ISSN 2389-8186
E-ISSN 2389-8194

Vol. 5, No. 2
Septiembre de 2018
doi: http://dx.doi.org/10.16967/rpe.v5n2a6

ARTÍCULOS ORIGINALES

Determination of the Representative Socioeconomic Level by BSA in the Mexican Republic

pp. 83-100

EDITH CECILIA MACEDO-RUÍZ[a], DOLORES LUQUÍN-GARCÍA[b], OMAR ROJAS[c], CARLOS LÓPEZ-HERNÁNDEZ[d]

ABSTRACT The aim of this article is to determine the socioeconomic level (SEL) with disaggregation of the Basic Statistical Area (BSA) in the Mexican Republic. The methodology used is the one established by the Mexican Association of Market Research Agencies (AMAI) along with the National Institute of Statistics and Geography (INEGI). The clustering of the BSAS was carried out according to variables contained in the Population and Housing Census of 2010, through Gaussian mixture models, learning neural networks and, finally, by defining the labels corresponding to each SEL. We found the existence of a representative SEL for each BSA. In addition, the definition of each socioeconomic level shows good results with an average of 90.86 % of correctly labeled elements.

KEYWORDS segmentation, clustering, SEL, BSA, Gaussian mixture, neural networks, labeling.

HISTORIA DEL ARTÍCULO

¿CÓMO CITAR?:
Macedo-Ruíz, E.C., Luquín-García, D., Rojas, O. & López-Hernández, C. (2018). Determination of the Representative Socioeconomic Level by BSA in the Mexican Republic. *Perspectiva Empresarial*, 5(2), 83-100. http://dx.doi.org/10.16967/rpe.v5n2a6

RECIBIDO: 10 de noviembre de 2017
APROBADO: 28 de mayo de 2018

CORRESPONDENCIA:
Dolores Luquín-García, Calzada
Circunvalación Poniente No. 49
Ciudad Granja CP 45010 Zapopan,
Jalisco, México.

a *Magíster, profesora tiempo completo en la Universidad Panamericana, México. Correo electrónico:* emacedo@up.edu.mx

b *Magíster, profesora tiempo completo en la Universidad Panamericana, México. Correo electrónico:* dluquin@up.edu.mx

c *Doctor, profesor tiempo completo en la Universidad Panamericana, México. Correo electrónico:* orojas@up.edu.mx

d *Doctor, profesor tiempo completo en la Universidad Panamericana, México. Correo electrónico:* calopez@up.edu.mx

ISSN 2389-8186
E-ISSN 2389-8194

Vol. 5, No. 2
Septiembre de 2018
doi: http://dx.doi.org/10.16967/rpe.v5n2a6

Determinación del nivel socioeconómico representativo mediante AEB en la República Mexicana

RESUMEN El objetivo de este artículo es determinar el nivel socioeconómico (NSE) con la desagregación del Área Estadística Básica (AEB) en la República Mexicana. La metodología utilizada es la que estableció la Asociación Mexicana de Agencias de Investigación de Mercados (AMAI) junto con el Instituto Nacional de Estadística y Geografía (INEGI). El agrupamiento de las AEB se llevó a cabo de acuerdo con las variables incluidas en el Censo de Población y Vivienda de 2010, a través de modelos de mezcla gaussiana, redes neuronales de aprendizaje y, por último, mediante la definición de las etiquetas correspondientes a cada NSE. Se encontró la existencia de un NSE representativo para cada AEB. Además, la definición de cada nivel socioeconómico muestra buenos resultados con un promedio de 90,86 % de elementos etiquetados de forma correcta.

PALABRAS CLAVE segmentación, agrupamiento, NSE, AEB, mezcla gaussiana, redes neuronales, etiquetado.

¿CÓMO CITO EL ARTÍCULO?
HOW TO CITE THIS PAPER?

CHICAGO:
Macedo-Ruíz, Edith, Luquín-García, Dolores, Rojas, Omar y López-Hernández, Carlos. 2018. “Determination of the Representative Socioeconomic Level by BSA in the Mexican Republic”. *Perspectiva Empresarial* 5(2): 83-100. http://dx.doi.org/10.16967/rpe.v5n2a6

MLA:
Macedo-Ruíz, Edith, Luquín-García, Dolores, Rojas, Omar y López-Hernández, Carlos. “Determination of the Representative Socioeconomic Level by BSA in the Mexican Republic”. *Perspectiva Empresarial* 5.2 (2018): 83-100. Digital. http://dx.doi.org/10.16967/rpe.v5n2a6

Determinação do nível socioeconômico representativo mediante AEB na República Mexicana

RESUMO O objetivo deste artigo é determinar o nível socioeconômico (NSE) com a desagregação da Área Estatística Básica (AEB) na República Mexicana. A metodologia utilizada é a que estabeleceu a Associação Mexicana de Agências de Pesquisa de Mercado e Opinião Pública (AMAI, na sigla em espanhol) junto com o Instituto Nacional de Estatística e Geografia (INEGI, na sigla em espanhol). O agrupamento das AEB foi feito de acordo com as variáveis incluídas no Censo Demográfico de 2010, por meio de modelos de mistura gaussiana, redes neuronais de aprendizagem e, por último, a partir da definição das classificações correspondentes a cada NSE. Foi descoberta a existência de um NSE representativo para cada AEB. Além disso, a definição de cada nível socioeconômico mostra bons resultados, com uma média de 90,86% de elementos classificados de forma correta.

PALAVRAS CHAVE segmentação, agrupamento, NSE, AEB, mistura gaussiana, redes neuronais, classificação..

Introduction

Market segmentation is one of the most popular and important tools in marketing because it allows the market to be divided into small consumer groups with similar characteristics. In this way, companies can focus on specific customers for their products and/or services in order to maximize their profits. Market stratification has become a marketing strategy that can make companies sustainable and profitable, becoming an important method for achieving targeted communication with target and/or potential customers (Hiziroglu, 2013; Murray, Agard, & Barajas, 2017; Nosi, Pratesi, & D'agostino, 2014; Pridmore & Hämäläinen, 2017). However, the effort required to find information from these potential consumers is enormous since the data required for this type of analysis are generally scattered in a large number of sources, both governmental and private. Although there are tools that are capable of combining segmentation variables together with these sources of information, these belong to private companies, such as Nielsen, Experian, Mapcity and Psyte, that charge for their use. Therefore, it is out of the reach of some small- and medium-sized enterprises.

To determine the socioeconomic level (SEL)—the group to which an individual or family belongs according to variables such as income, education, occupation, etc.—has been for many years a problem for both marketers and sociologists (Gutiérrez, 2016; Hollingshead, 1975; Murray et al., 2017; Pridmore & Hämäläinen, 2017). The closest to such an indicator is the Human Development Index, which is calculated by the United Nations Development Program. It assesses people and their capabilities as the most relevant criteria for measuring the development of a society. It is based on three dimensions: health, education, and decent standard of living, that is, the living conditions that can be met from basic goods and/or consumer services. In this way, a high human development index would mean a high quality of life (UNDP, 2017). However, some countries calculate their own indexes to measure the SEL. For example, in the United States, Socioeconomic Status (SES) conceptualizes the social position of an individual through the combination of education, income, and occupation (Berzofsky, Smiley-McDonald, Moore, & Krebs, 2014).

On the other hand, the classification of ESOMAR (European Society of Marketing Research) has been used in Europe since 1997 and recently applied to Chile (Lizana, González, Lera, & Leyton, 2017). The social classes proposed by this methodology are obtained from the occupation of the head of the household or the person who brings the most income to the household and the age at which the person finished school. On the other hand, if the head of the household is not active, the economic level of the household is determined by the possession of consumer goods such as cars, television sets, computers, etc.

ESOMAR establishes five main categories and three subcategories of social class, which refer to the main supporter of the household, ranging from professionals with a high educational level or senior management to the lowest level, that is, skilled and unskilled manual workers, and people who are engaged in agricultural and fishery work with a low educational level. This last category is divided into three subcategories (Lizana et al., 2017).

In Mexico, the Mexican Association of Market Research Agencies (AMAI) is responsible for defining the SEL index. This index is updated based on the National Household Income and Expenditure Survey (ENIGH), with this last national representation. For Mexico, seven socioeconomic levels have been defined, each with an income profile and specific consumption habit.

Based on information obtained from the Population and Housing Census 2010, it is possible to obtain an approximation of the different SELs that exist in each state of the republic. However, one of the first limitations has been the degree of confidentiality of the Census data (2010), and therefore, seeking to extend the knowledge boundary, the information collected has a disaggregation up to the level of Basic Statistical Urban Area (STUA). By using this unit of information, we sought to estimate the predominant SEL by STUA. It will finally seek to approximate the result of the SEL at the national level, comparing it with the calculations published by the AMAI.

The structure of the present article is as follows: in Section 2, we carry out a review of concepts related to the socioeconomic level (SEL), the basis for its calculation, which was established by the National Institute of Statistics and Geography (INEGI), the Gaussian mixture models, the neural networks and their labeling. Section 3 describes the way in which socio-economic levels were calculated using clusters and neural networks.

Section 4 shows the findings obtained when applying the models to the Population and Housing Census database at the STUA level. Finally, in Section 5, final comments are made with possible future lines of research.

Theoretical Framework

Segmentation

In the process of generating reliable information, it is essential that it can be used to solve problems. According to Heath (2012), the work of processing information does not end with the production of statistics, but when converting it into knowledge that serves as "input to make decisions". The methodologies used for the creation of indicators are based on procedures, which, in order to have statistical validity, should be susceptible to standardization and replication. They are, in the end, the basis for the relevant decision-making in the industry.

The process of market segmentation, since it first appeared with Smith (1956), remains practically the same to this day. It was defined by the author as the heterogeneous visualization of the market within a homogeneous number of smaller markets in response to different product preferences among consumers. In other words, market segmentation is fundamentally the acquisition of information on purchasing behavior (Allenby et al., 2002) in terms of consumer demand (Dickson & Ginter, 1987), with the aim of explaining and predicting the response of consumer purchasing (Nosi et al., 2014). In this way, companies have moved from mass marketing strategies to specific strategies for a target market (Kotler & Armstrong, 2012). To achieve a more efficient segmentation, four groups of variables are mainly used: geographic, demographic, psychographic, and behavioral. Market segmentation is an essential element in the industrialization of countries; goods and services cannot be produced without first considering the needs of consumers, while recognizing their heterogeneity (Wedel & Kamakura, 2012). The spatial-level study of market segmentation has drawn the attention of marketers in recent years, under the assumption that people are similar to their nearest neighbor in socio-demographic characteristics, lifestyle, and, of course, buying behavior (Wedel & Kamakura, 2012), as well as allowing a better understanding of the consumers' attraction towards a specific market (Cliquet, 2013).

Locating the best segment of the market in which companies or industries are to concentrate to implement effective marketing strategies (Aghdaie, Zolfani, & Zavadskas, 2013; Lopes, Machado, Rabêlo, Fernandes, & Lima, 2016; Momeni, Yazdani, & Khorshidi, 2016) is another important point that arises after segmentation. However, most of the existing literature focuses on the characteristics, techniques and validation of the optimum number of market segments and very little focuses on the selection of the best segment. Social class segmentation allows customers to be divided according to preferences for specific products (Kotler & Armstrong, 2012; Larsen, 2010), and helps to determine with more precision the attitude towards specific products of the individuals belonging to each socioeconomic level. For companies, social class segmentation is useful for determining positioning strategies (Mihić & Čulina, 2006). According to Hollingshead (1975), education level is associated with an individual's lifestyle plus prestige within the social ladder and correlates positively with more specific patterns and consumer behaviors (Bukhari, 2011; Fisher, Bashyal, & Bachman, 2012). Better education can lead individuals to better-paid jobs, which translates into more income to spend on more goods and services.

Socioeconomic level

A critical point within market segmentation for academics, marketers, and professionals has been how to find the best way to subdivide the market. Different approaches and techniques have emerged to this end. Beane & Ennis (1987) point out that there is no universal technique to segment the market; it will depend on the objective pursued and the data available. The most common forms are: geographic, demographic, psychographic and behavioral (Aghdaie et al., 2013).

Socioeconomic segmentation plays an important role in the choice of a target market. This is defined as the division of the market according to the mix of certain characteristics, such as income, occupation and education with the purpose of inferring consumers' behavior and/or lifestyle, as well as influencing the operation and performance of an organization (Kotler & Armstrong, 2012). In addition, as Hiziroglu (2013) points out, socioeconomic segmentation is important for companies because the customers of the same segment tend to require products and services customized to their lifestyle. This makes it easier for companies to determine how profitable the customers of certain segments are.

The SEL is a total measure that combines economic and sociological aspects of a person's job preparation and social position, either individually, or by family, compared to other families and/or people (Bradley & Corwyn, 2002; Gottfried, 1985; Vera-Romero & Vera-Romero, 2015) according to variables such as income, education, occupation, etc., *cfr.* Hollingshead (1975) and Gutierrez (2016). Its correct determination has not been easy for either marketers or sociologists. Not being a physical nor an easily obtainable characteristic, its adequate definition and standardization for calculation are complicated.

AMAI-INEGI methodology

The National Institute of Statistics and Geography (INEGI), in order to provide greater precision on the diverse economic and social conditions of Mexico, in conjunction with AMAI, carried out a study titled *The Socioeconomic Regions of Mexico* (INEGI, 2002). This work has not been updated and provides a poor stratification by socioeconomic level since there is only disaggregation to the local level, but municipalities can accommodate more than one socioeconomic level, while stratification by Basic Statistical Area (BSA) allows greater depth in the application of market strategies. The AMAI considers that SEL determines the well-being of households but not that of particular individuals. It does so through dimensions such as human capital, connectivity, entertainment, practical infrastructure, basic sanitation and space. Since 2011, AMAI has calculated the SEL with the 8 × 7 rule from data collected in the National Survey of Household Income and Expenditure (ENIGH). This classifies households into seven levels, which consider eight characteristics of the above-mentioned dimensions. For AMAI, SEL does not define a social class or lifestyle and does not only consider income; it defines SEL as the level of household welfare with which families meet their needs according to their economic and social well-being (AMAI, 2015). The AMAI classifies Mexican households into seven SELs (see Table 1). It is important to underline that the AMAI methodology classifies households, not people. Given that the available data are by BSA, what is sought is a representative level by this geographical unit. These questions are concrete and very specific, helping to describe the characteristics of households. Questions range from the number of light bulbs present in a house to how many cars belong to the same family.

Since it is the interest of this project to represent the different SELs in the most disaggregated way (BSA level), the sample of data used was the Population and Housing Census (2010). However, for confidentiality reasons, the most representative results at the BSA level of the Census (2010) were used. For the construction of the SEL with Census data (2010), questions were used that cover the dimensions established by AMAI:

TABLE 1. Socioeconomic Levels

SEL	DESCRIPTION
ab	It is the level with the highest standard of living in the country. Households have covered all welfare needs and are the only level that has the resources to invest and plan for the future.
c+	Households have covered all quality of life needs; however, they have limitations to invest and save for the future.
c	Households are characterized by having reached a standard of practical living with certain amenities. It has a basic infrastructure in entertainment and technology.
c-	Households are characterized by having covered the needs of space and health and by having the appliances and equipment that ensure the minimum of practicality and comfort in the home.
d+	Households have covered the minimum sanitary infrastructure of their home.
d	Households are characterized by having achieved a property but lacking most of the satisfactory services and goods.
e	This is the level with lowest quality of life or well-being. Households lack all satisfactory services and goods.

Note. Created by the authors with data from AMAI (2015).

- *Human capital:* Average schooling by BSA, which takes into consideration groups of people of equivalent ages. For this study, we took the average level of schooling of people at 15 years of age and above and a population of 18 years and above with post basic education. *Practical infrastructure:* If the houses had electricity, refrigerator, gas stove, computer and internet access.
- *Sanitary infrastructure:* If the houses had sanitation and running water.
- *Basic infrastructure and space:* Type of floor and number of rooms.

Clusters

Clustering refers to the process of dividing a set of data or objects into smaller groups. Objects that have similarities in their characteristics tend to belong to the same group, whereas those with different characteristics tend to belong to different groups (Krawczyk, 2016). In addition, it is a useful exploratory method when it comes to solving classification and segmentation problems (Aghdaie et al., 2013; Aparna & Nair, 2015).

Clustering can be classified into two different types of methods: hierarchical and non-hierarchical. Algorithms of non-hierarchical methods have been the most popular in cluster analysis, K-means being the most used because of its easy implementation, rapidity and efficiency in the clustering of large databases (Adnan, Longley, Singleton, & Brunsdon, 2010; Capó, Pérez, & Lozano, 2017; Dickson & Ginter, 1987). Its main limitation and one of the reasons why this method was discarded is that the number of clusters should be supplied as a parameter *a priori* (Aghdaie et al., 2013; Gan, Ma, & Wu, 2007). The diffuse and C-means algorithms also pertain to non-hierarchical methods, but unlike K-means, C-means is much more efficient when working with large multidimensional databases and with geodemographic information (Grekousis & Thomas, 2012; Müller & Hamm, 2014; Musyoka, Mutyauvyu, Kiema, Karanja, & Siriba, 2007), and this algorithm allows the localization of data in boundary clusters (Aghdaie et al., 2013; Momeni et al., 2016).

Unlike K-means, where data belong exclusively to a single cluster, with C-means there is the possibility that some data are located between the borders of one or more clusters and cannot be pigeonholed into only one segment (Everitt, Landau, Leese, & Stahl, 2001; Grekousis & Thomas, 2012; Ruiz, Angulo, & Agell, 2008; Sánchez-hernández, Chiclana, Agell, & Carlos, 2013).

In geographic segmentation, the data are clustered or categorized according to criteria such as neighborhoods, regions, states, countries, etc. However, the process runs the risk of overlapping clusters, making geospatial analysis inefficient. Suhaibah et al. (2016) use a variant of the K-means algorithm, called K-means ++; this variation prevents overlapping of the clusters, allowing the existence of boundary clusters. K-means ++ achieves an adequate initialization in a primary set of centers for K-means through a random seeding. This is crucial to finding an optimal grouping. In addition, this algorithm improves both the speed and the accuracy of K-means (Arthur & Vassilvitskii, 2007; Bahmani, Moseley, Vattani, Kumar, & Vassilvitskii, 2012) One of the biggest problems of cluster analysis is not to find the appropriate or more efficient method of segmentation, but rather to interpret it (Lopes, Machado, & Rabelo, 2014). The correct definition of each cluster is not a trivial task, which makes it necessary to identify each element that composes it, in such a way that a label can be assigned to each group. The labeling of clusters allows a better compression of these by being able to combine attributes and value ranges representative of each cluster (Lopes et al., 2016).

Gaussian mixture models

With the use of cluster analysis, market segmentation develops meaningful groupings of individuals or objects, with the aim of forming mutually exclusive but collectively exhaustive groups. In this way, when confronting a particular element of a group with respect to another, the limits of each grouping are seen clearly (Garza García, 1995; Hair, Black, Babin, & Anderson, 2010; Müller & Hamm, 2014; Wedel & Kamakura, 2012; Winston, 2014). Cluster analysis is considered an exploratory method and is used for the identification of market segments seeking to subsequently become strategies for companies (Wedel & Kamakura, 2012; Winston, 2014).

Mixture models have recently attracted the attention of academics and experts because they are more efficient in identifying market segments in the face of consumer heterogeneity (Brochado & Martins, 2015; Wedel & Kamakura, 2012), providing a principle-based statistical approach to determine the number of clusters present and how observations should be assigned in the available clusters (O'Hagan, Murphy, Gormley, McNicholas, & Karlis, 2016). Likewise, Andrews, Brusco, Currim, & Davis, (2010) as well as Kim & Lee (2011) found that mixture models are superior to other methods in terms of potential marketing strategies, because these models assume a function of specific density for each segment, from which it is possible to predict them more reliably. These models are a type of latent variable model that express the global distribution of one or more variables as a mixture of a finite number of component distributions; e.g. the heterogeneity of a population with respect to a set of variables is a result of the existence of one or more distinct homogeneous subgroups, or latent classes of individuals (Masyn, 2013).

Mclust-Binary is a software package based on clusters, with a classification and density based on the infinite normal mixture models that belong to the field of unsupervised learning (Pan, Shen, & Liu, 2013). According to the main assumption of these models, the data are generated from a mixture distribution, where each group is described by one or more components of the mixture (Scrucca, Fop, Murphy, & Raftery, 2016; Scrucca & Raftery, 2015), so that the estimation of the parameters is usually carried out through EM algorithms. This is an iterative process that is sensitive to the partitions of hierarchical grouping.

In addition, it implements Gaussian hierarchical clustering algorithms under the Bayesian Information Criterion (BIC) in compression strategies for clustering, density estimation and discriminant analysis. Likewise, this criterion is used to identify the optimum value and number of components of the mixture (O'Hagan et al., 2016). On the other hand, Mclust, unlike K-means or main components, has demonstrated greater accuracy and functionality to show and visualize classification and clustering results (Scrucca et al., 2016).

Neural Networks and Labeling

The correct definition of a cluster is not trivial; it is necessary to identify them, assigning a label or name to each one of them. Through techniques such as supervised and unsupervised learning and discretization of models, it is possible to achieve such a task. Clustering algorithms has as a main objective to classify the data into small groups so that they are the most similar to each other within specific metrics. However, one of the major limitations of cluster analysis has been its tendency to concentrate only on finding the correct number of segments and not on understanding the set of elements that integrate each cluster as a whole. Additionally, there is a lack of consensus in the definition of its properties, as well as the lack of formal categorization (Fahad et al., 2014). With cluster labeling, it is possible to guarantee the analysis of each attribute and identify the main characteristics that define each group, as well as the most appropriate strategies or models according to the individual needs of each cluster. For this task, the use of neural networks (ANNs), has been commonly chosen due to its learning ability, tolerance toward errors, and data organization. ANNs are known to treat nonlinear and/or dynamic problems; additionally, they have a flexible structure that makes them capable of solving a wide variety of complex problems (Hiziroglu, 2013; Lopes et al., 2014; Samarasinghe, 2016; Ultsch, 2002).

The perceptron is the most basic neural network. It is composed of an artificial neuron that receives incoming signals. That is, links are connected between input neurons and output neurons—referred to as "weights"—that facilitate a learning structure, which allows a network to freely follow the patterns of data (Samarasinghe, 2016). Once processed, these signals are offered by the neuron as an output value. The weights are commonly called free parameters, and the neural networks are therefore parametric models that involve the estimation of optimal parameters. The perceptron used is the Multilayer Perceptron network (MLP); this is a feed-forward network where there are two labels (one input and one output). Typically, the output values of one neuron serve as input only from the next neuron. MLP is used to find a possible relation between the attributes, that is to say, each one of the variables with respect to the others of the same cluster. When finding the relation, how much a variable belongs to a certain cluster is analyzed, establishing the significance and the degree of precision to predict another value. The first step of data labeling is to choose the type of data discretization for the attributes that can take different values in a specific domain, establishing themselves as new discrete values. In this way, the supervised learning algorithm will be able to identify more easily a possible relationship between the attributes that show better results in their classification in exchange for loss of information. However, discretization is not necessary in all cases.

The most commonly used methods for discretizing are: Equal Width Discretization (EWD) and Equal Frequency Discretization (EFD) (Ruiz et al., 2008; Wang & Zaniolo, 2000). The first EWD method uses some measures to discretize the data. For example, if you need to discretize the data in four ranges of values, there will be only three means of values, where the first measure is the simple arithmetic mean between the maximum and minimum values of the analyzed attribute. The second and third measures are simple arithmetic means, which can be calculated using the first average with the highest and lowest value respectively. The EWD model deals with value ranges that contain the same amount of different values among the elements provided. Given a number E of different elements and a number R of different ranges, different elements can be defined in each D (E / R). Care must be taken that E must be equal to or greater than R and both values must be greater than 0. The first range will have the lowest rated value and its maximum as the value indicated by the Dth value or ordered Dth which can be represented as [minDth].

The use of a discretization model allows the unsupervised algorithm to work with ranges of values that facilitate the detection of relevant attributes. This also makes it possible to deduce a set of values for the generated label. A supervised algorithm will be applied to each generated cluster. Here we try to detect which attributes are relevant. For each attribute of the elements of a given cluster, an ANN will be created. This ANN will be presented as an output of the estimated value for the relative attribute and will have as input the other attributes. Each ANN of the same cluster works with the same element variation only in the way that the values of its attributes are used in the neuron. Considering any cluster, the database in this step is divided into training and testing. These steps are used in a stage known as cross validation and are used by the neuron in its own learning. The evaluation process will be used to measure efficiency during the training process in relation to learning. After learning, during the testing phase, if the resulting value corresponds to the attribute range for that value, there is a hit. If not, there is an error. Therefore, each ANN is created to represent and evaluate the importance of the results. More broadly, each cluster will have a hit rate for each ANN at a hit rate for each criterion evaluated. Thus, we can know which attributes are relevant in relation to others for a given cluster. For more confidence regarding the attribute in this step, there is an average of M performances. Each time a performance (M) is executed, an ANN is created for each attribute and the final value used is the average of all M means.

After the training phase, each cluster will have the average of attributes in M performances. The ANN will have the highest average hit rates (most relevant attributes). Another variation parameter is V; it will select other attributes that have a hit rate of the majority of V in relation to the main attribute. This gives a set of attributes that can be used as relevant for the definition of each cluster. Then, it is possible to detect which range of values of the attributes is more frequent in any cluster and which is taken as relevant. Therefore, we know precisely the importance of each attribute in each cluster. Once the group of relevant attributes has been configured, it is possible to confirm which of the values dominates the group. That is, the range of the most frequent attributes in any cluster is detected where that attribute is taken as relevant. This provides precision of each attribute, as well as its possible ranges. These two pieces of data are the most important to take into account when naming the clusters.

Variables of the Model

The Socioeconomic Level Index is currently calculated by the Mexican Association of Market Research (AMAI) based on the National Household Income and Expenditure Survey (ENIGH); the latter is a national survey. This disaggregation in many cases represents a limitation because in the same geographic area it is possible to find different SELs. The AMAI 8 × 7 methodology classifies households according to the satisfaction of the needs coverage that encompasses six dimensions: human capital, planning and future, connectivity and entertainment, practical infrastructure, sanitary infrastructure, and basic infrastructure and space. From these dimensions the 8 × 7 methodology segments households into seven levels with eight indicators. In order to assign each BSA to a predominant SEL, the variables that could be adapted to the AMAI questionnaire covering the abovementioned dimensions were selected. These were taken from the Population and Housing Census (2010) at the urban BSA level.

Table 2 shows the selected variables vs. the AMAI questionnaire question.

Given the information limitations, the 2010 Population and Housing Census questions were adapted to comply with the 8 × 7 methodology. For the question: number of rooms in the household excluding bathrooms, corridors, patios and washrooms, three questions of the census were considered: total of houses with one, two and more than three rooms. For the questions regarding the exclusive use of the toilet and showers, the questions used were housing with running water in the house and sanitation. Since no information was available on the total number of light bulbs in a house, the availability of electric energy in the dwellings inhabited by BSA was taken as proxy variable. For the question on the material of the floors of the dwelling, there were no problems of identification given that this question coincides both with the AMAI questionnaire and with the 2010 Census. The question regarding the number of cars per dwelling took the proxy question of whether the household has a private vehicle. In the question concerning whether the homes have a gas stove, five types of property were taken into account that would correspond to the needs met by a stove: computer, washing machine, refrigerator, television, and internet access. Due to the lack of information on each household registered in the census, the average schooling level of each BSA and the total population over 18 years of age with post-basic education was considered, as well as the level of schooling of the population over 15 years. In this way, the AMAI 8 × 7 methodology was adapted and followed. Once the information was collected, there was a total of 58,859 urban BSAs for the entire Mexican Republic. The method used to achieve the segmentation of the different SEL types was the Mclust method, which is proposed by INEGI in the Stratifier-INEGI (2013) project. The above variables are highly correlated to the economic welfare of each BSA, so that if the indicator is close to one, the related SEL should also be related to the higher socioeconomic levels.

New variables

To obtain the seven levels defined by the AMAI in the SEL index and with the limited Census information (2010), the normal finite mixture modeling was used from the Mclust package of R. One feature of this method is that it facilitates modeling of clusters when they are formed from heterogeneous samples (Lin, Lee, & Yen, 2007). Because the BSAs are not all the same size, the information collected from each of them cannot be compared to each other. On the other hand, each variable also has a different scale, which would have undesirable effects in the calculation of the covariances. For this reason, new indicators were created from the previous variables so that the size of the BSA does not influence the weight of the variables and so that they share the same scale from 0 to 1. Table 3 shows the variables that were selected and the form in which each of the indicators used was created.

TABLE 2 . INEGI Variables Used *vs*. AMAI Questions

INEGI VARIABLE	AMAI QUESTION
Average schooling for population over 15 years of age	• Level of schooling for the head of household
Population older than 18 years with post basic education	
Availability of running water inside the house	• Bathrooms for exclusive use of the family • Availability of a shower
Availability of electricity inside the house	• Number of light bulbs in the house
Non-dirt floors	• Flooring material different from dirt
Availability of an automobile for domestic use	• Number of automobiles for domestic use
Housing with one room different to a roof, washroom, corridor and/or hallway	• Number of rooms in the dwelling different to a roof, washroom, corridor and/or hallway
Housing with two rooms different to a roof, washroom, corridor and/or hallway	
Housing with three rooms different to a roof, washroom, corridor and/or hallway.	

Note. Created by the authors with data from AMAI and Population Census 2010.

TABLE 3 . Construction of New Variables

INDEX	CONSTRUCTION
Occupied dwellings with only one room	Occupied dwellings with only one room in the bsa / (Total of occupied dwellings in the bsa)
Occupied dwellings with two rooms	Occupied dwellings with two rooms in the bsa / (Total of occupied dwellings in the bsa)
Occupied dwellings with more than three rooms	Occupied dwellings with more than three rooms in the bsa / (Total of occupied dwellings in the bsa)
Dwellings with running water	(Private dwellings with running water in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with electricity	(Dwellings with electricity in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with access to sanitation	(Dwellings with access to sanitation in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with non-dirt floors	(Dwellings with non-dirt floors in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with a refrigerator	(Dwellings with a refrigerator in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with a television set	(Dwellings with a television set in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with a telephone	(Dwellings with a telephone in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with a computer	Dwellings with a computer in the bsa / (Total of occupied dwellings-occupied private dwellings with no goods)
Dwellings with a washing machine	(Dwellings with a washing machine in the bsa) / (Total of occupied dwellings in the bsa)
Dwellings with a car	(Dwellings with a car in the bsa) / (Total of occupied dwellings in the bsa)
Average level of schooling	Average of years studied by persons over 15 in the bsa/ (Max. average of years studied in all bsas)
Population older than 18 years with secondary and tertiary education	Number of persons over 18 years old with secondary and tertiary education / (Number of persons over 18 in the bsa)

Note. Created by the authors with data from the Population and Housing Census of 2010.

Once the variables were standardized, the next step was to analyze the correlation of the variables through the correlation matrix. This matrix helps to determine the level of statistical relationship of the variables among themselves. Thus, the socioeconomic levels described by AMAI will be found. To this end, the normal finite mixture model was used because the sample of available data is heterogeneous, and these models facilitate the modeling of clusters when the data are of this nature. Once the seven socioeconomic levels were calculated, the last step was to evaluate whether the variables that make up each one of them are relevant through a label. Assigning a label to each cluster is a guarantee that each element of each segment has been analyzed and that the main characteristics that define each group have been identified.

Empirical Results

As indicated in the previous section, after standardizing the variables, the correlation matrix of the variables was obtained, in order to analyze the degree of statistical dependence of each variable. The matrix can be observed in Table 4.

As can be seen in Table 4, the correlation is high for some variables but low for the vast majority of them. This is largely due to the heterogeneity of data and variables. Because of this nature, the best fit model of clustering is the so-called normal finite mixture model, since this model captures unobserved heterogeneity in the predictive effects of the results (George et al., 2013). Table 5 shows the medians of each variable.

By using this criterion, it is possible to reduce all information contained in each of these indices

TABLE 4 . Correlations Matrix

	SCHOOLING	WATER	PC	SANITATION	ELECTRICITY	FLOORS	AUTO	PB EDUCATION	TV	REFRIGERATOR	TEL	WASHER	1 ROOM	2 ROOMS	3 ROOMS
Schooling	1	0.3	0.56	0.43	0.13	0.33	0.46	0.75	0.31	0.49	0.52	0.49	-0.29	-0.43	0.49
Water	0.3	1	0.14	0.35	0.34	0.43	0.18	0.18	0.34	0.36	0.24	0.35	-0.29	-0.24	0.38
PC	0.56	0.14	1	0.19	0.06	0.15	0.43	0.49	0.14	0.29	0.47	0.35	-0.12	-0.23	0.34
Sanitation	0.43	0.35	0.19	1	0.3	0.42	0.2	0.25	0.43	0.48	0.3	0.4	-0.28	-0.32	0.37
Electricity	0.13	0.34	0.06	0.3	1	0.41	0.05	0.04	0.61	0.46	0.16	0.31	-0.2	-0.14	0.27
Floors	0.33	0.43	0.15	0.42	0.41	1	0.24	0.19	0.52	0.48	0.24	0.43	-0.31	-0.23	0.45
Auto	0.46	0.18	0.43	0.2	0.05	0.24	1	0.45	0.26	0.43	0.35	0.53	-0.13	-0.17	0.46
PB Education	0.75	0.18	0.49	0.25	0.04	0.19	0.45	1	0.18	0.36	0.42	0.39	-0.13	-0.25	0.41
TV	0.31	0.34	0.14	0.43	0.61	0.52	0.26	0.18	1	0.61	0.25	0.49	-0.23	-0.22	0.42
Refrigerator	0.49	0.36	0.29	0.48	0.46	0.48	0.43	0.36	0.61	1	0.38	0.7	-0.25	-0.29	0.5
Tel	0.52	0.24	0.47	0.3	0.16	0.24	0.35	0.42	0.25	0.38	1	0.41	-0.23	-0.32	0.41
Washer	0.49	0.35	0.35	0.4	0.31	0.43	0.53	0.39	0.49	0.7	0.41	1	-0.23	-0.28	0.52
1 Room	-0.29	-0.29	-0.1	-0.28	-0.2	-0.31	-0.13	-0.13	-0.23	-0.25	-0.23	-0.23	1	0.5	-0.37
2 Rooms	-0.43	-0.24	-0.2	-0.32	-0.14	-0.23	-0.17	-0.25	-0.22	-0.29	-0.32	-0.28	0.5	1	-0.45
3 Rooms	0.49	0.38	0.34	0.37	0.27	0.45	0.46	0.41	0.42	0.5	0.41	0.52	-0.37	-0.45	1

Note. Created by the authors with data from Population and Housing Census 2010.

TABLE 5. Clusters and Median of Each Variable

CLUSTER	0	1	2	3	4	5	6
Schooling	0.45	0.372	0.379	0.375	0.462	0.578	0.657
Water	0.956	0.95	0.854	0.806	0.975	0.98	0.957
PC	0.257	0	0.117	0.375	0.335	0.523	0.652
Sanitation	0.985	0.983	0.907	0.889	0.994	1	1
Electricity	0.985	0.988	0.975	0.919	0.985	0.985	0.968
Floors	0.954	0.945	0.903	0.842	0.966	0.973	0.95
Auto	0.443	0.287	0.301	0.36	0.553	0.639	0.788
PB Education	0.337	0.245	0.216	0.199	0.357	0.593	0.73
TV	0.954	0.945	0.913	0.857	0.965	0.973	0.955
Refrigerator	0.873	0.812	0.757	0.7	0.922	0.95	0.942
Tel	0.429	0.194	0.228	0.231	0.489	0.681	0.667
Washer	0.705	0.619	0.547	0.529	0.788	0.844	0.857
1 Room	0.062	0.103	0.111	0.186	0.027	0.012	0.013
2 Rooms	0.16	0.25	0.238	0.31	0.118	0.045	0.041
3 Rooms	0.756	0.639	0.625	0.556	0.824	0.918	0.896

Note. Created by the authors with data from Population and Housing Census 2010.

to a single value: that of the median, which can be used to rank the BSAs according to the welfare level of each of them based on the 8 × 7 rule of AMAI. Once sorted by descending order, we analyze which percentage of the total BSAs represents a particular SEL. This is compared to the percentages of each AMAI SEL to assign to each BSA a particular SEL. The results can be seen in Table 6.

Because the SEL index is designed so that high values (close to 1) are related to high economic welfare, the SELs were assigned in descending order. Thus, by ordering the clusters obtained from highest to lowest and comparing the proportion of the population accumulated in each SEL representative of the AMAI, the results of Table 7 were obtained.

The INEGI database had information on 54,028 BSAs, but only 51,830 BSAs were considered because BSAs were excluded where the average level of schooling was equal to zero. This was done because, when analyzing the data contained in these BSAs, it was observed that most of their information was confidential, empty or zero, which did not help the algorithm to learn.

In order to ensure that there was at least a moderate accuracy, the coefficient of variation of each variable for each socioeconomic level was analyzed. The coefficient of variation allows evaluation of the quality and statistical accuracy of the estimates from the dispersion of the data with respect to its mean. The coefficient of variation is an indicator that varies in a range from 0 to 1 and allows to evaluate the statistical quality of the estimates. In an estimate, a coefficient of variation less than 0.2 is considered acceptable but should be used with caution. When the coefficient is greater than 0.2 and less than 0.5, the estimate is imprecise and should only be used for descriptive

TABLE 6. Clusters Sorted in Descending Order

CLUSTER	AB	C+	C	C-	D+	D	
Schooling	0.657	0.578	0.462	0.45	0.372	0.379	0.375
Water	0.957	0.98	0.975	0.956	0.95	0.854	0.806
PC	0.652	0.523	0.335	0.257	0	0.117	0.375
Sanitation	1	1	0.994	0.985	0.983	0.907	0.889
Electricity	0.968	0.985	0.985	0.985	0.988	0.975	0.919
Floors	0.95	0.973	0.966	0.954	0.945	0.903	0.842
Auto	0.788	0.639	0.553	0.443	0.287	0.301	0.36
PB Education	0.73	0.593	0.357	0.337	0.245	0.216	0.199
TV	0.955	0.973	0.965	0.954	0.945	0.913	0.857
Refrigerator	0.942	0.95	0.922	0.873	0.812	0.757	0.7
Tel	0.667	0.681	0.489	0.429	0.194	0.228	0.231
Washer	0.857	0.844	0.788	0.705	0.619	0.547	0.529
1 Room	0.013	0.012	0.027	0.062	0.103	0.111	0.186
2 Rooms	0.041	0.045	0.118	0.16	0.25	0.238	0.31
3 Rooms	0.896	0.918	0.824	0.756	0.639	0.625	0.556

Note. Created by the authors with data from Population and Housing Census 2010.

TABLE 7. Comparison of SEL and Clusters *vs*. AMAI

SEL	CLUSTER	TOTAL BSAS	PERCENTAGE	AMAI 8X7 FOR LOCALITIES OF MORE THAN 50,000 INHABITANTS
ab	6	3,776	7.3	6.8
c+	5	8,070	15.6	14.2
c	4	8,762	16.9	17.0
c-	0	8,929	17.2	17.1
d+	1	9,589	18.5	18.5
d	2	9,952	19.2	21.4
e	3	2,763	5.3	5.0

Note. Created by the authors with data from Population and Housing Census 2010 and AMA purposes. On the other hand, if the coefficient is greater than 0.5, there is no statistical precision in the estimate. Below is presented table 8 with the coefficients of variation of each of the variables by SEL. The lighter color represents acceptable probability, while the intermediate color indicates its accuracy is moderate, and the darker one shows that there is no precision.

The accuracy of the variables as a whole was either acceptable or moderate 73 % of the time, which indicates that the construction of the clusters was correct (INEC, 2013, Levin & Rubin, 2004). Finally, after having estimated the clusters through unsupervised algorithms, a supervised algorithm was applied in each of the generated clusters. It was intended to detect relevant variables in each segment. Data labeling was implemented using Python. The parameters present in the approach were selected after a series of preliminary tests. In these tests, different variations were used in the parameter values until it was estimated with the help of the literature (Lopes, Machado, & Rabelo, 2014; Vajda, Rangoni, & Cecotti, 2015) which parameters were optimal for the data set. The results were as follows: the parameter M was 10 and refers to the number of iterations performed to obtain the means of the ANNs. Since the discretization method was of the EWD type, the V was 15 and the R was 4.

Identification of the socioeconomic level

The database comes from the Population and Housing Census 2010, with a total of 51,830 observations, which are divided into seven groups. Each cluster is composed of 15 variables, which are the proxy described above. 75 % of the data was used for the training phase and 25 % for the testing phase. The clustering method followed was the finite normal mixture model through the Mclust package. The results obtained are shown in Table 9. The first column mentions the variable (also called attribute) that is being analyzed, the second contains the number of elements, which in this case are the BSAs. The third column, analysis, indicates the average percentage of relevance of the attribute, i.e., how important that variable is with regards to the group. In other words, it represents the average success rate of the neural networks of an attribute relative to its group. The range is the label assigned to each variable by the learning of neural networks and is based on the values of the database. It is based on these ranges comparing them with the percentage of analysis, which establishes the importance of each variable or attribute in a given cluster. Errors are the absolute number of times that the neuron does not hit a certain attribute, whereas the hits represent

TABLE 8. Coefficients of Variation

CLUSTER	AB	C+	C	C-	D+	D	E
Schooling	0.19	0.12	0.19	0.16	0.26	0.18	0.24
Water	0.02	0.09			0.22	0.03	0.17
PC	0.2	0.27	0.67	0.43	1.71	0.66	0.77
Sanitation	0.13	0.02	0.1	0.04	0.25	0.11	0.22
Electricity	0.15	0.26	0.32	0.25	0.86	0.51	0.61
Floors	0.18	0.22	0.53	0.41	0.98	0.48	0.78
Auto	0.12	0.01	0.13	0.02	0.37	0.2	0.33
PB Education	0.13	0.02	0.2	0.07	0.46	0.41	0.51
TV	0.13	0.02	0.09	0.02	0.23	0.08	0.2
Refrigerator	0.13	0.04	0.13	0.09	0.39	0.23	0.37
Tel	0.2	0.16	0.56	0.32	1.13	0.77	1.01
Washer	0.09	0.2	0.14	0.21	0.53	0.36	0.49
1 room	1.93	1.12	2.4	0.71	1.38	0.73	1.05
2 rooms	1.1	0.84	1.15	0.37	0.89	0.36	0.74
3 rooms	0.14	0.08	0.22	0.13	0.45	0.24	0.42

Note. Created by the authors with data from Population and Housing Census 2010.

Por favor revisar las cifras y su ubicación, la tabla no estaba bien estructurada y se alteraron las columnas the degree of precision of the label, which is the relative number of times the neuron hits a certain attribute.

Table 9 shows that the variables (attributes) that define the highest socioeconomic level are housing in terms of having 1 or 2 rooms, although one might think that it should contain the variable 3 or more rooms. Regarding the availability of basic satisfiers, sanitation, running water, electricity, and different types of flooring are representative indicators to conform this level. There is also a high incidence of other satisfiers, such as a television set and a refrigerator.

Both the socioeconomic level AB and the C+ have in their description of the label a large number of variables. According to AMAI, the higher the socioeconomic level, the greater the number of satisfiers that help to cover needs.

Table 10 shows the results of the SEL C+; it coincides in all the attributes with the level AB, but also the attributes of a washing machine and 3 rooms are included. That is, although for SEL AB it is not representative to have 3 rooms, for SEL C+ it is.

The next SEL, which corresponds to the typical C, the middle class, have barely covered their needs. Here only the variables of electricity, 1 room and television are representative. In comparison, SEL C- has more descriptors for labeling and with a greater number of hits. This may indicate that what characterizes the typical SEL C is more dispersed than what helps to define other strata.

Table 11 corresponds to the typical C, the middle class, have barely covered their needs. Here only the variables of electricity, 1 room and television are representative. In comparison, SEL C- (which is shown in Table 12) has more descriptors for labeling and with a greater number of hits. This may indicate that what characterizes the typical SEL C is more dispersed than what helps to define other strata.

SEL D+ groups the largest proportion of the population. The only attribute that represents them is to have a computer; the PC has 9589 elements with a 100 % analysis and a range of (-0.001, 0.25) with 0 errors and 100 % hits. There is no other descriptor that defines them. On the other hand, BSAs belonging to level D (Table 13)

TABLE 9. Results SEL AB

ATTRIBUTE	# ELEMENTS	ANALYSIS (%)	RANGES	ERRORS	HITS (%)
1 Room	3776	100	(-0.001, 0.25]	0	100
Sanitation		99.89	(0.75, 1.0]	4	99.89
Electricity		99.19	(0.75, 1.0]	12	99.68
Water		99	(0.75, 1.0]	14	99.64
TV		98.98	(0.75, 1.0]	20	99.46
Floors		98.87	(0.75, 1.0]	18	99.52
Refrigerator		97.79	(0.75, 1.0]	68	98.21
2 Rooms		87.01	(-0.001, 0.25]	473	87.48

Note. Created by the authors with data from Population and Housing Census 2010.

TABLE 10. Results SEL C+

ATTRIBUTE	# ELEMENTS	ANALYSIS (%)	RANGES	ERRORS	HITS (%)
Sanitation	8070	100	(0.75, 1.0]	0	100
TV		100	(0.75, 1.0]	0	100
Electricity		100	(0.75, 1.0]	0	100
Water		100	(0.75, 1.0]	0	100
1 Room		100	(-0.001, 0.25]	0	100
Floors		100	(0.75, 1.0]	0	100
Refrigerator		100	(0.75, 1.0]	5	99.94
2 rooms		99.11	(-0.001, 0.25]	71	99.12
3 rooms		90.67	(0.75, 1.0]	730	90.95
Washer		82.36	(0.75, 1.0]	1110	86.25

Note. Created by the authors with data from Population and Housing Census 2010.

have various attributes: electricity, television, computer, 1 room, type of floor, and level of schooling. The schooling attribute is unique to this SEL because it represents basic levels of schooling. Most have a secondary education (complete and incomplete) which is compulsory education mandated by the State.

SEL E (Table 14) is described by two attributes: electricity and television. It is possible to emphasize that, although they do not have basic satisfiers, such as sanitation and flooring, they do have with electricity and television, (even if they obtain it by means of illegal connections to the grid) and, thus, to information and entertainment.

Following Lopes, Machado, & Rabelo (2014), we consider the main attributes to define the label. It can be seen that the attributes that were 100 % relevant have no error. However, if only these attributes were considered, there would be the possibility of ambiguity between the labels, as occurs between clusters between C+ and C- clusters or between AB and C+ clusters. Because of this, it is necessary to observe whether the other attributes suggested within the V variation are sufficient to distinguish all the labels. However, the cost of avoiding ambiguity is dependent on less relevant attributes, as suggested by Lopes et al. (2016). In order to be able to differentiate C+ from C-, we can

TABLE 11. Results SEL C

ATTRIBUTE	# ELEMENTS	ANALYSIS (%)	RANGES	ERRORS	HITS (%)
Electricity	8361	93.72	(0.75, 1.0]	716	91.44
1 Room		85.78	(-0.001, 0.25]	1056	87.37
TV		85.34	(0.75, 1.0]	1214	85.48

Note. Created by the authors with data from Population and Housing Census 2010

TABLE 12. Results SEL C-

ATTRIBUTE	# ELEMENTS	ANALYSIS (%)	RANGES	ERRORS	HITS (%)
Sanitation	8979	100	(0.75, 1.0]	0	100
TV		100	(0.75, 1.0]	0	100
Electricity		100	(0.75, 1.0]	0	100
Floors		99.99	(0.75, 1.0]	0	100
1 Room		98.5	(-0.001, 0.25]	94	98.95
Water		97.53	(0.75, 1.0]	228	97.46
2 rooms		87.39	(-0.001, 0.25]	1028	88.55
Refrigerator		84.89	(0.75, 1.0]	1076	88.01

Note. Created by the authors with data from Population and Housing Census 2010

TABLE 13. Results SEL D

ATTRIBUTE	# ELEMENTS	ANALYSIS (%)	RANGES	ERRORS	HITS (%)
Electricity	9952	100	(0.75, 1.0]	0	100
TV		95.64	(0.75, 1.0]	460	95.37
PC		89.66	(-0.001, 0.25]	762	92.35
1 Room		88.75	(-0.001, 0.25]	956	90.39
Floors		89.74	(0.75, 1.0]	1092	89.03
Schooling		93.24	(0.23, 0.45]	1282	87.12

Note. Created by the authors with data from Population and Housing Census 2010

TABLE 14. Results SEL E

ATTRIBUTE	# ELEMENTS	ANALYSIS (%)	RANGES	ERRORS	HITS (%)
Electricity	3169	85.76	(0.75, 1.0]	494	84.42
TV		75.74	(0.75, 1.0]	655	79.34

Note. Created by the authors with data from Population and Housing Census 2010 suggest the following label for C+: sanitation, TV, electricity, water, floors, refrigerator, 3 rooms and washing machine, all with a label (0.75, 1.0), and the attributes for the SEL of C- would be considered sanitation, electricity, floors, water, refrigerator labeled (0.75, 1.00), while 1 room and 2 rooms (-0.001, 0.25). Even with this alternative, the success rate remains high. Refrigerator, which is the lowest attribute for C-, shows that only 1,076 estimated observations do not obey the label. In general, the result of classification of the attributes in all the SELs studied was very satisfactory, obtaining an average of 90.86 % hits for all labels.

Conclusions

In the present study, the BSAs of the Mexican Republic were classified into seven fully differentiated groups through socioeconomic characteristics. Congruence was observed with data previously reported by the AMAI, which is the reference institution to carry out this division. The advantage of the applied methodology is that by using public information sources it is possible to replicate the results at any level of disaggregation of the National Geostatistical Framework, which can be carried out at the local, municipal, state or even city block level. The difficulty would be to have the necessary information for the classification. However, the only cost to companies would be the time invested in finding such data.

During the unsupervised cluster ordering process, it was observed that the results behaved according to the literature, since it is documented, on the one hand, that the median is a better descriptor than the average when the data are scattered. In the correlation matrix, it was observed that there are pooled variables with different characteristics. Thus, we adopted the normal finite mixture model, which allows comparing such diverse information. According to this methodology, the seven mutually exclusive and collectively exhaustive clusters established in the AMAI methodology were generated. However, when performing the labeling, it was established that the variables considered relevant in each cluster are different among themselves. In addition, each relevant variable has a high degree of accuracy, which helps to accurately predict the cluster to which a BSA belongs. We have confidence in the results since, following authors like Lopes et al. (2016), the bulk of the data was used for the learning stage of the algorithm, resulting in a 90.86 % accuracy in the prognosis of all clusters.

Further research can be done to improve the degree of disaggregation, provided that information is available per block. Another area to explore is the application of the SEL by Metropolitan Zones, since at present information can only be found for the three most important cities in Mexico: The Metropolitan Zone of the Valley of Mexico, the Metropolitan Zone of Guadalajara and the Metropolitan Zone of Monterrey. SEL information in smaller cities is practically nonexistent, or else, access to information is collected by private companies.

References

Adnan, M., Longley, P. A., Singleton, A. D., & Brunsdon, C. (2010). Towards real-time geodemographics: Clustering algorithm performance for large multidimensional spatial databases. *Transactions in GIS*, *14*(3), 283–297. https://doi.org/10.1111/j.14679671.2010.01197.x

Aghdaie, M. H., Zolfani, S. H., & Zavadskas, E. K. (2013). A hybrid approach for market segmentation and market segment evaluation and selection: An integration of data mining and MADM. *Transformations in Business and Economics*, *12*(2 B).

Allenby, G., Fennell, G., Bemmaor, A., Bhargava, V., Dawley, J., Dickson, P., … Yang, S. (2002). Market Segmentation Research: Beyond within and across Group Differences. *Marketing Letters*, *13*(3), 233–243.

AMAI. (2015). *Actualización Regla AMAI de los Niveles Socioecónomicos 8x7*. México, D.F. Retrieved from http://amai.org/privado/niveles.php

Andrews, R. L., Brusco, M., Currim, I. S., & Davis, B. (2010). An empirical comparison of methods for Clustering problems: Are there benefits from having a statistical model? *Review of Marketing Science*, *8*(1).

Aparna, K., & Nair, M. K. (2015). Comprehensive study and analysis of partitional data Clustering techniques. *International Journal of Business Analytics (IJBAN)*, *2*(1), 23–38.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027–1035).

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. (2012). Scalable k-means++. *Proceedings of the VLDB Endowment*, *5*(7), 622–633.

Beane, T. P., & Ennis, D. M. (1987). Market Segmentation: A Review. *European*

Journal of Marketing, *21*(5), 20–42. https://doi.org/10.1108/EUM0000000004695

Berzofsky, M., Smiley-McDonald, H., Moore, A., & Krebs, C. (2014). Measuring Socioeconomic Status (SES) in the NCVS: Background, Options, and Recommendations. *Report*. Washington, DC: Bureau of Justice Statistics.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Reviews in Psychology*, *53*, 371–399. https://doi.org/10.1146/annurev.psych.53.100901.135233

Brochado, A. O., & Martins, F. V. (2015). Identifying Small Market Segments with Mixture Regression Models. *International Journal of Latest Trends in Finance and Economic Sciences*, *4*(4), 9.

Bukhari, S. S. (2011). Green Marketing and its impact on consumer behavior. *European Journal of Business and Management*, *3*(4), 375–384.

Capó, M., Pérez, A., & Lozano, J. A. (2017). An efficient approximation to the K-means Clustering for massive data. *Knowledge-Based Systems*, *117*, 56–69. https://doi.org/10.1016/j.knosys.2016.06.031

Cliquet, G. (2013). *Geomarketing: Methods and strategies in spatial marketing*. John Wiley & Sons.

de la Garza García, J. (1995). *Análisis de la información mercadológica: a través de la estadística multivariante*. Alhambra Mexicana.

Dickson, P. R., & Ginter, J. L. (1987). Market segmentation, product differentiation, and marketing strategy. *The Journal of Marketing*, 1–10.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2001). *Cluster analysis*. https://doi.org/10.1177/014662167800200315

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., … Bouras, A. (2014). A survey of Clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279.

Fisher, C., Bashyal, S., & Bachman, B. (2012). Demographic impacts on environmentally friendly purchase behaviors. *Journal of Targeting, Measurement and Analysis for Marketing*, *20*(3–4), 172–184.

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications. ASASIAM Series on Statistics and Applied Probability* (Vol. 20). https://doi.org/10.1111/j.1751-5823.2007.00039_2.x

George, M. R. W., Yang, N., Jaki, T., Feaster, D. J., Lamont, A. E., Wilson, D. K., & Van Horn, M. L. (2013). Finite mixtures for simultaneously modeling differential effects and nonnormal distributions. *Multivariate Behavioral Research*, *48*(6), 816–844.

Gottfried, A. W. (1985). Measures of socioeconomic status in child development research: Data and recommendations. *Merrill-Palmer Quarterly (1982-)*, 85–92.

Grekousis, G., & Thomas, H. (2012). Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The fuzzy C-means and Gustafson-Kessel methods. *Applied Geography*, *34*. https://doi.org/10.1016/j.apgeog.2011.11.004

Gutiérrez, B. (2016). *Antropología del consumidor tapatío*. Guadalajara, Jalisco, México.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis. *Vectors*. https://doi.org/10.1016/j.ijpharm.2011.02.019

Heath, J. (2012). *Lo que indican los indicadores: cómo utilizar la información estadística para entender la realidad económica de México*.

Hiziroglu, A. (2013). A neuro-fuzzy two-stage Clustering approach to customer segmentation. *Journal of Marketing Analytics*, *1*(4), 202–221.

Hollingshead, A. (1975). *Four Factor index of social status* (No. 208265). New Haven.

INEC (2013). *Determinación de los coeficientes de variación*. Quito, Ecuador.

INEGI (2002). *Regiones Socioecónomicas de México*. México, D.F.

Kim, T., & Lee, H.-Y. (2011). External validity of market segmentation methods: a study of buyers of prestige cosmetic brands. *European Journal of Marketing*, *45*(1/2), 153–169.

Kotler, P., & Armstrong, G. (2012). *Marketing*.

Krawczyk, B. (2016). Knowledge-Based Systems Dynamic classifier selection for one-class classification, *107*, 43–53. https://doi.org/10.1016/j.knosys.2016.05.054

Larsen, N. (2010). *Market Segmentation - A Framework for Determining the Right Target Customers*. *Aarhus School of Business*. Retrieved from http://pure.au.dk/portal/files/11462/ba.pdf

Levin, R. I., & Rubin, D. S. (2004). *Estadística para administración y economía*. Pearson Educación.

Lin, T. I., Lee, J. C., & Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 909–927.

Lizana, P. A., González, S., Lera, L., & Leyton, B. (2017). Association between body composition, somatotype and socioeconomic status in Chilean children and adolescents at different school levels. *Journal of Biosocial Science*, 1–17.

Lopes, L. A., Machado, V. P., Rabêlo, R. A. L., Fernandes, R. A. S., & Lima, B. V. A. (2016). Knowledge-Based Systems Automatic labelling of Clusters of discrete and continuous data with supervised machine learning. *Knowledge-Based Systems*, *106*, 231–241. https://doi.org/10.1016/j.knosys.2016.05.044

Lopes, L. A., Machado, V. P., & Rabelo, R. D. A. L. (2014). Automatic Cluster labeling through Artificial Neural Networks. *Proceedings of the International Joint Conference on Neural Networks*, 762–769. https://doi.org/10.1109/IJCNN.2014.6889949

Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

Mihić, M., & Čulina, G. (2006). Buying behavior and consumption: social class versus income. *Management*, *11*(2), 77–92.

Momeni, M., Yazdani, S., & Khorshidi, M. F. (2016). Clustering customers by C-mean method (Case study: Golestan company). *International Business Management*, *10*(8). https://doi.org/10.3923/ibm.2016.1406.1413

Müller, H., & Hamm, U. (2014). Stability of market segmentation with Cluster analysis – A methodological approach. *Food Quality and Preference*, *34*, 70–78. https://doi.org/10.1016/j.foodqual.2013.12.004

Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers & Industrial Engineering*, *109*, 233–252.

Musyoka, S. M., Mutyauvyu, S. M., Kiema, J. B. K., Karanja, F. N., & Siriba, D. N. (2007). Market segmentation using geographic information systems (GIS): A case study of the soft drink industry in Kenya. *Marketing Intelligence & Planning*, *25*(6), 632–642. https://doi.org/DOI: 10.1108/02634500710819987

Nosi, C., Pratesi, C. A., & D'agostino, A. (2014). A benefit segmentation of the Italian market for full electric vehicles. *Journal of Marketing Analytics*, *2*(2), 120–134.

O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P. D., & Karlis, D. (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics & Data Analysis*, *93*, 18–30.

Pan, W., Shen, X., & Liu, B. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research*, *14*(1), 1865–1889.

Pridmore, J., & Hämäläinen, L. E. (2017). Market Segmentation in (In) Action: Marketing and 'Yet to Be Installed' Role of Big and Social Media Data. *Historical Social Research/Historische Sozialforschung*, 103–122.

Ruiz, F. J., Angulo, C., & Agell, N. (2008). IDD: A supervised interval distance-based method for discretization. *IEEE Transactions on Knowledge and Data Engineering*, *20*(9), 1230–1238. https://doi.org/10.1109/TKDE.2008.66

Samarasinghe, S. (2016). *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. CRC Press.

Sánchez-hernández, G., Chiclana, F., Agell, N., & Carlos, J. (2013). Knowledge-Based Systems Ranking and selection of unsupervised learning marketing segmentation, *44*, 20–33. https://doi.org/10.1016/j.knosys.2013.01.012

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal*, *8*(1), 289.

Scrucca, L., & Raftery, A. E. (2015). Improved initialisation of model-based Clustering using Gaussian hierarchical partitions. *Advances in Data Analysis and Classification*, *9*(4), 447. Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies Published. *Journal of Marketing*, *21*(1), 3–8. Retrieved from www.jstor.org/stable/1247695

Suhaibah, A., Uznir, U., Rahman, A. A., Anton, F., Mioc, D., Estate, R., & Segmentation, M. (2016). 3d geomarketing segmentation: a higher spatial dimension planning perspective, *XLII*(October), 3–5. https://doi.org/10.5194/isprs-archivesXLII-4-W1-1-2016

Ultsch, A. (2002). Emergent self-organising feature maps used for prediction and prevention of churn in mobile phone markets. *Journal of Targeting, Measurement and Analysis for Marketing*, *10*(4), 314–324.

UNDP. (2017). Human Development Reports.

Vajda, S., Rangoni, Y., & Cecotti, H. (2015). Semi-automatic ground truth generation using unsupervised Clustering and limited manual labeling: Application to handwritten character recognition. *Pattern Recognition Letters*, *58*, 23–28.

Vera-Romero, O. E., & Vera-Romero, F. M. (2015). Evaluación del nivel socioeconómico: presentación de una escala adaptada en una población de Lambayeque. *Rev. Cuerpo Méd. HNAAA*, *6*(1), 41–45.

Wang, H., & Zaniolo, C. (2000). CMP: a fast decision tree classifier using multivariate predictions. *Proceedings of 16th International Conference on Data Engineering (Cat. No.00CB37073)*. https://doi.org/10.1109/ICDE.2000.839444

Wedel, M., & Kamakura, W. A. (2012). *Market Segmentation: Conceptual and Methodological Foundations*. Springer Science & Business Media.

Winston, W. (2014). *Marketing Analytics*. *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004