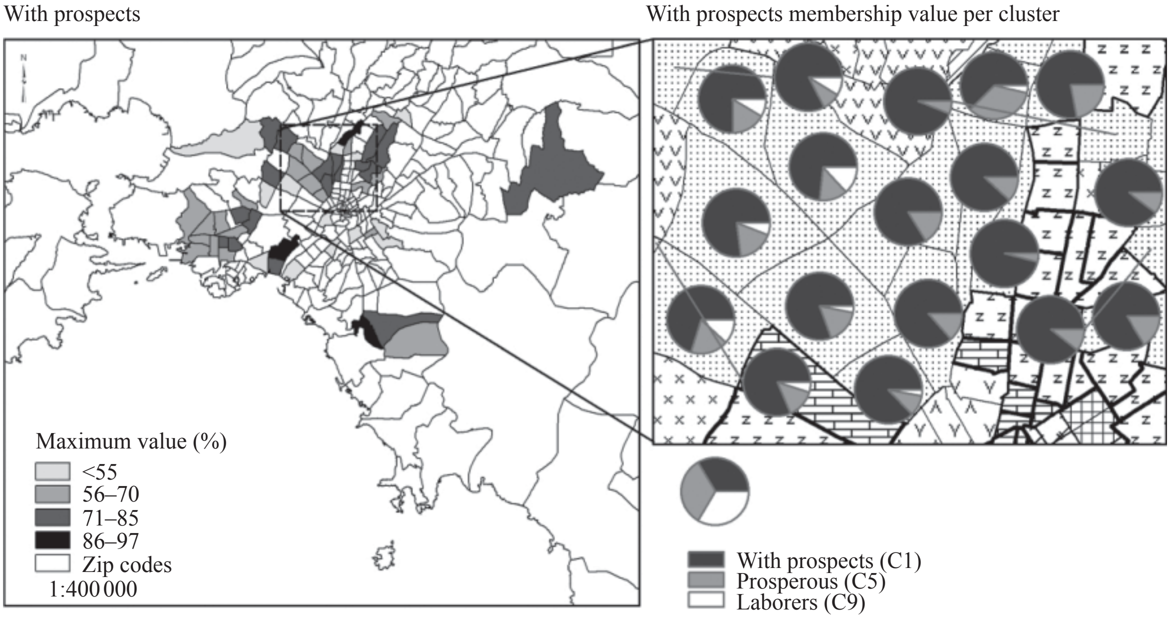

Figure 7. Membership values for zip codes of the same cluster.

Finally, to give a quick and easy view of the consumer segmentation, we create a map of the clusters where each zip code is assigned to the cluster with the highest membership value (figure 8). This is a simple but suitable presentation for tasks such as the selection of a specific trading area, comparison with other clustering methods, or simplification for nonspecialists in spatial analysis, for example, marketers.

Clusters C2, C6, C7, and C8 can be found inside the city centre and in closely neighbouring areas. Clusters C1, C11, and C9 can be found mainly in the western suburbs, while clusters C3, C4, and C5 are found in northern and eastern suburbs. Finally, C10 and C12 are found scattered far away from the city centre. Furthermore, three clusters (C1, C5, and C6) contain 53% of the zip codes, while clusters C2, C8, and C10 contain 10% of the total zip codes (figure 9).

4.5 Discussion

The results generated by the proposed methodology are consistent with other work in this area. For example, Hatzichristos (2004) used the fuzzy Kohonen algorithm for the segmentation of fifty-two postal codes in the municipality of Athens into eight demographic clusters. The clusters generated are similar to those presented in this study, for example 'prospering' with 'affluent centre C8' that are found near the Greek parliament. Furthermore, the deprived areas are highly relevant, because in both studies these areas are located in the so-called 'commercial triangle' that lies beside the historical centre.

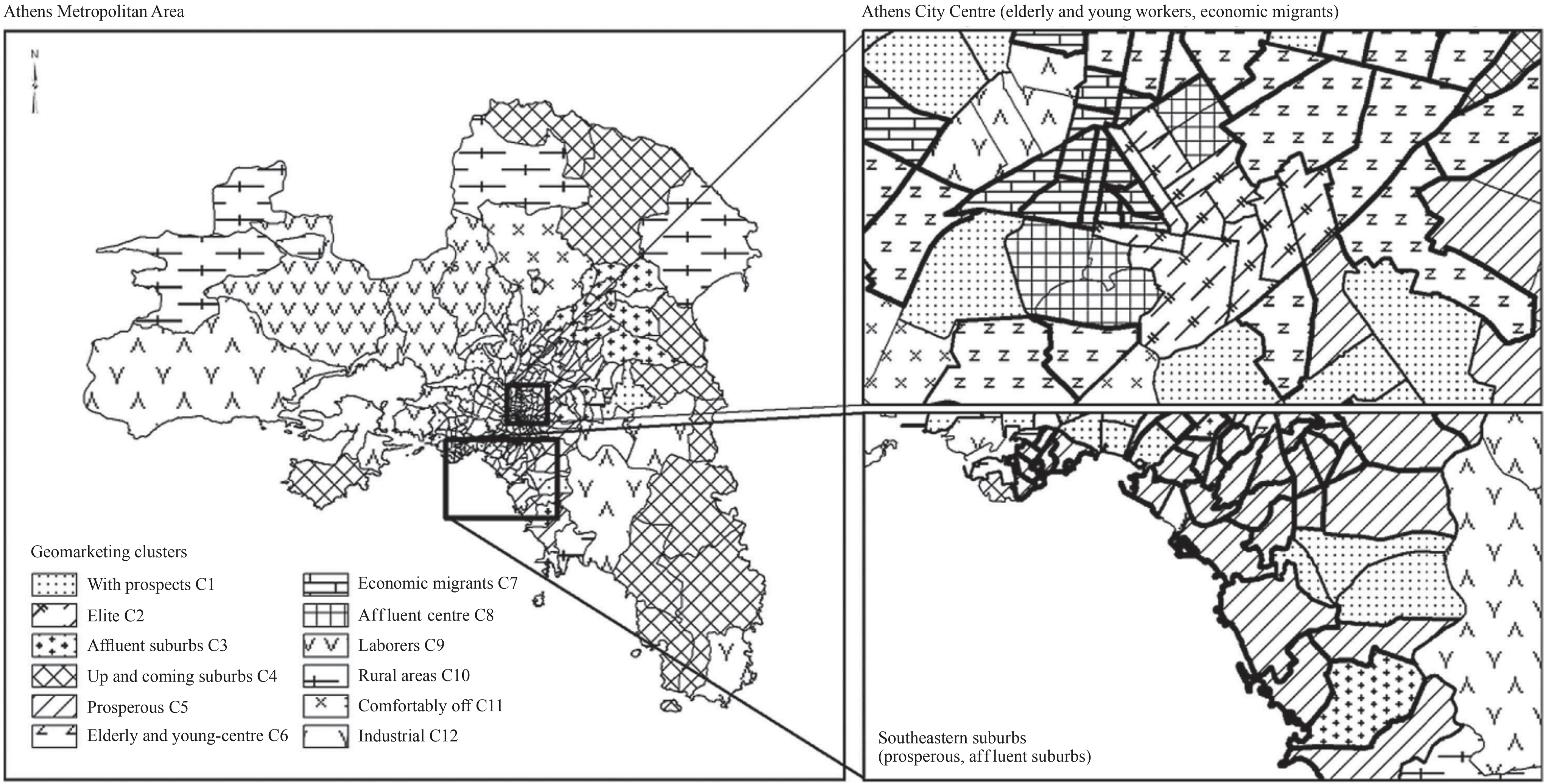

Figure 8. Geomarketing clusters.

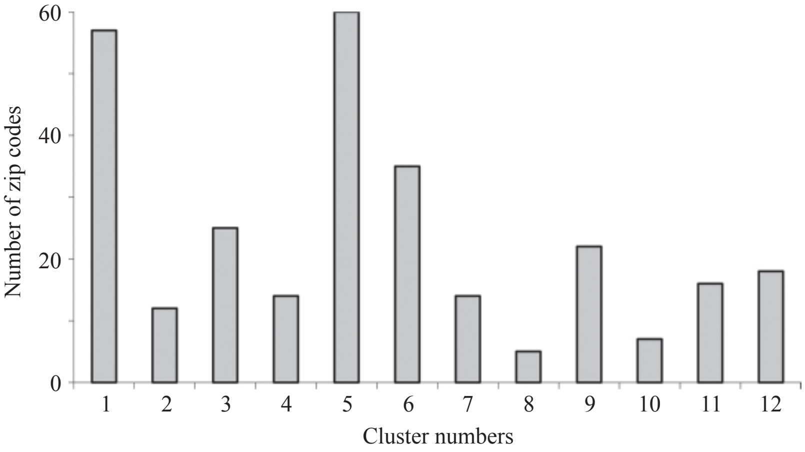

Figure 9. Number of zip codes per cluster.

Emmanuel (2004) conducted a study on the socioeconomic inequalities and housing in Athens, and many of his remarks agree with the conclusions of the present segmentation. In Emmanuel's study, which is based mainly on socially driven methods and techniques, households in Athens were clustered in four basic levels and twenty-two subgroups according to basic demographic variables such as the size of household, profession, educational level, and economic data such as annual income or consuming expenses. The results are closely related to our findings, especially in areas in the southern and eastern regions of the metropolitan area.

Kandylis (2006) studied immigrant settlements in Greece and Athens, and his study showed that there is a high concentration of immigrants in the city centre which is labelled in our study as 'economic migrants C7'. According to Kandylis, a high percentage of immigrants can also be found in regions located to the east along the coast where agricultural activity is high, something that is also described in the pen portrait of cluster (C4) in our study.

Arapoglou (2006), Arapoglou and Sayas (2009), and Maloutas (2007) studied Athens city centre disparity from the suburbs. In these studies, the inhabitants of the western suburbs and municipalities are mainly employees or unskilled labourers with low incomes. This area is labelled as 'labourers C9' in our study and is located in the same geographic region. According to the above studies, affluent people live in northern and eastern suburbs labelled as 'affluent suburbs C3' and 'prosperous C5'.

Sayas (2006) studied the urban sprawl in the periurban coastal zone of Athens, showing the demographic and socioeconomic changes over the last decades. Beriatos and Gospodini (2004) draw the new emerging landscape of Athens due to the Olympic Games of 2004. These studies present remarks and conclusions that, in many cases, agree with the results of the geomarketing analysis for the metropolitan area of Athens presented in this paper.

Maloutas and Karadimitriou (2001) studied vertical social differentiation in a small spatial entity containing sixty-two apartment buildings lying on the northern part of Athens. In their study, they provide a very interesting discussion of how apartment size and occupation status vary by floor level. For example, for their specific case study they argue that the social rank of the households is correlated with the floor of their residence. The topic of vertical social differentiation in Athens, as well as in other southern European cities, has been studied by other researchers such as Leontidou (1990), Maloutas (2007), and Arbaci (2008). These studies are based on household-level data available for small parts of the city and do not aggregate data to zip-code level. Although household-level analysis is preferred for any geodemographic and geomarketing analysis it is not always feasible since these data are not easily accessible for larger areas, such as the metropolitan area of Athens. As fuzzy clustering allows an object to belong to more than one cluster with varying membership values, a zip code may belong to more than one cluster, thus revealing differentiations. A more extensive and thorough study of these areas may lead to spotting vertical social differentiations.

For example, in our case study, the zip codes lying nearby the Athens city centre present low maximum membership values revealing a synthetic profile which combines characteristics of more than one cluster (figure 5). More specifically, the area studied by Maloutas and Karadimitriou (2001) is a part of zip code 11523 which in our study has a maximum membership value of 49.71% in C6 cluster, also sharing characteristics with C1, C5, and C11, denoting a big differentiation in this spatial unit. As far as our study is concerned, we cannot jump directly to the conclusion that this is due to vertical social differentiation. Still, a more thorough analysis, which is beyond the scope of this study, may conclude that this is the case, as argued by Maloutas and Karadimitriou. On the other hand, areas in the outer suburbs present high maximum membership values revealing limited differentiation (figure 5). These areas mainly include 1, 2, or 3 floor self-owned buildings and as a result vertical differentiation is not a big issue. As a general conclusion, fuzzy clustering reveals trends that cannot be revealed with nonfuzzy clustering techniques, which should be studied further in a second level-analysis.

5 Conclusions

In this study, GIS technology is combined with unsupervised fuzzy clustering for the delineation of geomarketing regions. Fuzzy clustering provides the user with enhanced possibilities that are not available with classical cluster analysis. The advantages of fuzzy clustering lie in the fact that each spatial unit is not clustered in a single cluster but in more than one on the basis of the membership value. The exclusive assignment of an object or customer to a cluster has the disadvantage that it loses most of the richness in the information because a crisp assignment does not accurately reflect reality. However, in fuzzy clustering each spatial unit can be analysed at a greater depth because it can combine characteristics of two or more groups.

Especially in the case of geomarketing analysis in which the interest in tracing even small declinations of the individual habits of the general trends is maximised, fuzzy clustering gives a more detailed description showing that a customer of an area in one cluster exhibits, at some percentage, the same behaviour as in another region that belongs to a different cluster. Using this method, we avoid aggregating results on the basis of only one label. Furthermore, when the clustering process is repeated regularly, tendencies can be traced in the movements of customers from one cluster to another. Finally, by mapping membership values, more information is provided than is in a dual-logic conventional map, and this information can be used as a guide for further analysis and conclusions.

As an example of the use of this approach, an application for the definition of geomarketing regions in the metropolitan area of Athens was developed. A total of 285 zip codes were clustered with a total of 130 variables (demographic, economic, and lifestyle). Clustering performed using the Gustafson–Kessel algorithm produced twelve clusters, and the results show that for this application and dataset this method is reliable, functional, and has many advantages. The methodological framework and the technological tools used in this study can be applied to other regions.

References

Arapoglou V P, 2006, "Immigration, segregation and urban development in Athens: the relevance of the LA debate for southern European metropolises" *The Greek Review of Social Research, Special Issue in English* **121** (C) 11–38

Arapoglou V P, Sayas J P, 2009, "New facets of urban segregation in southern Europe: gender, migration and social class change in Athens" *European Urban and Regional Studies***16** 345–362

Arbaci S, 2008, "(Re)viewing ethnic residential segregation in southern European cities: housing and urban regimes as mechanisms of marginalisation" *Housing Studies* **23** 589–613

Babuska R, 1998 *Fuzzy Modelling for Control* (Kluwer, Boston, MA)

Balasko B, Abonyi J, Feil B, 2003 *Fuzzy Clustering and Data Analysis Toolbox: For Use with Matlab* PhD thesis, Department of Process Engineering, University of Veszprem

Batey P, Brown P, 1995, "From human ecology to customer targeting: the evolution of geodemographics", in *GIS for Business and Service Planning* Eds P Longley, G P Clarke (GeoInformation, Cambridge) pp 77–103

Bensaid A M , Hall L O, Bezdek J C, Clarke L P, Silbiger M L, Arrington J A, Murtagh R F, 1996, "Validity-guided (re)clustering with applications to image segmentation" *IEEE Transactions on Fuzzy Systems* **4** 112–123

Beriatos E, Gospodini A, 2004, " 'Glocalising' urban landscapes: Athens and the 2004 Olympics" *Cities* **21** 187–202

Bezdek, 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum Press, New York)

Blake M, Openshaw S, 1995, "Selecting variables for small area classifications of 1991 UK Census Data", RP 95-2, School of Geography, University of Leeds, http://www.geog.leeds.ac.uk/papers/95-2/

Boyer L, Burgaud D, 2000 *Le marketing avance: du one to one au e-business* (Éditions d'Organisation, Paris)

Cliquet G (Ed.), 2006 *Geomarketing: Methods and Strategies in Spatial Marketing* translated by E Hughes (ISTE, London)

Douard J P, 2006, "Geomarketing and consumer behaviour", in *Geomarketing: Methods and Strategies in Spatial Marketing* Ed. G Cliquet, translated by E Hughes (ISTE, London) chapter 4

Emmanuel D, 2004, "Socio-economic inequalities and housing in Athens: impacts of the monetary revolution of the 1990s" *The Greek Review of Social Research* **113** (A) 121–144

Exter T G, Mosley I, 2004, "Geodemographic segmentation: new methods, new results", paper presented at the Annual Meeting of the Population Association of America, Boston, MA, http://paa2004.princeton.edu/download/ asp?submission.Id=42142

Exter T G, Mosley I, 2005, "Geodemographic segmentation: the development of PSYTE Canada", paper presented at the International Union for the Scientific Study of Population Conference, Tours, http://iussp2005.princeton.edu/ download.aspx?submissionId=50378

Feng Z, Flowerdew R, 1998, "Fuzzy geodemographics: a contribution from fuzzy clustering method", in *Innovations in GIS 5* Ed. S Carver (Taylor and Francis, London) pp 119–127

Feng Z, Flowerdew R, 1999, "The use of fuzzy classification to improve geodemographic targeting", in *Innovations in GIS 6* Ed. B Gittings (Taylor and Francis, London) pp 133–144

Foxall R G, 2010, "Invitation to consumer behavior analysis" *Journal of Organizational Behavior Management* **30** 92–109

Fritz S, See L, Carver S, 2000, "A fuzzy modeling approach to wild land mapping in Scotland", in *GIS and Geocomputation: Innovations in GIS 7* Eds P Atkinson, D Martin (Taylor and Francis, London) pp 219–230

Gabor M R, 2010, "Explanatory methods of marketing data analysis—theoretical and methodological considerations" *Management and Marketing Journal* **2** 373–384

Gallopel K, 2006, "Advertising policy and geographic information", in *Geomarketing: Methods and Strategies in Spatial Marketing* Ed. G Cliquet, translated by E Hughes (ISTE, London) chapter 16

Gerla G, 2006, "Effectiveness and multivalued logics" *Journal of Symbolic Logic* **71** 137–162

Grekousis G, Photis N Y, 2011, "A fuzzy index for detecting spatiotemporal outliers" *Geoinformatica* **16** 597–619

Gustafson D E, Kessel W C, 1978, "Fuzzy clustering with fuzzy covariance matrix", in *Proceedings of the IEEE CDC, San Diego* **17** 761–766

Hagenbuchner M, Sperduti A, Chung Tsoi Ah, 2003, "A self-organizing map for adaptive processing of structured data" *IEEE Transactions on Neural Networks* **14** 491–492

Hall G B, Wang F, Subaryono, 1992, "Comparison of Boolean and fuzzy classification methods in land suitability analysis by using geographical information systems" *Environment and Planning A* **24** 497–516

Harris R, Longley P, 2004, "Targeting clusters of deprivation within cities", in *Applied GIS and Spatial Analysis* Eds J Stillwell, G Clarke (John Wiley, Chichester, Sussex) pp 89–110

Harris R, Sleight P, Webber R, 2005 *Geodemographics, GIS and Neighborhood Targeting* (John Wiley, Chichester, Sussex)

Hatzichristos T, 1999 *Delineation of Ecoregions Using GIS and Computational Intelligence* PhD thesis, School of Rural and Surveying Engineering, National Technical University of Athens (in Greek)

Hatzichristos T, 2004, "Delineation of demographic regions with GIS and computational intelligence" *Environment and Planning B: Planning and Design* **31** 39–49

Jennings R, Dubitsky T, 2003, "How to prospect for customers using ZIP code profile models" *Quirk's Marketing Research Review* http://quirks.com/articles/ a2003/20030110.aspx?searchID=23688480

Kandylis G, 2006, "From assimilation to national hierarchy: changing dominant representations in the formation of the Greek city" *The Greek Review of Social Research* **121** 157–174

Kohonen T, 1982, "Self-organised formation of topologically correct feature maps" *Biological Cybernetics* **43** 59–69

Latour P, Le Floch J, 2001 *Geomarketing: Principles, Methods and Applications* (Éditions d'Organisation, Paris)

Lazarides G, Psimenos I, 2000, "Migrant flows from Albania to Greece: economic, social and spatial exclusion", in *Eldorado or Fortress? Migration in South Europe* Eds R King, G Lazaridis, C Tsardanidis (Mamillan, London) pp 170–185

Lefebure G, Venturi R, 2000 *La gestion de la relation client* (Éditions d'Organisation, Paris)

Leontidou L, 1990 *The Mediterranean City in Transition* (Cambridge University Press, Cambridge)

Longley P, Goodchild M, Maguire D, Rhind D, 2001 *Geographical Information Systems and Science* (John Wiley, Chichester, Sussex)

McBratney A, de Gruijter J, 1992, "A continuum approach to soil classification by modified fuzzy *k*-means with extra grades" *Journal of Soil Science* **43** 159–175

McLuhan R, 2003, "Mapping customers" *Database Marketing* **56** 21–24

Maloutas T, 2004, "Segregation and residential mobility. Spatially entrapped social mobility and its impact on segregation in Athens" *European Urban and Regional Studies* **11** 195–211

Maloutas T, 2007, "Segregation, social polarization and immigration in Athens: theoretical expectations and contextual difference" *International Journal of Urban and Regional Research* **31** 733–758

Maloutas T, Karadimitriou N, 2001, "Vertical social differentiation in Athens: alternative or complement to urban segregation?" *International Journal of Urban and Regional Research* **25** 699–716

Morphet C, 1993, "The mapping of small-area census data—a consideration of the effects of enumeration district boundaries" *Environment and Planning A* **25** 1267–1277

Odeh, 1990, "Design of optimal sample spacings for maping soil using fuzzy *k*-means and regionalized variable theory" *Geoderma* **47** 93–122

Openshaw S, 1989, "Making geodemographics more sophisticated" *Journal of the Market Research Society* **31** 111–131

Openshaw S, 1994a "Neuroclassification of spatial data", in *Neural Nets: Applications in Geography* Eds B Hewitson, R Crane (Kluwer Academic, Dordrecht) pp 53–70

Openshaw S, 1994b, "Developing smart and intelligent target marketing systems: part I" *Journal of Targeting, Measurement and Analysis for Marketing* **2** 289–301

Openshaw S, 1995 *Census Users' Handbook* (Geoinformation International, Cambridge)

Openshaw S, Turton I, 1996, "A parallel Kohonen algorithm for the classification of large spatial data sets" *Computers and Geosciences* **22** 1019–1026

Openshaw S, Wymer C, 1991, "A neural net classifier system for handling census data", in *Proceedings of the Neural Networks for Statistical and Economic Data Conference Dublin* Ed. F Murtagh (Munotec Systems, Dublin) pp 73–86

Openshaw S, Blake M, Wymer C, 1995 "Using neurocomputing methods to classify Britain's residential areas", in *Innovations in GIS 2* Ed. P Fisher (Taylor and Francis, London) pp 97–112

Parthasarathy S, 2003 *Lecture Notes for CIS 694Z: Introduction to Datamining* Ohio State University, Colombus, OH, http://www.cse.ohio-state.edu/~srini/

Petersen J, Gibin M, Longley P, Mateos P, Atkinson P, Ashby D, 2010, "Geodemographics as a tool for targeting neighbourhoods in public health campaigns" *Journal of Geographical Systems* **13** 173–192

Petrakos G, Economou D, 1999, "Internationalisation and structural changes in the European urban system", in *The Development of Greek Cities* Eds D Economou, G Petrakos (Gutenberg and University of Thessaly Publications, Athens) (in Greek) pp 13–44

Punj G, Stewart D W, 1983, "Cluster analysis in marketing research: review and suggestions for application" *Journal of Marketing Research* **20** 134–148

Ross J T, 2004 *Fuzzy Logic with Engineering Applications* (John Wiley, Chichester, Sussex)

Sammon J W Jr, 1969, "A nonlinear mapping for data structure analysis" *IEEE Transactions on Computers* **18** 401–409

Sayas J P, 2006, "Urban sprawl in the periurban coastal zones of Athens" *The Greek Review of Social Research* **121** (C) 71–104

Schürmann J, 1996 *Pattern Classification. A Unified View of Statistical and Neural Approaches* (John Wiley, New York)

See L, Openshaw S, 2001, "Fuzzy geodemographic targeting" , in *Regional Science in Business* Eds G Clarke, M Madden (Springer, Berlin) pp 269–282

Shevky E, Bell W, 1955 *Social Area Analysis: Theory, Illustrative Application, and Computational Procedures* (Stanford University Press, Stanford, CA)

Singleton A, Longley P, 2008, "Creating open source geodemographics—refining a national classification of census output areas for applications in higher education" *Papers in Regional Science* **88** 643–666

Singleton A, Longley P, 2009, "Geodemographics, visualisation, and social networks in applied geography" *Applied Geography* **29** 289–298

Sleight P (Ed.), 2004 *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business* (World Advertising Research Centre, Henley on Thames, Oxon)

Spielman S E, Thill J C, 2008, "Social area analysis, data mining and GIS" *Computers, Environment and Urban Systems* **32** 110–122

Van Gaans P F M, Burrough P A, 1993, "The use of fuzzy logic and continuous classification in GIS applications", in *Proceedings of the 5th European Conference on Geographical Information Systems Paris* Eds J J Harts, H F L Ottens, H J Scholten (EGIS Foundation, Utrecht) pp 1025–1033

Vickers D, Rees P, 2007, "Creating the UK national statistics 2001 output area classification" *Journal of the Royal Statistical Society Statistics in Society A* **170** 379–403

Wang W, Yunjie Zhang Y, 2007, "On fuzzy cluster validity indices" *Fuzzy Sets and Systems* **158** 2095–2117

Xie X L, Beni G A, 1991, "Validity measure for fuzzy clustering" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **3** 841–846

Zadeh L A, 1965, "Information and control" *Fuzzy Sets* **8** 338–353