Contributions of Clustering Variable Selection: Methods for International Segmentation

Talibi, Abdelghafour

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Contributions of Clustering Variable Selection: Methods for International Segmentation

Abdelghafour TALIBI, *(PhD, Professor)*
Laboratory of Innovation in Management and Engineering for the Enterprise
Higher Institute of Engineering and Business, ISGA Groupe, Morocco

Correspondence address :	Institut Supérieur d'InGénierie et des Affaires (Higher Institute of Engineering and Business) Laboratoire d'Innovation en Management et en Ingénierie pour l'Entreprise 27, Avenue Oqba, 27 Av. Oqba Ibn Naafi, Rabat Maroc, 10090 Tel and fax of faculty or business school or author: +212 5 37 77 14 68
Disclosure Statement :	Author is not aware of any findings that might be perceived as affecting the objectivity of this study
Conflict of Interest :	The author report's no conflicts of interest.
Cite this article :	TALIBI, A. (2022). Contributions of Clustering Variable Selection: Methods for International Segmentation. International Journal of Accounting, Finance, Auditing, Management and Economics, 3(4-3), 498-530. https://doi.org/10.5281/zenodo.6969989
License	**This is an open access article under the CC BY-NC-ND license**

Received: June 15, 2022 *Published online: August 06, 2022*

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Contributions of Clustering Variable Selection: Methods for International Segmentation

Abstract:

Performing international activities is a challenging operation given the heterogeneity of the international market which makes practically impossible the development of successful standardized strategies for the entire world's population. Finding homogeneous international customer segments helps companies to better communicate with the targeted customers by concentrating on a few units, a group, or several groups. Depending on the study purpose, the segmentation results may help to select potentially attractive international markets, to develop in the context of a global marketing standardized strategies for a segment of countries, or to develop in the context of an international marketing a totally or partially differentiated strategy for several groups. Thus, international segmentation has become an indispensable task in the strategic decision-making process for various international business research questions. Consequently, choosing the relevant segmentation bases and the statistical method represent crucial steps to carry out to identify segments of customers. Actually, research studies in which an international market is segmented mainly employ as bases socio-economic or cultural variables. Moreover, in these studies, since the purpose of the analysis is usually to discover a priori unknown segments in an international population, the segmentation task is performed by clustering techniques. Typically, in this scientific research, to facilitate the interpretation of the results, the segmentation task is preceded by factor analysis to reduce a large number of the initial variables into a few dimensions or factors. However, on the one hand, factor analysis usually generates a loss of information and distortion of reality. On the other hand, the set of the variables initially considered may contain irrelevant variables that might lead to incorrect classification. Therefore, to retain only relevant information for the clustering task: variable selection should be performed to reduce the data dimension before considering a factor analysis.As shown by the numerical experiments,conducted on the basis of two secondary databases: the 03/07/2018 updates of the structure of consumption expenditure published by Eurostat including 32 countries of the European Union and its neighboring countries and the 15/04/2016 version of the updated European Values Study data including customers from 48 European countries, it will allow discovering the accurate groups and facilitate result interpretation. As a result, variable selection allows discovering relevant segments that are easy to interpret. Thus, once the variable selection is performed, the segmentation results will enable relevant and accurate analysis and support correct decision-making.

Keywords: Clustering; Variable selection; International segmentation;International marketing; Global marketing**.**
JEL Classification: C38
Paper type: Empirical research

1. Introduction

Engage in international activities in the context of global or international marketing has become an essential step for the company's survival and expansion.Indeed, entering new foreign markets, exporting the company's product, and expanding the company's activities across national borders, enables targeting consumers in different countries.So, this allows companies to increase the quantity sold, reduce production costs through economies of scale, manage risk, as well as to improve the quality of the product to reply to the needs of international consumers (Goodman (1983); Yip (1995); Steenkamp & Ter Hofstede (2002)).

Given the size of the world population and the international market heterogeneity, several research works have performed international market segmentation to find groups of similar countries or international customers. This concept is defined as the process of identifying specific segments, composed of countries or individual customers, which are composed of potential individuals with homogeneous attributes that are likely to present similar responses to the company's marketing composition (Hassan & Katsanis (1991), p.17).

International segmentation is necessary for decision-making and marketing strategy development at the international level (Helsen et al. (1993); Sethi (1971)). Indeed, it is a tool that allows companies, for instance, to select attractive markets (Cavusgil et al. (2004)) and to apply a global marketing program to a group of countries or international customers (Kramer & Herbig (1994)). It can also help to compromise between the standardization and differentiation of marketing mix strategies in the context of international marketing (Sethi (1971)), to duplicate success and transfer the experience previously got in a market to a similar market (Sethi and Holton (1973); Ye Sheng & Mullen (2011)), and to define the company's international positioning strategy (Brooksbank (1994)).

Consequently, several authors, such as Day et al. (1988), NachumNachum (1994), and Papadopoulos & Martín Martín (2011) have confirmed the usefulness and importance of international segmentation and have mentioned the importance of the variables chosen as its basis. Choosing the segmentation bases to assign consumers or countries to groups represents a significant factor for a successful international segmentation (Steenkamp & Ter Hofstede (2002)).

A rarely used approach consists in collecting information called primary data for domain-specific variables by conducting a data collection method, mainly a survey (Cleveland et al. (2011)). The other approach, widely used by marketing researchers, is to exploit secondary databases.The latter are composed of general variables that do not have a direct relationship to the problem at hand, mainly socio-economic and cultural variables such as those available in accessible sources (e.g. databases of the World Bank and the UN, the indices of Hofstede et al. (1990)) (Helsen et al. (1993); Cleveland et al. (2011)).

Thus, a crucial step in the international market segmentation process, after the problem is well defined and the segmentation objectives are named, is to identify the necessary information and determine the relevant variables to be used as bases for the segmentation task.

Typically, research studies in which international segmentation has been conducted to obtain groups of homogeneous countries or international customers use clustering techniques based on socio-economic, cultural, or psycho-graphical variables. In these works, to help interpret the groups obtained (e.g. Day et al. (1988); Sriram & Gopalakrishna (1991); Peterson & Malhotra (2000); Steenkamp (2001); Cavusgil et al. (2004); Dubois et al. (2005); Budeva & Mullen (2014)), factor analysis techniques have been frequently performed before the clustering task to reduce the initial number of variables into a small number of factors without proceeding by variable selection to keep only the relevant variables. Nevertheless, dimension reduction by a factor analysis usually generates a loss of information and a distortion of reality.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Other authors (e.g. Law et al. (2004); Raftery & Dean (2006); Wang & Zhu (2008); Xie et al. (2008); Maugis et al. (2009); Meynet & Maugis-Rabusseau (2012); Sun et al. (2012); Silvestre et al. (2015); Arias-Castro & Pu (2017)) considers that clustering methods should include variable selection. Their proposed methods are based on the hypothesis that only a subset of the available variables might contain the relevant information for the correct classification. Eliminating irrelevant variables allows, according to these authors, to improve the clustering results and facilitates the interpretation and analysis of the results that will be conducted based on a few variables.

To improve the results of an international market segmentation performed by clustering methods. Firstly, for an initial exploration of a given research topic, for which the researcher ignores the potentially relevant variables, we propose to use variable selection methods to consider only the set of relevant variables from high-dimensional databases. Secondly, we consider that it is indispensable, as done in the previous works, to use several variables potentially relevant to studying a given theme or problem. Yet, to apply the variable selection methods to select the set of relevant variables from the initially considered variables.

Thus, we claim that variable selection for clustering should be conducted to perform an international segmentation to explore a research topic for which the researcher downright ignores a posteriori relevant variables, or, to select the relevant ones from a set of potentially relevant variables.Mainly, our objective is to apply variable selection methods for an international segmentation to investigate their advantages in terms of clustering results and the degree of interpretability ease.

This article is organized as follows. First, Section 2 presents a selective literature review of works in which international segmentation was conducted. Second, Section 3 presents our research methodology. Third, in Section 3, we applied clustering techniques and variable selection methods on quantitative and qualitative data. Finally, a conclusion is presented in section 5.

2. Literature Review

Several researches work in international or global marketing have used statistical techniques, particularly clustering methods, for the international segmentation task. Predominantly, the segmentation bases used were general variables, mainly socio-economic and cultural variables. Nevertheless, some works have used a domain-specific variables In the current section, we will present a selective literature review of studies in which international segmentation was performed.

In one of the first international marketing works, Sethi (1971) has explored the opportunities of using clustering methods for international marketing purposes. Sethi (1971) has performed two analyzes on a database containing 91 countries described by 29 socio-economic variables. First, V-analysis (Tryon and Bailey (1966)) was performed to obtain groups of variables. Secondly, based on the V-analysis results, an O-analysis (Tryon & Bailey (1966)) was executed to form groups of countries. The author concludes that countries should be classified according to several variables to identify relevant international marketing opportunities.

Based on the assumption that segmentation is useful for national and international marketing, Day et al. (1988) examined the limitations and advantages of country clustering to identify standardization opportunities in industrial marketing. The authors have selected a final sample composed of 96 countries, described by 18 economic variables, on which a factor analysis was performed and has resulted in retaining tree factors uncorrelated with two variables. These two variables were eliminated from the second factor analysis conducted on the other 16 variables, resulting in the same factors being retained as in the first analysis. The clustering task was

carried out using FASTCLUS, a SAS statistical software procedure (Fernandez (2010)), based on the dimensions obtained from the second factor analysis combined with the two variables removed from the analysis and secondly based on the dimensions obtained from the first factor analysis. Both clustering analyses resulted in 6 clusters being chosen as the optimal solution. The authors point out that similarities between countries need to be explored to identify standardization opportunities and that marketers looking for global marketing strategies should select the economic variables relevant to the product or service in question.

In the work of Lee (1990), based on a cross-sectional study measuring innovation, the author uses the ownership of white and black televisions and color televisions per thousand people in 1981 to measure the degree of adoption of a new product class for 70 countries. The main objective was to examine the relevant factors and determinants of innovation based on 10 socio-economic variables and to divide countries into groups with different levels of innovation. First, the author conducted correlation analyses and stepwise regressions to analyze the determinants of innovation based on the a priori selected variables. Then a hierarchical classification of the countries into 5 clusters was made. The author has claimed that the results obtained are potentially useful for international marketers to develop global marketing strategies for new products and can be used by international marketers to target each group with specific communication methods or messages. Therefore, he concluded that the cultural variables that influence the level of innovation in a country should be combined with economic variables to achieve a more relevant segmentation.

To explore the possibilities of standardization for international marketing strategies, Sriram & Gopalakrishna (1991) classified 40 countries described by 9 economic variables, 4 cultural dimensions, and 7 media-related variables. First, a factor analysis was conducted based on the standardized values of the variables to reduce the 20 variables to a small number of factors. Secondly, the country scores for these factors were used as the basis for classifying the countries through hierarchical classification, resulting in the selection of 6 clusters as the best solution. Finally, the authors analyzed the stability of the clustering results and the relevant variables from the variables initially considered using discriminant analysis. The authors concluded that cultural and media-related variables should be used simultaneously as a basis for the segmentation task to develop appropriate standardized advertising strategies.

Kale (1995) classified 17 Western European countries based on Hofstede's 4 cultural dimensions (see Hofstede (1980)) to identify strategic marketing opportunities for European marketers. As in several other papers, the author first performed a hierarchical classification using SAS software to determine the number of groups. In a second step, a non-hierarchical method with 3 groups as input was used to classify the countries. One consequence of the results obtained by the author is the possibility of determining a specific promotion for each group, which can be carried out differently for the countries of the same group.

In the work of Zandpour & Harich (1996), the authors attempted to estimate the best advertising appeals for each country using cultural and market-related variables for 23 countries. Based on a sample of 1914 TV commercials for different categories of randomly selected products from 8 countries, in which each TV commercial was described by the type of appeal and the advertising information, several regression analyses were conducted considering the indices of types of appeal and type of advertising information as dependent variables and the cultural and market-related variables as independent variables. Based on the results of the regression analysis, the 23 countries were analyzed in terms of the predominant type of appeal and relevant advertising information. The authors concluded that marketers should use cultural, market-related, and media-related variables instead of geographical variables to find ways to standardize advertising.

The aim of Peterson & Malhotra's (2000) work was to present an international segmentation based on the quality of life variables (The IL QoL survey). Their study used 6 quality of life

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

variables for 65 countries, with data measured over three years (1990-1992). The authors first analyzed for each variable the correlation between their values for two different years. Second, an exploratory factor analysis was conducted using the maximum likelihood technique, which resulted in 2 factors being retained for each year's data. Third, a confirmatory factor analysis was conducted using structural equations, which confirmed the existence of these 2 factors. Finally, a clustering analysis was conducted based on the standardized values of the six variables using a hierarchical and nonhierarchical technique which resulted in the selection of 12 clusters as the optimal outcome. The authors concluded that the clustering results can be used by marketers to gain strategic advantages in terms of promotional strategy and that researchers can use the "IL QoL survey" data as a reference for research in international trade.

Steenkamp (2001) has examined the interrelationships between the two main cultural frameworks: Hofstede's framework (Hofstede (1984); Hofstede et al. (1991)) and Schwartz's framework (Schwartz (1994, 1997)). The author applied a factorial analysis with principal component analysis to a database containing Hofstede's 4 cultural dimensions and Schwartz's 7 cultural domains for 24 countries included in both databases (Hofstede et al. (1991); Schwartz (1994)). The factorial analysis has resulted in maintaining a unified cultural framework composed of 4 factors. The author has also explored the usefulness of segmentation for international marketing. Based on the values for the 4 factors of the unified cultural framework, the 24 countries were classified into 7 groups through two-stage cluster analysis.

The aim of the work by Gupta et al. (2002) was to classify 61 countries participating in the GLOBE survey based on variables already considered relevant in previous work, such as language, geography, religion, ethnicity, values, and professional attitudes. The authors proposed a classification into 10 groups and tested the validity of the predefined groups using a linear discriminant function based on 9 social attitudes scales "AS IS" and 9 social values scales "Should be.

The aim of the work by Cavusgil et al. (2004) was to analyze the role and importance of clustering and country ranking techniques for market selection, overcome the limitations of previous studies, and apply these techniques to the latest available data. The data used to correspond to 90 countries described by 29 variables, with four variables not used in previous work. To perform the clustering task, the authors first conducted an exploratory factor analysis using principal component analysis and selected five factors that summarized much of the data variance. Second, based on these results, a hierarchical classification was performed to determine the number of clusters. Third, a 10 clusters solution was selected and used as input to the K-Means algorithm. The authors have found that clustering and country ranking techniques allow marketers to evaluate international market opportunities. However, they conclude that the clustering technique, unlike country ranking, finds groups of similar countries that can help marketers determine relevant strategies for a given country.

In the context of the international financial market, the aim of the work of Bijmolt et al. (2004) was to combine country and consumer segmentation in an international segmentation using a Multi-Level Latent Class Model (Vermunt (2003)). The data used to correspond to the Eurobarometer 56:0 (Christensen (2002)), which measures the ownership of 8 financial products for 15 countries in the European Union, with each country represented by about 1000 individuals. The authors conclude that international segmentation is an important tool for formulating relevant international policies if some difficulties can be overcome.

Dubois et al. (2005) first conducted an exploratory study to analyze consumers' experiences with luxury, which led to the development of a set of 33 items measuring consumer attitudes and serving as the basis for segmentation. Second, the authors collected data from 1848 subjects corresponding to a sample of management students from 20 countries across the four continents. Third, the authors conducted a factor analysis, which revealed that the 33 items could not be reduced to a small number of factors. Finally, a mixture model was applied to the data, resulting

in the selection of 3 classes as the best outcome. The authors claimed that besides cultural variables, psychological variables could also influence customers' attitudes towards luxury.
The aim of Budeva & Mullen's (2014) work was the first to analyze the difference between segmentation based on economic variables, based on cultural variables, and based on both types of variables used simultaneously. Secondly, to test the stability of the clustering result over time. The original sample consisted of 34 countries covered by the 1990-1991 and 1999-2001 World Value Survey. These countries were described by Inglehart & Baker's two cultural dimensions (Inglehart & Baker (2000)): 'traditional versus secular-rational orientation' and 'survival versus self-expression', and 12 economic variables selected based on a literature review to which principal component analysis was applied for the two periods to reduce the original number of economic variables to three factors. Based on the country scores for the three economic factors and the two cultural dimensions, several classifications were made for the two periods studied. These clustering analyses were carried out using a two-step method, i.e. a hierarchical method to select the number of groups, followed by a non-hierarchical method, the K-means algorithm, first based on the cultural dimensions, then based on the economic factors, and finally based on both types.
The objective of the study of Hernani-Merino et al. (2020) was to develop a better understanding of global customer culture with regard to standardize or adapt a global brand strategy for a specific international customer segment that shares the same desires and preferences. The data considered was collected through a non-probabilistic online survey in the United States, Brazil, Peru, France and the Czech Republic.The final simple is composed of 412 participants, 77 of which were from the United States, 122 from Brazil, 78 from Peru, 121 from France and 14 from the Czech Republic. These participants were asked questions about variables from the measurable theoretical model proposed by Hernani-Merino et al. (2015), which has seven dimensions: conformity to consumption trends, quality perception, social prestige, social responsibility, brand credibility, perceived risk and information costs saved. To find segments composed of customers sharing the same cultural characteristics, the authors have used a probabilistic clustering method in which each individual have a probability to belong to every group: The fuzzy C-Means. The method chosen two groups as the best solution but to differentiate more the characteristics of customers belonging to different groups the authors proposed a classification into 3 groups. The authors concluded that customers from different countries have common beliefs about the social responsibility of global brands, confirming the existence of the fragmentation of the needs of customers within and between countries.
Table A.1 (Appendix) gives a summary of these works.

3. Research method or methodology

3.1. Research design

To investigate the contributions of variable selection for clustering we will compare, in terms of clustering results and degree of ease of interpretation, standard clustering methods and clustering methods where the clustering process involves variable selection. First and foremost, clustering methods are used for quantitative data. Secondly, clustering methods are used for qualitative data to perform international segmentation.

3.2. Data Description

For the application of clustering techniques on numerical variables, we will use secondary data published by Eurostat corresponding to the 03/07/2018 updates of the structure of consumption expenditure according to the 2010 consumption function. The database considered includes 53 variables measured for 32 countries of the European Union and its neighboring countries.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

We have cleaned the data by eliminating the values for the Netherlands with many missing values and the values for Germany (until 1990, former territory of the FRG) characterized by the presence of several zero values for several variables. The final number of countries considered is 30, and the number of variables is 53. To eliminate the effect of population size, we have chosen the "purchasing power standard" (PPS) per household as the unit. Table A.2 describes the variables in the database.
Clustering techniques based on qualitative data are applied to the 15/04/2016 version of the updated European Values Study data (EVS, 2016). The observations from the EVS (2016) are older people aged 18 years and older for all countries considered (see Table A.5) except Armenia (15 years and older) and Finland (between 18 and 74 years). These observations were selected by stratified random sampling with a net sample size of 1500 per country, except for Northern Cyprus and Northern Ireland (500 observations each), Iceland (808), the Republic of Cyprus (1000), Ireland (1013), the Kingdom of Norway (1090), Finland (1134), Sweden (1187), Switzerland (1272), France (random sample: 1501, two additional quota samples: 1570) and Germany (disproportionate sample, East: 1004, West: 1071).
Data collection was conducted on the basis of the uniform instructions prepared by the EVS Advisory Groups, and by the administration of a questionnaire conducted with a face-to-face interview in the appropriate national language. From the question 52 of the family life and marriage items of this questionnaire, we have extracted data corresponding to ten variables (v170 - v180) as presented in the table A.6.

3.3.Analysis

For quantitative data we will use as standard clustering methods, the standard Gaussian mixture models and the K-Means algorithm. While as variable selection clustering methods, we will use the SRUW method of Maugis et al. (2009) and the regularized K-Means algorithm (Talibi et al., 2017a).
For qualitative data we will use as a standard clustering method, the traditional latent class model. While as variable selection clustering method, we will use the penalized latent class model (Talibi et al., 2017b).
Our numerical experiments will be performed using R software (R Core Team (2020)). In particular, for standard methods, we will use the R Stats package (R Core Team (2020)) to run the K-Means algorithm, the Rmixmod package (Langrognet et al. (2019)) to run Gaussian mixture models, and the R package poLCA (Linzer & Lewis (2011)) to run the traditional latent class model. For clustering methods including variable selection, the R package SelvarMix (Sedki et al. (2014)) is used to run the SRUW method of Maugis et al. (2009), while the regularized K-Means algorithm (Talibi et al., 2017a) and the penalized latent class model (Talibi et al., 2017b) are performed using R scripts with functions provided by the authors.

3.3.1. Mixture models

Mixture models (Wolfe (1963), Mclachlan & Basford (1988)) have attracted much attention recently because they provide an intuitive notion of a population consisting of multiple groups and are flexible for modeling a variety of phenomena.
Gaussian mixture models are model-based clustering models for multivariate numerical data. The idea is that each group is represented by a multivariate Gaussian distribution, since each observationx_i, $\mathrm{i} = \{1, \ldots, \mathrm{n}\}$, is a vector $(\mathrm{x_{i1}}, \ldots, \mathrm{x_{ij}})$ with $\mathrm{x_{ij}}$, the value of the variable for the observation, with different parameters of the distributions of other groups, while the entire population is represented by a mixture of these Gaussian distributions.
The general form of the likelihood of a Gaussian mixture model with K components for a single observation $\mathrm{x_i}$ is as follows:

$$\mathbf{L}(\mathbf{x_i}, \boldsymbol{\theta}) = \sum_{\mathbf{k=1}}^{\mathbf{K}} \boldsymbol{\pi_k} \mathbf{f_k}(\mathbf{x_i}, \boldsymbol{\theta_k}) \quad (1)$$

With,
$\pi_1, \ldots, \pi_K$: the mixture proportions,
f_k: the Gaussian distribution of the component k,

$$\mathbf{f_k}(\mathbf{x_i}, \boldsymbol{\theta_k}) = \frac{\mathbf{1}}{(\mathbf{2\pi})^{\mathbf{J/2}} |\mathbf{V_k}|^{\mathbf{1/2}}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2}} (\mathbf{x_i} - \mathbf{m_k})^{\mathbf{t}} \mathbf{V_k^{-1}} (\mathbf{x_i} - \mathbf{m_k})\right) \quad (2)$$

$\theta_k = \{m_k, V_k\}$: the parameters of f_k the Gaussian distribution of the component k,
m_k: the means vector of the component k,
V_k: the covariance matrix of the component k.
And the general form of the likelihood of a Gaussian mixture model with K components for n observations x_i is as follows:

$$\mathbf{L}(\mathbf{x}, \boldsymbol{\theta}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \left[\sum_{\mathbf{k=1}}^{\mathbf{K}} \boldsymbol{\pi_k} \mathbf{f_k}(\mathbf{x_i}, \boldsymbol{\theta_k}) \right] (3)$$

The parameter vector is then $\theta = (\pi_1, \ldots, \pi_K, m_1, \ldots, m_K, V_1, \ldots, V_K)$.

3.3.2. SRUW Method

Maugis et al. (2009) proposed a variable selection method formulated as a model selection problem for Gaussian mixture models, where they consider a parsimonious models based on a decomposition of the covariance matrix proposed by Fraley & Raftery (1998) and Celeux & Govaret (1995):

$$\mathbf{V_k} = \boldsymbol{\lambda_k} \mathbf{D_k} \mathbf{A_k} \mathbf{D_k^t} \quad (4)$$

Where λ_k is the largest eigenvalue of V_k which controls the volume of the k^{th} cluster, D_k is the eigenvectors matrix of V_k, which control the orientation of that cluster and A_k is a diagonal matrix with the scaled eigenvalues as entries, which control the shape of that cluster. By imposing constraints on the various elements of this decomposition, a large range of models is available, ranging from the simple spherical models that have fixed shape to the least parsimonious model where all elements of the decomposition are allowed to vary across the clusters.
Maugis et al. (2009) consider firstly the subset S which represents the relevant variables, and which includes a subset R of the relevant variables related to a subset of irrelevant variables, and secondly, Sc the complement of the subset S which is divided into two subsets: a subset U of irrelevant variables that can be explained by linear regression to the subset R and subset W of irrelevant variables that is completely independent of all relevant variables. The proposed method called SRUWattemptsto identify the subsets F = (S; R; U; W). The selected model maximizes the following quantity:

$$\left(\widehat{\mathbf{K}}, \widehat{\mathbf{m}}, \widehat{\mathbf{r}}, \widehat{\mathbf{l}}, \widehat{\mathbf{F}}\right) = \arg \max_{\mathbf{K,m,r,l,F}} \left\{ \mathbf{BIC_{Clustering}}(\mathbf{x^S}|\mathbf{K}, \mathbf{m}) + \mathbf{BIC_{Regression}}(\mathbf{x^U}|\mathbf{r}, \mathbf{x^R}) + \mathbf{BIC_{Ind}}(\mathbf{x^W}|\mathbf{l}) \right\} \quad (5)$$

The quantity (5) includes three terms. The first one corresponds to model-based clustering by a Gaussian mixture model with K components on the subset S and m its shape chosen from a collection of 28 parsimonious models available in Mixmod software (Biernacki et al. (2006)).

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

The second term represents a BIC approximation of the linear regression of the subset U of irrelevant variables to the subset R of relevant variables, where r is the form of the covariance matrix of the regression assumed to be spherical, diagonal or unconstrained. The last term corresponds to the BIC of a Gaussian distribution of the subset of the irrelevant variables W that are assumed to be independent of all the relevant variables where l is the shape of its variance matrix assumed to be diagonal or spherical.

3.3.3. K-Means

The k-means algorithm as all methods of clustering aims to classify the observations x_i, $i = \{1, \dots, n\}$ representing a population or a sample composed of n observations with all observations x_i are measured on J variables $(x_{i1}, \dots, x_{iJ})$ to K groups$G_1, \dots, G_K$ by minimizing the within-cluster sum of squares (WCSS) which is the distance between the observations belonging to the same cluster. The mathematical formulation of the algorithm is as follows:

$$\min_{G} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{x_i, x_{i'} \in G_k} \sum_{j=1}^{J} d_j(x_i, x_{i'}) \quad (6)$$

Where n_k is the number of observations in the cluster k, $d_j(x_i, x_{i'})$ is a dissimilarity measure based on the variable j between the observation x_i and the observation x_i, which can be expressed by:

$$d_j(x_i, x_{i'}) = \|x_{ij} - x_{i'j}\|^2 \quad (7)$$

Where $\|.\|$ is the standard Euclidean norm.
Then (6) is equivalent to:

$$\min_{G, m_k} \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{J} \|x_{ij} - m_{kj}\|^2 \quad (8)$$

Where $m_k = (m_{k1}, \dots, m_{kJ})^t$ the means vector of the cluster k and m_{kj} is the mean of the variable j in the cluster k.
The k-means algorithm starts with random centers and minimizes the function (8) by iterating between two steps. It classifies the observations into groups by storing each observation in the group with the closest center, and then computes new values for the centers given the last classification obtained.

3.3.4. Adaptive L∞-norm Regularized K-Means clustering

In this model Talibi et al. (2017a) applied the L∞-norm penalty (Wang & Zhu (2008)) on the means of the variables in the clusters to the K-Means algorithm to perform the clustering and to select the relevant variables.
Talibi et al. (2017a) proposed a modified version of K-Means in which the L∞-norm (Wang & Zhu (2008)) was applied to the means of variables within clusters to perform clustering and select the relevant variables.
The ALR-K-means algorithm is formulated as follows:

$$\min_{G, m_k} \frac{1}{2} \sum_{k=1}^{K} \sum_{x_i \in G_k} \sum_{j=1}^{J} \|x_{ij} - m_{kj}\|^2 + \lambda \sum_{j=1}^{J} w_j \max_{k \in \{1, \dots, K\}} (|m_{kj}|) \quad (9)$$

The first term of the quantity (9) to be minimized corresponds to the WCSS. The second term is the penalty function, formulated as an adaptive L∞-norm penalty applied to the means of the

variables within clusters. The penalty function applied to the centered data forces the irrelevant variables to have a means of zero within the clusters. With a tuning parameter λcontrolling the desired degree of sparsity andw_j The weights of the variable j, the adaptive L∞-norm penalty takes into account the relative importance of each variable (Zou, H. (2006)), so that the informative variables are easily regularized in contrast to the non-informative variables.
As with K-means, Talibi et al. (2017a) use an iterative approach to minimize (9). First, their algorithm initializes the values with standard K-means. Then, their algorithm iterates between two steps. (9) is minimized with respect to the clustering assignment$G = \{G_1, \dots, G_K\}$ by assigning each observation to the closest cluster; then the values of m_{kj} are computed using the last clustering result by minimizing (9).

3.3.5. Latent Class Model

The latent class model (Clogg (1995)) which can be formalized by two different and completely equivalent parameterizations; probabilistic and log-linear, was initially introduced by Lazarsfeld & Henry (1968), based on the idea that the dependence between categorical variables is in fact the result of a latent variable, which its modalities represent the classes in clustering.
The traditional latent class model is a model based clustering for multivariate categorical data, for which the classes have a multinomial distribution and the variables are independent given the knowledge of the class label.
The general form of the likelihood of a traditional latent class model with K components for n observationsx_i measured on J categorical variables can be formulated as follows:

$$\mathbf{L}(\mathbf{x}, \boldsymbol{\theta}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \left[\sum_{\mathbf{k=1}}^{\mathbf{K}} \boldsymbol{\pi}_{\mathbf{k}} \mathbf{f}_{\mathbf{k}}(\mathbf{x}_{\mathbf{i}}, \boldsymbol{\theta}_{\mathbf{k}}) \right] \quad (10)$$

$$\mathbf{f}_{\mathbf{k}}(\mathbf{x}_{\mathbf{i}}, \boldsymbol{\theta}_{\mathbf{k}}) = \prod_{\mathbf{j=1}}^{\mathbf{J}} \prod_{\mathbf{m_j=1}}^{\mathbf{d_j}} \mathbf{p}_{\mathbf{jm_jk}}^{\mathbf{1\{x_i=m_j\}}} \quad (11)$$

Where d_j is the number of categories that the variable j can takes, $1\{x_i = m_j\}$ is an indicator function that equals 1 if the variable j take the modality m_j as value and 0 otherwise and p_{jm_jk} is the probability that the variable j take the value m_j in the class k.

3.3.6. Variable Selection for Latent Class Model

Inspired and with the same motivation as in penalized model-based clustering approach, and by using the relationship between latent class model and the log-linear model (Goodman (1974), Haberman (1979)), Talibi et al. (2017b) proposed a penalized latent class model approach that selects the relevant variables and perform clustering by penalizing the log-likelihood function of the log-linear model to be minimized.
In fact, the conditional probabilities of the latent class model can be formulated by log-linear model parameters for the complete data, which include as interactions only those between the latent variable and each of the indicator variables.
In the case of four indicator variables ($J = 4$), the log-linear model for the expected cell counts $N_{m_1,m_2,m_3,m_4,k}$ of the complete data that include the classes label values can be expressed as:

$$\log\left(\mathbf{N}_{\mathbf{m_1,m_2,m_3,m_4,k}}\right) = \boldsymbol{\lambda} + \boldsymbol{\lambda}_{\mathbf{k}}^{\mathbf{LC}} + \boldsymbol{\lambda}_{\mathbf{m_1}}^{\mathbf{1}} + \boldsymbol{\lambda}_{\mathbf{m_2}}^{\mathbf{2}} + \boldsymbol{\lambda}_{\mathbf{m_3}}^{\mathbf{3}} + \boldsymbol{\lambda}_{\mathbf{m_4}}^{\mathbf{4}} + \boldsymbol{\lambda}_{\mathbf{m_1,k}}^{\mathbf{1,LC}} + \boldsymbol{\lambda}_{\mathbf{m_2,k}}^{\mathbf{2,LC}} + \boldsymbol{\lambda}_{\mathbf{m_3,k}}^{\mathbf{3,LC}} + \boldsymbol{\lambda}_{\mathbf{m_4,k}}^{\mathbf{4,LC}} \quad (12)$$

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

As expressed in (12), the log-linear model includes the first order effect of the indicator variables, the first order effect of the latent variable, and the interaction parameters between the latent variable and each one of the indicator variables.
With restrictions,

$$\sum_{k=1}^{K}\lambda_{k}^{LC}=\sum_{m_1=1}^{d_1}\lambda_{m_1}^{1}=\sum_{m_2=1}^{d_2}\lambda_{m_2}^{2}=\sum_{m_3=1}^{d_3}\lambda_{m_3}^{3}=\sum_{m_4=1}^{d_4}\lambda_{m_4}^{4}=\sum_{m_1=1}^{d_1}\lambda_{m_1,k}^{1,LC}=\sum_{m_2=1}^{d_2}\lambda_{m_2,k}^{2,LC}$$
$$=\sum_{m_3=1}^{d_3}\lambda_{m_3,k}^{3,LC}=\sum_{m_4=1}^{d_4}\lambda_{m_4,k}^{4,LC}=\sum_{k=1}^{K}\lambda_{m_1,k}^{1,LC}=\sum_{k=1}^{K}\lambda_{m_2,k}^{2,LC}=\sum_{k=1}^{K}\lambda_{m_3,k}^{3,LC}=\sum_{k=1}^{K}\lambda_{m_4,k}^{4,LC}$$
$$=\mathbf{0}\ (13)$$

A general formulation of the log-linear model can be expressed as follows:

$$\mathbf{log}(\mathbf{N})=\mathbf{X\Lambda}\ (14)$$

Where N is a vector of expected cell counts, X a design matrix composed by 0s and 1s, depending on the parameters included in the calculation of each one of the expected cell counts, and Λ is the vector of the unknown log-linear parameters.
The relationship between the parameters of the latent class model and the log-linear parameters is formulated as follows when calculating the conditional probabilities:

$$\mathbf{p}_{\mathbf{jm_jk}}=\frac{\exp\left(\lambda_{m_j}^{j}+\lambda_{m_j,k}^{j,LC}\right)}{\sum_{m_j=1}^{d_j}\exp\left(\lambda_{m_j}^{j}+\lambda_{m_j,k}^{j,LC}\right)}\ (15)$$

In the latent class model, a variable*j* is considered irrelevant if its distribution is the same in all classes. In log-linear parameterization, the total interaction parameters between the latent variable and an irrelevant variable*j* are all equal to 0, so that the variable has the same distribution in all classes.
To enforce that the irrelevant variables have the same distribution in all classes, Talibi et al. (2017b) proposed a penalized function that includes a penalty function for the interaction parameters. The penalized function, which must be minimized to estimate the log-linear parameters in the case of four explanatory variables$J = 4$, has the following form:

$$-\sum_{m_1,m_2,m_3,m_4,k}n_{m_1,m_2,m_3,m_4,k}\times\log N_{m_1,m_2,m_3,m_4,k}+\sum_{m_1,m_2,m_3,m_4,k}N_{m_1,m_2,m_3,m_4,k}$$
$$+P_w\left(\lambda_{m_j,k}^{j,LC}\right)\ (16)$$

Where $n_{m_1,m_2,m_3,m_4,k}$ is the observed cell count and P_w is a penalty function on the log-linear interaction parameters which have the following form:

$$P_w\left(\lambda_{m_j,k}^{j,LC}\right)=W\sum_{j=1}^{J}w_j\left\|\lambda_{m_j,k}^{j,LC}\right\|_2\ (17)$$

Where W is an hyper parameter which controls the level of the desired sparsity, w_j is the weight of the variable j estimated by the overall average variance of categories probabilities across the classes and $\|.\|_2$ the l2-penalty with $\left\|\lambda_{m_j,k}^{j,LC}\right\|_2^2=\sum_{m_j}\sum_{k}\left(\lambda_{m_j,k}^{j,LC}\right)^2$. Thus, a small value of the interactions parameters automatically will be regularized to be equal to 0, and if the overall

interaction parameters between the latent variable and a variable j are all equal, $\lambda_{m_j,1}^{j,LC} = \lambda_{m_j,2}^{j,LC} = \cdots = \lambda_{m_j,K}^{j,LC} = 0$ for all $m_j = 1, \ldots, d_j$, its distribution will be the same across the clusters and will be considered as irrelevant.

4. Results and discussion

4.1. Results

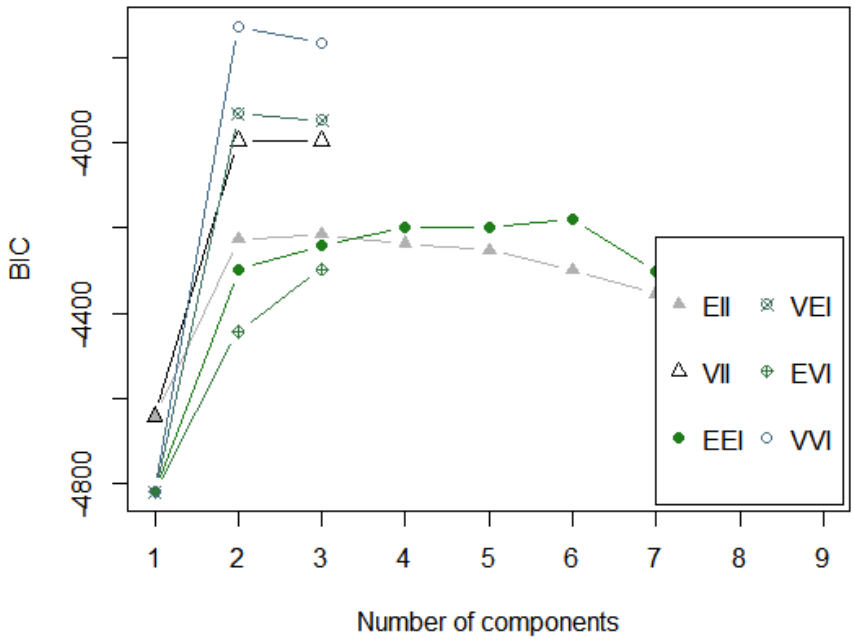
4.1.1. Clustering application on quantitative data

For the application of clustering techniques on numerical variables, we will use secondary data published by Eurostat corresponding to the 03/07/2018 updates of the structure of consumption expenditure according to the 2010 consumption function.

4.1.1.1.Gaussian mixture models and SRUW method

First, we applied a standard Gaussian mixture model to the standardized data, resulting in the choice of a model with 2 classes (BIC = -3727.824; ICL = -3727.824) with a diagonal variance matrix where each cluster has a different volume and shape (VVI) (see Figure 1).

Figure 1: Selection of the number of classes for the general mixture model



Source: Author

The clustering result of applying the general mixture model is shown in Table 1.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Table 1: The clustering result of the application of the general mixture model

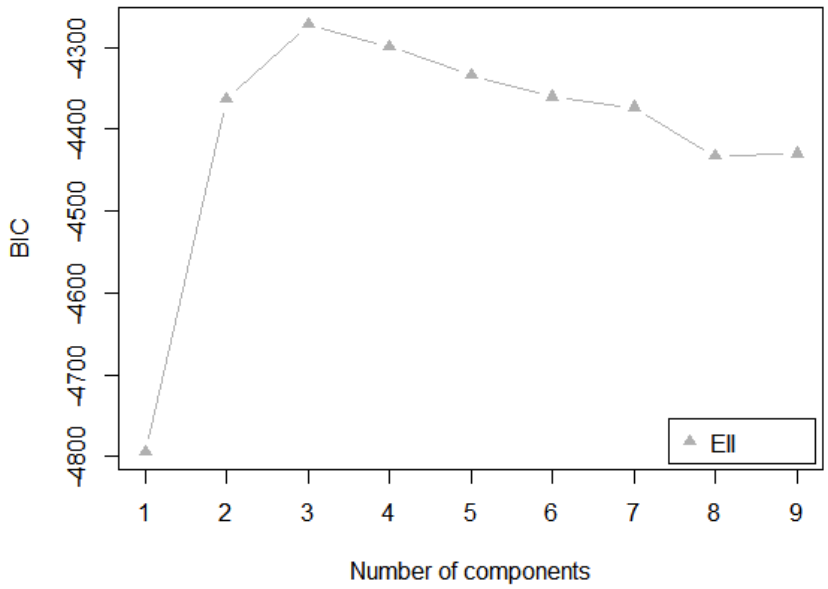
Group	
1 (ps = 0.6)	**2 (ps = 0.4)**
Belgium	Bulgaria
Denmark	Czech Republic
Ireland	Estonia
Greece	Croatia
Spain	Latvia
France	Lithuania
Italy	Hungary
Cyprus	Poland
Luxembourg	Romania
Malta	Slovakia
Austria	Montenegro
Portugal	Old Republic Yugoslav of Macedonia
Slovenia	
Finland	
Sweden	
UK	
Norway	
Turkey	

ps: population share

Source: Author

`We also forced the model to have a common spherical variance between all clusters (EII) corresponding to the K-Means algorithm for which the number of classes is selected by an information criterion. The model selected based on the information criteria BIC and ICL (BIC = -4216.93; ICL = -4216.936) consists of 3 classes (see Figure 2).

Figure 2: Selection of the number of classes for the spherical mixture model



Source: Author

The clustering result of applying the spherical mixture model is shown in Table 2.

Table 2: The clustering result of the application of the spherical mixture model

Group		
1 (ps = 0.4333314)	**2 (ps = 0.4334277)**	**3 (ps = 0.1332409)**
Belgium	Bulgaria	Greece
Denmark	Czech Republic	Spain
Ireland	Estonia	Cyprus
France	Croatia	Portugal
Italy	Latvia	
Luxembourg	Lithuania	
Malta	Hungary	
Austria	Poland	
Slovenia	Romania	
Finland	Slovakia	
Sweden	Montenegro	
UK	Old Republic Yugoslav of Macedonia	
	Turkey	

ps: population share

Source: Author

To compare the clustering results of applying the previous standard mixture models, we used the R SelvarMix package by Sedki et al. (2014) to perform variable selection under the assumptions of a general and a common covariance matrix. For variable selection for the mixture model, the R SelvarMix package combines the penalization method of Zhou et al. (2009) used firstly to rang the variables and secondly the SRUW method ofMaugis et al. (2009) based on the initially ranged variables.

Assuming a general covariance matrix, the method selects, on the one hand, a model with a general covariance matrix where the clusters have a common shape and orientation. On the other hand, with regard to variable selection, the method considers variables v1, v3, v5 and v6 as relevant, variables v52 and v53 as irrelevant, while the other variables are considered redundant.

The results obtained are presented in terms of variable selection in Table A.3, in terms of parameter values in Table 3, and in terms of clustering result in Table 4.

Table 3: The SRUW method parameters values under the assumption of a general covariance matrix

Cluster 1				
Proportion =	0.8667			
Means =	-0.0517	-0.3141	-0.1495	-0.1437
Variances =	0.9493	0.2951	0.3618	0.7728
	0.2951	0.3252	-0.1331	0.3198
	0.3618	-0.1331	0.8213	0.2323
	0.7728	0.3198	0.2323	0.8325
Cluster 2				
Proportion =	0.1333			
Means =	0.3358	2.0419	0.9720	0.9338
Variances =	0.9493	0.2951	0.3618	0.7728
	0.2951	0.3252	-0.1331	0.3198
	0.3618	-0.1331	0.8213	0.2323
	0.7728	0.3198	0.2323	0.8325

Source: Author

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Table 4: The clustering result of the SRUW method under the assumption of a general covariance matrix

Group	
1 (ps = 0.8667)	**2 (ps = 0.1333)**
Belgium	Greece
Bulgaria	Spain
Czech Republic	Italy
Denmark	Portugal
Estonia	
Ireland	
France	
Croatia	
Cyprus	
Latvia	
Lithuania	
Luxembourg	
Hungary	
Malta	
Austria	
Poland	
Romania	
Slovakia	
Slovenia	
Finland	
Sweden	
UK	
Norway	
Montenegro	
Old Republic Yugoslav of Macedonia	
Turkey	

ps: population share

Source: Author

Furthermore, under the assumption of a common spherical variance matrix, the method considers variables v1, v2, v3, v4, and v5 as relevant, variables v49, v52, and v53 as irrelevant, and the other variables as redundant.

The results obtained are presented in terms of variable selection in Table A.4, in terms of parameter values in Table 5, and in terms of clustering result in Table 6.

Table 5: The SRUW method parameters values under the assumption of a spherical covariance matrix

Cluster 1					
Proportion =	0.7608				
Means =	-0.3552	-0.4123	-0.1700	-0.4565	-0.3724
Variances =	0.5390	0.0	0.0	0.0	0.0
	0.0	0.5390	0.0	0.0	0.0
	0.0	0.0	0.5390	0.0	0.0
	0.0	0.0	0.0	0.5390	0.0
	0.0	0.0	0.0	0.0	0.5390
Cluster 2					
Proportion =	0.2392				
Means =	1.1299	1.3117	0.5408	1.4522	1.1848
Variances =	0.5390	0.0	0.0	0.0	0.0
	0.0	0.5390	0.0	0.0	0.0
	0.0	0.0	0.5390	0.0	0.0
	0.0	0.0	0.0	0.5390	0.0
	0.0	0.0	0.0	0.0	0.5390

Source: Author

Table 6: The clustering result of the SRUW method under the assumption of a spherical covariance matrix

Group	
1 (ps = 0.7608)	**2 (ps = 0.2392)**
Belgium	Greece
Bulgaria	Croatia
Czech Republic	Italy
Denmark	Cyprus
Estonia	Malta
Ireland	Montenegro
Spain	Old Republic Yugoslav of Macedonia
France	
Latvia	
Lithuania	
Luxembourg	
Hungary	
Austria	
Poland	
Portugal	
Romania	
Slovakia	
Slovenia	
Finland	
Sweden	
UK	
Norway	
Turkey	

ps: population share

Source: Author

Results interpretation

On the one hand, if we compare the clustering result based on a mixture model with full-covariance matrices with the result based on the SRUW variable selection method, we find that, with the exception of Greece, Spain, Italy and Portugal, all other countries change their groups labels when classified according to the SRUW method. This difference leads to an adjusted Rand index of -0.03360489 and a Rand index of 0.4850575, indicating a very low agreement between the two clustering results. On the other hand, the SRUW method shows that the relevant variables are only v1, v3, v5 and v6, which facilitates the description of the obtained

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

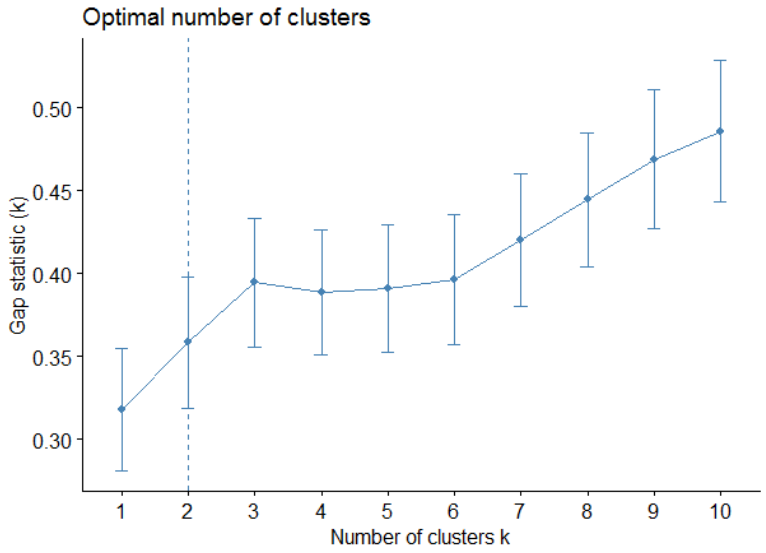
IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

groups. Group 1, which includes Greece, Spain, Italy and Portugal, are characterized by significant values for variables v1 (bread and cereals), v3 (fish and seafood), v5 (oils and fats) and v6 (fruit), unlike the other groups.
On the other hand, the comparison of the clustering result between the Gaussian matrix model, for which a number of 3 groups were chosen, and the SRUW variable selection method, for which the optimal number of groups is 2, assuming common spherical variance matrices between the clusters, shows that the number of groups is different and leads to different classifications. This difference is measured by an adjusted Rand index of 0.03413822 and a Rand index of 0.4850575, expressing weak agreement between the two clustering results. Regarding the selection of variables, the SRUW method selects v1 (bread and cereals), v2 (meat), v3 (fish and seafood), v4 (milk, cheese and eggs) and v5 (oils and fats), which facilitates the description and interpretation of the obtained groups. Group 1 is characterized by low values for v1, v2, v3, v4 and v5, in contrast to group 2.

4.1.1.2.Standard and regularized K-means algorithm

To eliminate the effects of redundant variables, we applied the standard K-Means algorithm and the regularized K-Means algorithm (Talibi et al., 2017a) to the data that contained only the relevant and irrelevant variables selected by the SRUW method assuming a common spherical variance matrix (v1, v2, v3, v4, v5, v53, v52, and v49).
To select the optimal number of classes for the standard K-Means algorithm, the gap statistic was used. Figure 3 shows that the optimal choice is two classes. Based on this result, we obtained the clustering result shown in Table 7.

Figure 3: Selection of the number of classes for the K-means algorithm by the Gap statistic



Source: Author

Table 7: The clustering result of the standard K-means

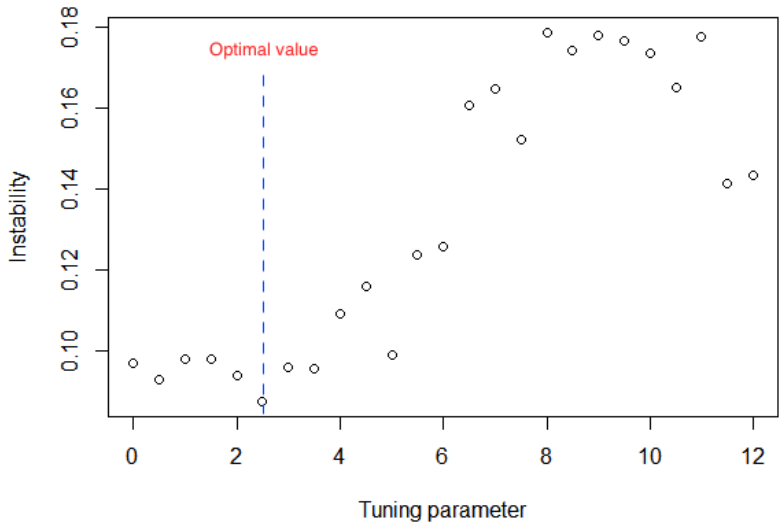
Group	
1 (ps = 0. 4666667)	**2 (ps = 0. 5333333)**
Bulgaria	Belgium
Czech Republic	Denmark
Estonia	Ireland
Croatia	Greece
Latvia	Spain
Lithuania	France
Hungary	Italy
Poland	Cyprus
Portugal	Luxembourg
Romania	Malta
Slovakia	Austria
Montenegro	Slovenia
Old Republic Yugoslav of Macedonia	Finland
Turkey	Sweden
	UK
	Norway

ps: population share

Source: Author

Moreover, for the regularized K-Means algorithm, we have considered that the number of groups is known and fixed at 2. In addition, the value of the regularization parameter that maximizes the stability of the clustering result leads us to choose a value of 2.5, as shown in the diagram in Figure 4, indicating that the means of some variables should be regularized.

Figure 4: Selection of the regularization parameter value for the regularized K-means algorithm



Source: Author

Based on these results, the variable selection presented in Table 8 leads to the retention of the four variables selected as relevant (v1, v2, v3, v4, v5) in addition to variable v52, which was already classified as irrelevant by the SRUW method, considering that a centered mean of 0.016 is not significantly different from a mean of -0.016.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Table 8: The centered means and variables weights used for the regularized K-means algorithm

Variable	Weight	Centered means	
		Cluster 1	Cluster 2
1	1.177698e+00	-0.3949173	0.5943352
2	6.489132e-01	-0.5320224	1.0611319
3	1.980829e+00	-0.2612431	0.2612431
4	7.562189e-01	-0.4928326	0.9398820
5	1.132542e+00	-0.4027133	0.6250692
49	5.577755e+04	0.0000000	0.0000000
52	3.601374e+00	0.01605299	-0.01605299
53	3.269620e+01	0.0000000	0.0000000

Source: Author

The clustering result of the regularized K-means algorithm is presented in the table 9.

Table 9: The clustering result of the regularized K-means

Group	
1 (ps = 0.7)	**2 (ps = 0.3)**
Belgium	Greece
Bulgaria	Spain
Czech Republic	Croatia
Denmark	Italy
Estonia	Cyprus
Ireland	Malta
France	Montenegro
Latvia	Old Republic Yugoslav of Macedonia
Lithuania	
Luxembourg	
Hungary	
Austria	
Poland	
Portugal	
Romania	
Slovakia	
Slovenia	
Finland	
Sweden	
UK	
Norway	
Turkey	

ps: population share

Source: Author

Results interpretation

The comparison between the clustering result of the standard K-Means algorithm and that of the regularized K-Means algorithm shows that using the irrelevant variables leads to a very different classification result. The adjusted Rand index between the two results is -0.03044755 (Rand index = 0.4827586), which expresses a very large disagreement between the two clustering results. However, the adjusted Rand index between the clustering result of the regularized K-Means algorithm and that of applying the SRUW method with common spherical variance matrices is 0.7334559 (Rand index = 0.8712644), reflecting a large agreement between the two results.

4.1.2. Clustering application on qualitative data

To apply clustering techniques to categorical data, the traditional latent class model and the penalized latent class model (Talibi et al., 2017b) we extracted data corresponding to ten variables (v170 - v180) from question 52 of the Family Life and Marriage Items of the questionnaire of the 15/04/2016 version of the updated European Values Study data (EVS, 2016), as shown in Table A.6.

4.1.2.1.Traditional latent class

For the traditional latent class model, we have assumed a maximum number of groups of 5 to facilitate the interpretation of the results. The choice of this maximum value is justified by an entropy value of 0.9, which we consider sufficient. The model chosen based on the BIC is indeed the 5-class model (see Table 10).

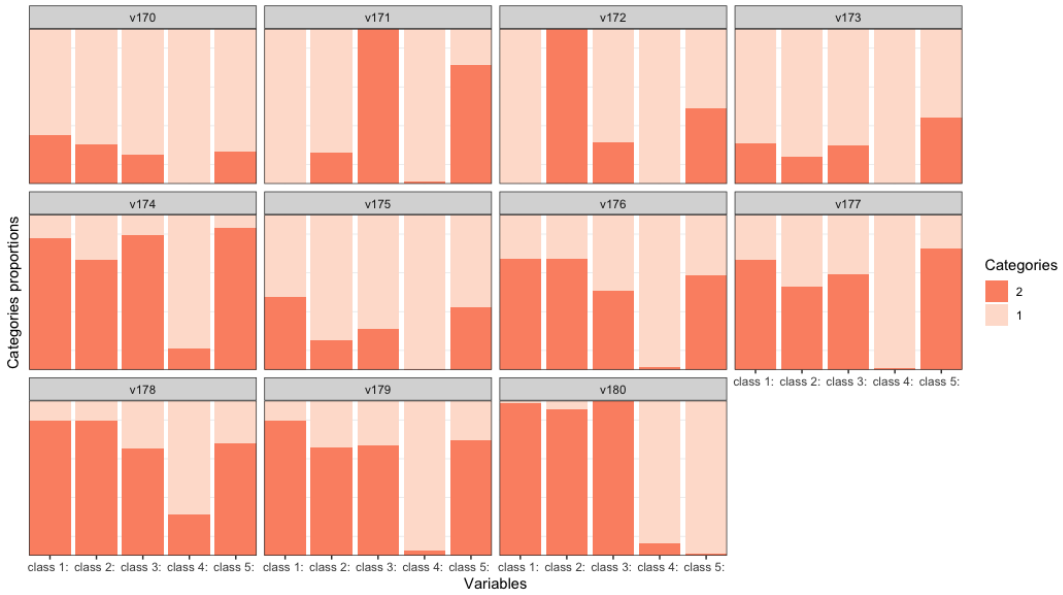
Table 10: Criteria for choosing the number of classes for the traditional latent class model

Model	Nbr of classes	Log- likelihood	Degree of freedom	G^2	BIC	Entropy
1	2	-387625.83	2024	102359.20	775504.40	0.46
2	3	-382565.76	2012	92239.07	765516.12	0.52
3	4	-380461.05	2000	88029.63	761438.54	0.67
4	5	-378658.33	1988	84424.20	757964.98	0.90

Source: Author

The proportions of categories of each variable in each class are shown in Figure 5, and the result of clustering using the traditional latent class model is shown in Figure 6. For the latter, the groups were ranged by size and described by the 10 most common nationalities.

Figure 5: The variables' categories proportions estimated by the traditional latent class model

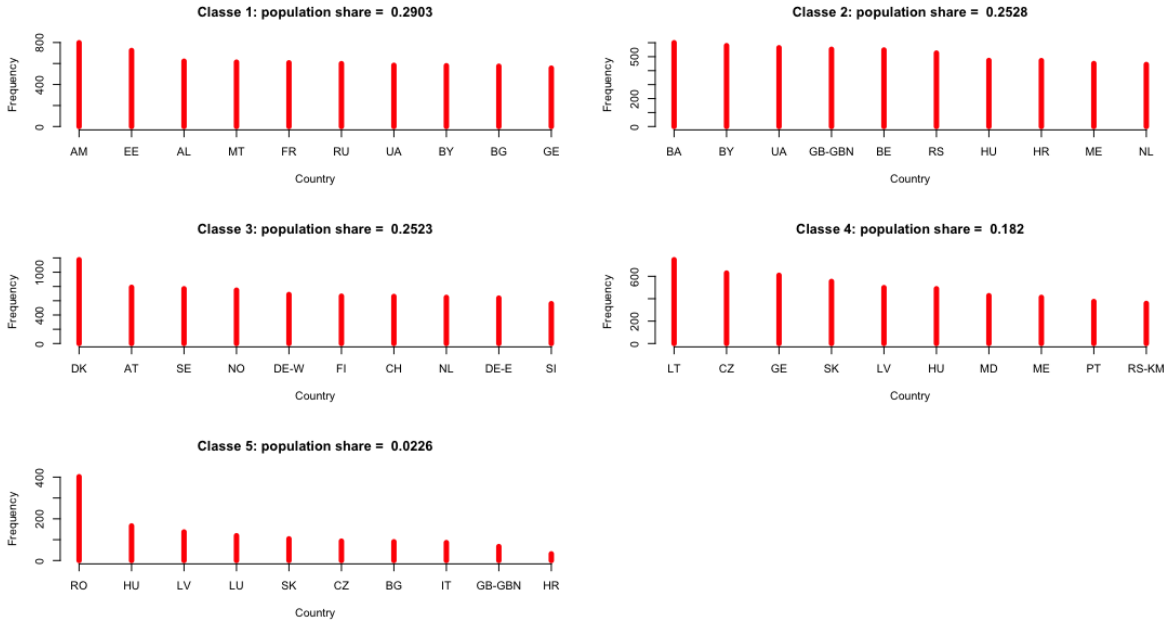


Source: Author

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Figure 6: The clustering result obtained by the traditional latent class model

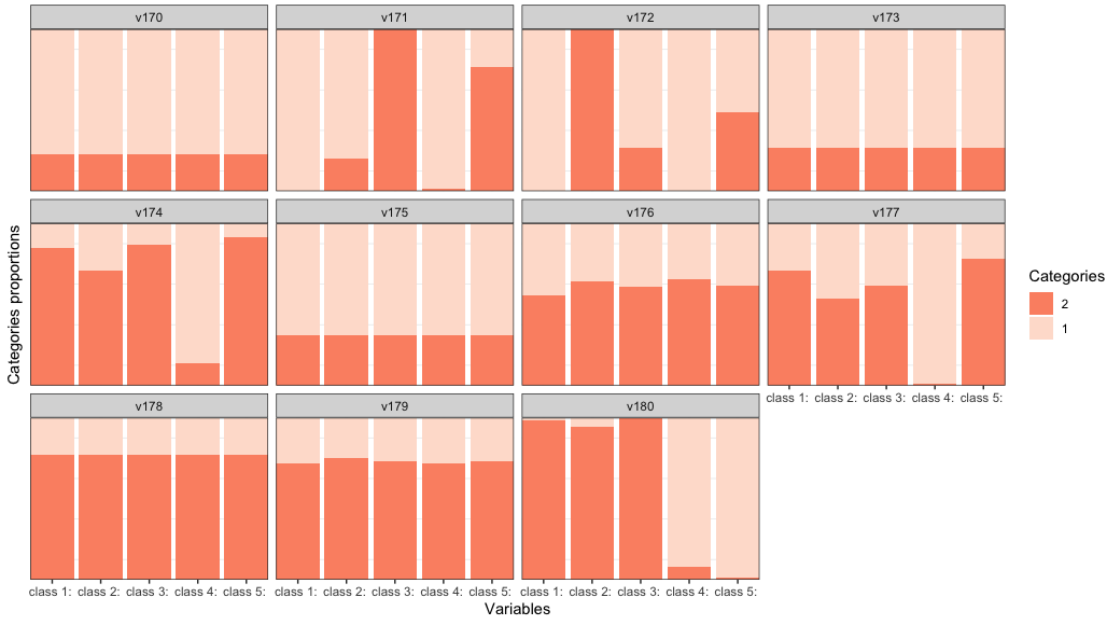


Source: Author

4.1.2.2.The penalized latent class model

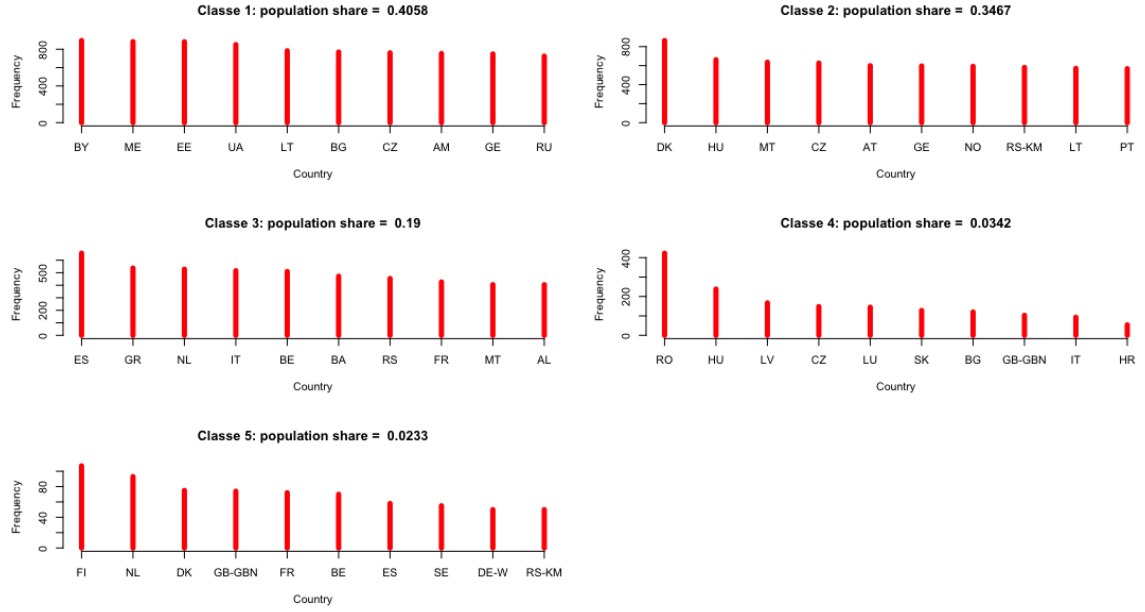
To compare the clustering result of applying the traditional latent class model with that of the penalized latent class model, we considered the same number of classes for the latter.
With a value of the regularization parameter of 1.8, the penalized latent class model, for which the value of the Pearson Chi^2-squared statistic is 553302.6, and the value of the Khi-squared likelihood ratio statistic G^2 is 165181.5 at 2009 degrees of freedom, has a BIC criterion value of 143105.9 and an AIC criterion value of 161163.5.
The proportions of the variable categories and the clustering result estimated by the penalized latent class model are shown in the diagram in Figure 7 and the graphs in Figure 8, respectively.

Figure 7: The variables' categories proportions estimated by the penalized latent class model



Source: Author

Figure 8: The clustering result obtained by the penalized latent class model



Source: Author

Results interpretation

Figure 7 shows that the penalized latent class model regularizes the estimation of the proportion values of the potentially irrelevant variables, which resulted in variables v170; v173; v175 and v178 being selected as irrelevant.

Thus, the groups obtained do not differ in their perceptions of the father's characteristics corresponding to these variables. Indeed, a large proportion of each group believe that parents should try to raise their children to be well-mannered, responsible, tolerant and respectful of others. Yet, parents should not encourage their children to have a particular religious belief.

Considering these variables as irrelevant results in a different clustering result. In contrast to the traditional latent class model, the groups obtained by the penalized latent class model are constructed so that the distributions of these irrelevant variables are the same in all groups.

4.2. Discussion

We conclude from the numerical experiments and applications that variable selection is necessary for international segmentation. In contrast to factorial analysis, which seeks to lower the number of initial variables in a reduced set of dimensions and results in information loss i.e. Day et al. (1988), Sriram & Gopalarishna (1991), Peterson & Malhotra (2000), Steenkamp (2001), Cavusgil et al. (2004) Dubois et al. (2005) and Budeva & Mullen (2014), variable selection allow us to maintain the relevant variables that most differentiate the groups formed.

In other words, variable selection allows for appropriate analyses and relevant decisions because it offers two advantages. First, clustering results are improved by eliminating irrelevant variables that may lead to a poor classification. Secondly, the description and interpretation of the results is facilitated as they are based on a limited number of relevant variables compared to the original number of variables. Indeed, the variable selection makes it possible to conduct a correct study of the subject of research at hand.

Therefore, variable selection is necessary for international segmentation. Creating segmentation through clustering methods that include variable selection enables the search for relevant segments and can allow relevant analyses to be carried out and support appropriate decision-making.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

5. Conclusion

In international or global marketing research, where international segmentation has been undertaken, a variety of data analysis techniques are used to understand, interpret and analyze a particular research topic. This is especially true for clustering to obtain homogeneous groups of countries or international consumers.

Clustering technics aims to classify objects of a population into groups, where the objects in the same group are similar to each other, and the objects in different groups are dissimilar. Unlike the supervised classification where the number of groups is known in advance, at least for a sample, in the case of clustering, it is unknown how many groups and it remains to be estimated. Given the clustering characteristics, many fields of research i.e. international marketing, global marketing, used clustering methods on data sets, in order to obtain a priori unknown groups that allow understanding and interpreting the phenomenon studied.

Usually, the database used to contain several socio-economic, cultural or psychographic variables for which factor analysis was often performed before clustering, i.e. Day et al. (1988), Sriram & Gopalarishna (1991), Peterson & Malhotra (2000), Steenkamp (2001), Cavusgil et al. (2004) Dubois et al. (2005) and Budeva & Mullen (2014), in order to reduce the original number of variables to a small number of dimensions and thus facilitate the interpretation of the results. However, factor analysis usually implies a loss of information.

In our study, we have applied standard clustering methods i.e. standard Gaussian mixture model, the standard K-Means algorithm, the traditional latent class model, and variable selection methods i.e. the SRUW method of Maugis et al. (2009), the regularized K-Means algorithm (Talibi et al., 2017a), the penalized latent class model (Talibi et al., 2017b). The purpose was to show the benefits and advantages of variable selection methods in terms of both clustering result and ease of interpretation of the groups obtained.

The use of variable selection methods, on the other hand, shows that unlike factor analysis techniques, which involve a loss of information, it is possible to keep only the relevant variables for which the differences between observations belonging to different groups are maximal. In this way, the clustering result is improved by considering only the information necessary for the segmentation task, while facilitating the interpretation of the groups obtained.

We believe that variable selection should be considered to reduce the data dimension before thinking of conducting a factor analysis.

Overall, we believe that variable selection is necessary for international segmentation as it allows for an adequate investigation of the research topic under discussion. Creating segmentation through clustering methods, taking into account the selection of relevant variables, makes it possible to find relevant segments and can lead to correct analysis and decision-making.

However, we consider it significant to use factor analysis techniques such as principal component analysis for quantitative data and multiple correspondence analysis for qualitative before the clustering task. Then, compare clustering results based on dimensions or factors retained by these dimension reduction techniques with these obtained by variable selection methods.

References:

(1) Arias-Castro, E. & Pu, X. (2017). A simple approach to sparse clustering. Computational Statistics & Data Analysis, 105:217–228.
(2) Biernacki, C., Celeux, G., Govaert, G., &Langrognet, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. Computational Statistics & Data Analysis, 51(2), 587-600.
(3) Bijmolt, T. H., Paas, L. J., & Vermunt, J. K. (2004). Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. International Journal of Research in Marketing, 21(4):323–340.
(4) Brooksbank, R. (1994). The anatomy of marketing positioning strategy. Marketing Intelligence & Planning, 12(4):10–14.
(5) Budeva, D. G. & Mullen, M. R. (2014). International market segmentation: Economics, national culture and time. European Journal of Marketing, 48(7-8):1209–1238.
(6) Cavusgil, S. T., Kiyak, T., & Yeniyurt, S. (2004). Complementary approaches to preliminary foreign market opportunity assessment: Country clustering and country ranking. Industrial Marketing Management, 33(7):607–617.
(7) Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern recognition, 28(5), 781-793.
(8) Christensen, T. (2002). Eurobarometer 56.0: Information and Communication Technologies, Financial Services, and Cultural Activities, August-September 2001. Inter-university Consortium for Political and Social Research.
(9) Cleveland, M., Papadopoulos, N., & Laroche, M. (2011). Identity, demographics, and consumer behaviors: International market segmentation across product categories. International Marketing Review, 28(3):244–266.
(10) Clogg, C. C. (1995). Latent class models. In Handbook of statistical modeling for the social and behavioral sciences (pp. 311-359). Springer, Boston, MA.
(11) Day, E., Fox, R. J., & Huszagh, S. M. (1988). Segmenting the global market for industrial goods: issues and implications. International Marketing Review, 5(3):14–27.
(12) Dubois, B., Czellar, S., & Laurent, G. (2005). Consumer segments based on attitudes toward luxury: Empirical evidence from twenty countries. Marketing letters, 16(2):115–128.
(13) EVS (2016). European values study (2008): Integrated dataset (evs 2008). Fernandez, G. (2010). Statistical data mining using SAS applications. CRC press.
(14) Fraley, C., & Raftery, A. E. (1998). MCLUST: Software for model-based cluster and discriminant analysis. Department of Statistics, University of Washington: Technical Report, 342.
(15) Goodman, L. A. (1983). The globalization of markets. Harvard Business Review. Yip, G. S. (1995). Total global strategy. Englewood Cliffs, 61(3):92--102.
(16) Gupta, V., Hanges, P. J., & Dorfman, P. (2002). Cultural clusters: Methodology and findings. Journal of world business, 37(1):11–15.
(17) Hassan, S. S. & Katsanis, L. P. (1991). Identification of global consumer segments: a behavioral framework. Journal of International Consumer Marketing, 3(2):11–28.
(18) Helsen, K., Jedidi, K., & DeSarbo, W. S. (1993). A new approach to country segmentation utilizing multinational diffusion patterns. The Journal of marketing, pages 60–71.
(19) Hernani-Merino, M., Lazo, J. G. L., López, A. T., Mazzon, J. A., & López-Tafur, G. (2020). An international market segmentation model based on susceptibility to globalconsumer culture. Cross Cultural & Strategic Management.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

(20) Hernani-Merino, M., Mazzon, J.A. & Isabella, G. (2015), "A model of susceptibility to global consumer culture", Review of Business Management, Vol. 17 No. 57, pp. 1212-1227.
(21) Hofstede, G. (1980). Motivation, leadership, and organization: do American theories apply abroad? Organizational dynamics, 9(1):42–63.
(22) Hofstede, G. (1984). Culture's consequences: International differences in work-related values, volume 5. sage.
(23) Hofstede, G. et al. (1991). Organizations and cultures: Software of the mind. McGrawHill, New York.
(24) Hofstede, G., Neuijen, B., Ohayv, D. D., & Sanders, G. (1990). Measuring organizational cultures: A qualitative and quantitative study across twenty cases. Administrative science quarterly, pages 286–316.
(25) Inglehart, R. & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. American sociological review, pages 19–51.
(26) Kale, S. H. (1995). Grouping euroconsumers: a culture-based clustering approach. Journal of International Marketing, 3(3):35–48.
(27) Kramer, H. E. & Herbig, P. A. (1994). Cultural differences in doing business: Germany and the sout. Review of Business, 16(2):33.
(28) Langrognet, F., Lebret, R., Poli, C., Iovleff, S., Auder, B., & Iovleff, S. (2019). Rmixmod: classification with mixture modeling. R package version, 2(2.2).
(29) Law, M. H., Figueiredo, M. A., & Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. IEEE transactions on pattern analysis and machine intelligence, 26(9):1154–1166.
(30) Lazarsfeld, P., & Henry, N. (1968). Latent Structure Analysis Houghton Mifflin. New York.
(31) Lee, C. (1990). Determinants of national innovativeness and international market segmentation. International Marketing Review, 7(5).
(32) Linzer DA, Lewis JB (2011). "poLCA: An R Package for Polytomous Variable Latent Class Analysis." Journal of Statistical Software, 42(10), 1–29.
(33) Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L.M. Cam and J. Neyman, editors, Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, volume 1 University of California Press.
(34) Maugis, C., Celeux, G., & Martin-Magniette, M.-L. (2009). Variable selection in model-based clustering: A general variable role modeling. Computational Statistics & Data Analysis, 53(11):3872–3882.
(35) Mclachlan, G.& Basford, K.E. (1988). BASFORD. Mixture models: inference and applications to clustering. New York: Marcel Dekker.
(36) Meynet, C. & Maugis-Rabusseau, C. (2012). A sparse variable selection procedure in model-based clustering. Research report.
(37) Nachum, L. (1994). The choice of variables for segmentation of the international market. International Marketing Review, 11(3):54–67.
(38) Papadopoulos, N. & Martín Martín, O. (2011). International market selection and segmentation: perspectives and challenges. International Marketing Review, 28(2):132–149.
(39) Peterson, M. & Malhotra, N. (2000). Country segmentation based on objective quality-of-life measures. International Marketing Review, 17(1):56–73.
(40) R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

(41) Raftery, A. E. & Dean, N. (2006). Variable selection for model-based clustering. Journal of the American Statistical Association, 101(473):168–178.
(42) Schwartz, S. H. (1994). Beyond individualism/collectivism: New cultural dimensions of values. Cross-cultural research and methodology series, Vol. 18. Individualism and collectivism: Theory, method, and applications:85–119.
(43) Schwartz, S. H. (1997). Values and culture. In D. Munro, J. F. Schumaker, & S. C. Carr (Eds.), Motivation and culture, pages 69–84.
(44) Sedki, M., Celeux, G., & Maugis-Rabusseau, C. (2014). Selvarmix: Ar package for variable selection in model-based clustering and discriminant analysis with a regularization approach. INRIA Techical report.
(45) Sethi, S. P. & Holton, R. H. (1973). Country typologies for the multinational corporation: a new basic approach. California Management Review, 15(3):105–118.
(46) Sethi, S. P. (1971). Comparative cluster analysis for world markets. Journal of Marketing Research, pages 348–354.
(47) Silvestre, C., Cardoso, M. G., & Figueiredo, M. (2015). Feature selection for clustering categorical data with an embedded modeling approach. Expert systems, 32(3):444–453.
(48) Sriram, V. & Gopalakrishna, P. (1991). Can advertising be standardized among similar countries? a cluster-based analysis. International Journal of Advertising, 10(2):137–149.
(49) Steenkamp, J.-B. E. & Ter Hofstede, F. (2002). International market segmentation: issues and perspectives. International journal of research in marketing, 19(3):185–213.
(50) Steenkamp, J.-B. E. (2001). The role of national culture in international marketing research. International Marketing Review, 18(1):30–44.
(51) Sun, W., Wang, J., Fang, Y., et al. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. Electronic Journal of Statistics, 6:148–167.
(52) Talibi, A., Achchab, B., & Gutiérrez-Sánchez, R. (2017a). Variable selection for clustering with regularized k-means algorithm. In Proceedings of Data Engineering In Bioinformatics, Image and Data Analysis, Conference of the Moroccan Classification Society (SMC'2017), Tangier (Morocco), pages 96–102.
(53) Talibi, A., Achchab, B., Nafidi, A., & Gutiérrez-Sánchez, R. (2017b). Penalized latent class model for clustering with application to variable selection. In First International Conference on Real Time Intelligent Systems, pages 55–65. Springer.
(54) Tryon, R. C. & Bailey, D. E. (1966). The bc try computer system of cluster and factor analysis. Multivariate Behavioral Research, 1(1):95–111.
(55) Vermunt, J. K. (2003). Multilevel latent class models. Sociological methodology, 33(1):213–239.
(56) Wang, S. & Zhu, J. (2008). Variable selection for model-based highdimensionalclustering and its application to microarray data. Biometrics,64(2):440–448.
(57) Wolfe, J. H. (1963). Object cluster analysis of social areas. In Masters thesis. University of California, Berkeley.
(58) Xie, B., Pan, W., & Shen, X. (2008). Variable selection in penalized modelbasedclustering via regularization on grouped parameters. Biometrics,64(3):921–930.
(59) Ye Sheng, S. & Mullen, M. R. (2011). A hybrid model for export marketopportunity analysis. International Marketing Review, 28(2):163–182.
(60) Yip, G. S. (1995). Instructor's Manual: Total Global Strategy: Managing for Worldwide Competitive Advantage. Prentice Hall.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

(61) Zandpour, F. & Harich, K. R. (1996). Think and feel country clusters: Anew approach to international advertising standardization. InternationalJournal of Advertising, 15(4):325–344.
(62) Zhou, H., Pan, W., & Shen, X. (2009). Penalized model-based clusteringwith unconstrained covariance matrices. Electronic journal of statistics,3:1473.
(63) Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

Appendix

Table A.1: Summary of the literature review

Authors	Sample size	Variables	Statistical Methods	Results	Main findings
Sethi (1971)	91 countries	29 socio-economic variables	V-Analysis, O-Analysis.	4 groups of variables, 7 groups of individuals.	There is no clearly defined continuum of economic development, Countries should be classified on the basis of several variables.
Day et al. (1988)	96 countries	18 economic variables measuring economic developme nt	Factor analysis, Clustering by FASTCLUS.	1st factor analysis: 3 factors uncorrelated with 2 variables, 2nd factor analysis: was carried out on the basis of 16 variables and also resulted in 3 factors, Two clustering analyses led to the selection of 6 clusters as the optimal solution.	To identify opportunities for standardization, commonalities between countries should be explored, Marketers looking for global marketing strategies should use the economic variables relevant to the product or service in question.
Lee (1990)	70 countries	The possession of white and black televisions and color televisions per thousand people in 1981 and 10 socio-economic variables	Correlation analysis, Stepwise regression, Hierarchical classification.	4 variables that are considered significant determinants of innovation, Based on the posterior determinants of innovation, a hierarchical classification was made to classify the countries into 5 clusters.	The results of clustering can be used by international marketing managers to target each group with specific communication tools or messages, Some cultural variables that influence innovation in a country should be combined with economic variables.
Sriram & Gopalaris hna (1991)	40 countries	9 economic variables, 4 cultural	Factor analysis (PCA), Hierarchical classification, Discriminant	Factor analysis led to the retention of 4 factors, 6 clusters were selected as the best clustering solution,	The use of cultural and media variables facilitates the interpretation of the results obtained, which can be used to standardize advertising strategies,

		dimensions and 7 media-related variables	analysis	All variables are relevant except radio and movie expenses.	Discriminant analysis, used as a method to select the best number of groups, can also be useful to measure the stability of clustering results.
Kale (1995)	17 countries of Western Europe	Hofstede's 4 cultural dimensions	Hierarchical classification, Non-hierarchical classification	The clustering result has led to select 3 groups as the best result.	Despite the emergence of the European Union, there are still cultural differences between European countries. Therefore, it is practical for marketers to look at Europe as groups of countries, The result of clustering can be useful in determining an appropriate type of advertising appeal for each group.
Zandpour & Harich (1996)	23 countries	Cultural Variables, National market-related variables, Media-related variables	Regression Analyzes	The 23 countries were analyzed in terms of the type of call and the corresponding commercial communication.	To identify opportunities for advertising standardization, marketers should use cultural variables, variables that describe a national market, and media-related variables instead of geographic variables.
Peterson & Malhotra (2000)	165 countries	6 variables measuring the quality of life	Correlation between the values of two different years, Factorial exploratory analysis with the maximum likelihood technique, Confirmatory factor analysis using structural equations, Hierarchical and non-hierarchical classification	Exploratory factor analysis led to the selection of 2 factors for each of the data periods considered, The confirmatory factor analysis confirmed the existence of the 2 factors, The clustering results led to the selection of 12 clusters as the best result.	Clustering results can be used by marketers to gain strategic advantages in terms of advertising strategy. Researchers can use the "IL QoL survey" data as a reference for research in international trade in general.
Steenkamp (2001)	24 countries	Hofstede's 4 cultural dimensions and the Schwartz's 7 cultural factors	Factor analysis (PCA), Hierarchical classification, K-means algorithm	Factor analysis led to the selection of 4 factors that explain a substantial part of the data variance, The scores of the 4 factors of the unified cultural framework were used to divide the	Culture is an extremely complex research topic that cannot be summarized in a few dimensions or factors.

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

				24 countries into 7 groups.	
Gupta et al. (2002)	61 countries of the GLOBE survey	Language, geography, religion and historical accounts, ethnicity, values and professional attitudes	Discriminant analysis	The authors have proposed a classification into 10 clusters, The validity of the proposed predefined grouping is tested by constructing a linear discriminant function.	It is useful to examine the commonalities between countries to identify expansion opportunities for businesses.
Cavusgil et al. (2004)	90 countries	29 socio-economic variables	Exploratory factor analysis (PCA), Hierarchical classification, K-means algorithm	Exploratory factor analysis has led to the selection of 5 factors that summarize a large part of the data variance, An optimal solution of 10 clusters was selected as the best clustering result and used as input to the K-means algorithm.	Clustering and country rankings allow marketers to assess international market opportunities, Clustering helps marketers determine relevant strategies for a particular group, Country rankings and clustering are extremely useful for screening markets and selecting a small group of potentially attractive markets.
Bijmolt et al. (2004)	15 EU countries	Possessing 8 financial products	Multilevel latent class model	7 country groups, 14 consumer groups.	International segmentation is an important tool for companies to formulate international strategies.
Dubois et al. (2005)	1848 management students from 20 countries	33 items measuring consumer attitudes	Factor analysis, Mixture model	The factor analysis showed that the 33 items could not be reduced to a small number of factors, The mixture model led to the retention of 3 classes as the best result.	In addition to cultural variables, psychological variables can also influence consumer attitudes towards luxury.
Budeva & Mullen (2014)	34 countries, included in the two surveys of the "World Value Survey"	Economic and cultural variables	Factor analysis (PCA), Hierarchical classification, K-means algorithm	Factor analysis reduced the original number of economic variables to 3 factors, Clustering based on the economic variables resulted in a solution with 4 groups for the two periods considered, Clustering based on the cultural variables also led to a solution with 4 groups for the two periods considered, Clustering based on the economic and	The clustering result based on cultural variables changes slowly compared to the result obtained based on economic variables, The clustering results are unstable over time, Countries should be classified on the basis of both economic and cultural variables.

				cultural variables led to a solution with 6 groups for the two periods considered.	
Hernani-Merino et al. (2020)	412 participants from 5 countries	seven dimensions of the theoretical model proposed by Hernani-Merino et al. (2015a)	Fuzzy C-Means	3 groups of individuals	Customers from different countries have common beliefs about the social responsibility of global brands.

Source: Author

Table A.2: Variables used in the average consumption expenditure (by consumption function) database

v1: Bread and cereals	v2: Meat	v3: Fish and marine food
v4: Milk, cheeses and eggs	v5: Oils and fats	v6: Fruits
v7: Vegetables	v8: Sugar, jams, honey, chocolate and confectionery	v9: food products n.e.
v10: Coffee, tea and cocoa	v11: Mineral water, soft drinks, fruit and vegetable juices	v12: alcoholic beverages
v13: Tobacco	v14: Articles of clothing	v15: Real residential rents
v16: Rent charged to housing	v17: Routine maintenance and repairs of the dwelling	v18: Water supply and other services related to housing
v19: Electricity, gas and other fuels	v20: Furniture, furnishings, carpets and other floor coverings and repairs	v21: Household textile articles
v22: Household appliances	v23: Glassware, dishes and household utensils	v24: Tools for home and garden
v25: Goods and services for routine maintenance of the dwelling	v26: Medical products, apparatus and equipment	v27: Outpatient services
v28: Hospital Services	v29: Vehicle purchases	v30: Use of personal vehicles
v31: Transportation Services	v32: Postal services	v33: Telephone and fax equipment
v34: Telephone and fax services	v35: Devices and accessories, including repairs	v36: Other durable goods important for recreation and culture
v37: Other recreational items and equipment, gardens and pets	v38: Recreational and cultural services	v39: Press, bookstore and stationery
v40: Package tours	v41: Nursery and primary education	v42: Secondary education
v43: Post-secondary education that is not higher education	v44: Higher Education	v45: Teaching not defined by degree
v46: Catering services	v47: Hosting Services	v48: Personal care
v49: Personal effects n.e.	v50: Social protection	v51: Insurance
v52: Financial Services n.e.	v53: Other services n.e.	

Source: Author

ISSN: 2658-8455
Volume 3, Issue 4-3 (2022), pp.498-530.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

Table A.3: The SRUW method results in terms of variable selection under the assumption of a general covariance matrix

Criterion:	**BIC**
Criterion value:	Inf
Number of clusters:	2
Gaussian mixture model:	Gaussian_pk_L_C
Regression covariance model:	LC
Independent covariance model:	LI

The SRUW model:	
S:	1 3 5 6
R:	1 3 5 6
U:	2 4 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
W:	53 52

Source: Author

Table A.4: The SRUW method results in terms of variable selection under the assumption of a spherical covariance matrix

Criterion:	**BIC**
Criterion value:	15035.09
Number of clusters:	2
Gaussian mixture model:	Gaussian_pk_L_I
Regression covariance model:	LC
Independent covariance model:	LI

The SRUW model:	
S:	1 2 3 4 5
R:	1 2 3 4 5
U:	6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 50 51
W:	53 52 49

Source: Author

Table A.5: Abbreviation of the countries names included in the European values study EVS

AL: Albania	AM: Armenia	AT: Austria
BA: Bosnia and Herzegovina	BE: Belgium	BG: Bulgaria
BY: Belarus	CA: Canada	CH: Switzerland
CY: Cyprus	CZ: Czech Republic	DE-E: Germany East
DE-W: Germany West	DK: Denmark	EE: Estonia
ES: Spain	FI: Finland	FR: France
GB-GBN: Great Britain	GE: Georgia	GR: Greece
HR: Croatia	HU: Hungary	IE: Ireland
IS: Iceland	IT: Italy	RS-KM: Kosovo
LT: Lithuania	LU: Luxembourg	LV: Latvia
MD: Rep. of Moldova	ME: Rep. of Montenegro	MK: Macedonia
MT: Malta	CY-TCC: Northern Cyprus	GB-NIR: Northern Ireland
NL: Netherlands	NO: Norway	PL: Poland
PT: Portugal	RO: Romania	RU: Russian Federation
SE: Sweden	SI: Slovenia	SK: Slovak Republic
TR: Turkey	UA: Ukraine	US: United States

Source: Author

Table A.6: Variables measuring qualities which children can be encouraged to learn at home

Variable		**Categories proportions**	
Number	**Name**	**Category 1 (Cited)**	**Category 2 (Not cited)**
v170	Good manners	0.7710375	0.2289625
v171	Independence	0.4829334	0.5170666
v172	Application at work	0.533339	0.466661
v173	Sense of responsibilities	0.7306522	0.2693478
v174	Imagination	0.1842007	0.8157993
v175	Tolerance and respect of others	0.6889321	0.3110679
v176	Saving spirit, do not waste money or things	0.384657	0.615343
v177	Determination, perseverance	0.3625718	0.6374282
v178	Religious faith	0.2271882	0.7728118
v179	Generosity	0.2716796	0.7283204
v180	Obedience	0.2772389	0.7227611

Source: Author