OPTIMAL CLUSTERING TECHNIQUES FOR THE SEGMENTATION OF TOURIST SPENDING. ANALYSIS OF TOURIST SURVEYS IN THE VALENCIAN COMMUNITY (SPAIN): A CASE STUDY

A. RABASA[1], A. PÉREZ-MARTÍN[1] & D. GINER[2]
[1]University Miguel Hernández of Elche, Spain
[2]Instituto Valenciano Tecnologías Turísticas, Agència Valenciana de Turismo, Spain

ABSTRACT

The Valencian Community (South-East Spain) is one of the most important tourist destinations in Europe. The Valencian Government has been carrying out surveys about the types of travel, the type of transport, the type of accommodation, the duration of the trip and the number of travellers, as well as other issues. The aim is to discover the different spending typologies incurred by foreign visitors.

In their task of drawing up more attractive tourist strategies, the following questions may become particularly relevant to the Valencian Public Services: what type of traveller spends more on transportation in their own country, or pays for it in the Valencian Community; visitors' nationalities and their higher or lower propensity to spend money on leisure; or the number of overnight stays in low-end destinations.

But the surveys gathering all this information consist of multiple and nested responses, distributed in thematic blocks that overlap, and whose translation to flat file systems (susceptible to being analysed with acceptable counting times) is a complex problem.

This paper presents a treatment process of the surveys, especially oriented towards having a suitable dataset to generate models of optimal segmentation of the different types of expenditure. Likewise, some results of such segmentation are shown, which are proving to be of great value to public managers in their challenge to offer suitable tourist alternatives to each type of traveller.

The paper includes an example of how open data sources can be incorporated into the original dataset in order to obtain better segmentation. A variation to the classical segmentation methods (algorithms of the K means family) is also provided, which leads to the establishment of the optimal number of groups for each computational experiment.

Keywords: big data, clustering, optimization, surveys analysis, tourism

1 INTRODUCTION AND OBJECTIVES

The Valencian Community (Spain) is one of the main tourist destinations for foreign visitors. In the last years, the tourist offer in the zone has diversified enormously. A wide range of activities and destinations has been incorporated, beyond the classic 'sun and beach' destination [1]. So, the target public has also been changing its profile.

For this reason, the correct segmentation of foreign tourists visiting Valencia has become an important objective for public tourism managers in the area. So, this segmentation is intended to achieve fundamentally, based on its different typologies of spending. The aim is to make groups of tourists of maximum similarity, with regard to the amounts they spend on different concepts (accommodation, transport, leisure, etc.), but also on how to carry out such expenditure (payments at source, payments at destination, group payments, etc.). In addition, all this information must also be related to the socio-economic data of the traveller and with the type of trip they have chosen (overnight stays, type of accommodation, etc.).

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V12-N4-482-491

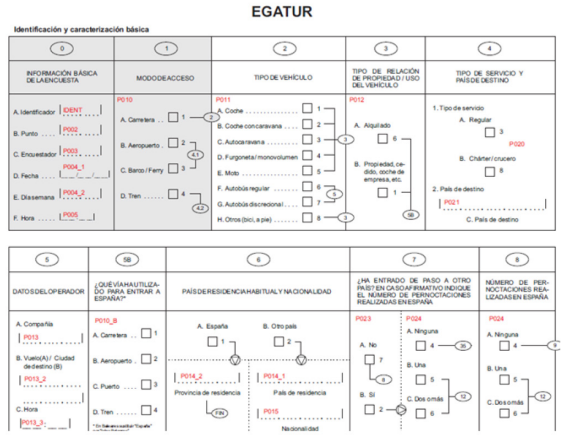

Figure 1: EGATUR survey.

2 INPUT DATA

The input dataset comes from surveys of foreign tourists visiting the Valencian Community (Spain), which are known as EGATUR surveys. They were designed by Turespaña [2] and the Spanish Statistical Office [3] and were carried out at different border points. The EGATUR surveys respond to a hierarchical design, using thematic blocks, where the response to one of the blocks automatically leads to another block. They gather different characteristics about the traveller (country, companions, employment status, etc.) and also about the trip (types of accommodation, overnight stays, excursions, etc.) and different typologies of expenditure (leisure, lodging, transport and total expenditure).

In general, and for the methodological explanations, the nomenclature of the variables (pXYZ) throughout the text has been respected. However, for the explanation of the results and to facilitate their interpretation, the nomenclature has been used by referring to the true meaning of the variable. Thus, for example, p097 is 'accompanied by'.

The input data consist of 57 data files, a survey per month from June 2011 to September 2015 (there was a readjustment of the survey in October 2015 by the Statistical Office). 139,700 records and 233 variables were imported. After carrying out the pre-processing steps described below, the dataset to be analysed consists of 97,442 registers and 110 variables.

3 ANALYTICAL METHODOLOGY

The proposed analytical methodology is as follows:

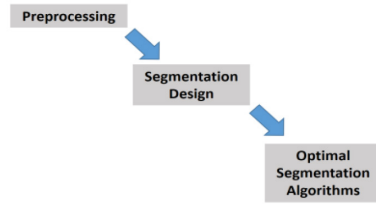

Figure 2: Analysis methodology. Phases.

a. Pre-processing

Several tasks that are usually faced on the preprocessing stage [4, 5] are as follows:

- Automatic selection of the most relevant characteristics

  There appear to be diverse variables, which describe realities from different points of view. In these cases, the most suitable variable for the analysis is chosen.

  On the other hand, for the subsequent clustering process quantitative variables are preferable, so this type of variable will predominate where there is a conflict

- Restructuring of variables

  We look for a series of variables that have a list of possible values, which can be restructured into different categories. This process involves, on the one hand, increasing the number of categories when it is of interest to contemplate other qualities within a specified category in the survey design; and on the other hand, decreasing the categories when the quantity of elements in one of these categories is so small that it is not representative.

  E.g.: Increasing the number of variables for the variable Reason (reason for the trip). The option 'leisure' breaks down into different types of leisure included in the variable p041, and by combining both variables greater accuracy is achieved.

  A reduction is also made in the number of possible values of the variable p097 (who people travel with), where the values (a total of seven values) included in the survey are based on four, which are Alone – Partner – Family – Friends. Besides, the variables – who paid for the package holiday (p103), the accommodation matrices (p213, p220, p232, p235), transport (p139_06, p144, p151, p162, p165, p181, p184, p195, p202) and other spending (p266, p268, p253) – are restructured into two single options:

  1 – Payment by the interested party

  2 – Payment by others (including family, company, free and others).

- Outlier or anomaly detection

  Outliers (out of range values) can become a serious problem for the correct representation of data because they distort the sample, so they have to be eliminated. Among the outliers that we withdraw from the rows analysed are, for example, all the trips that last for more than 180 days.

- Elimination of rows /and attributes (columns) with inconsistent values

  We also eliminate the variables p246_1 y p084_2_1, which refer to spending on own home accommodation and the amount spent for other extra expenses (respectively), since they include the imputation of values of the cost of the home itself and the accumulation of expenses. These variables are eliminated to avoid bias in the subsequent segmentation due to excessively high spending.

- Creation of variables of interest. Imputed variables

  Variables of interest derived from other variables given in the data or imported from open external sources are created. Among these generated variables we can highlight variables such as daily spending, or the GDP of the countries included in the data.

- Considerations regarding spending matrices

  After pre-processing and analysing the variables included in the spending matrices, one of the considerations that stands out is the impossibility of dividing the total amount by the number of days, since the price of the package holiday booked by this person can include accommodation as well as transport and other expenses. In the case of transport, there is a high percentage of consumers who have transport included in the price of their package holiday, so these consumers cannot be classified. On the other hand, we are able

to segment the price of transport for all those consumers who do not have transport included in their package holiday or who have not in fact booked a package holiday.

- File management and general considerations

  Based on 52 files with an average 2,700 records (rows) and 233 variables each. That is to say with a database of 139,700 surveys, which includes a total of 32.5 million data. After eliminating the records that correspond to day trippers (they do not stay overnight) and the repeated rows for each single identifier (according to agency instructions) there is a dataset of 109,387 surveys.

- Data files after pre-processing

  Finally, different survey files (annual and for the whole period of the study) were obtained with different aggregation levels. These pre-processed files are the input for the segmentation models. Different descriptive data from the series are obtained which can help establish the segmentation criteria to be applied. Below are some of the graphs obtained in this section (Figs. 3 and 4).

b. Segmentation design

This consists of generating segmentations (grouping according to similarity) of the different spending concepts, by considering the different characteristics of tourists and their holiday.

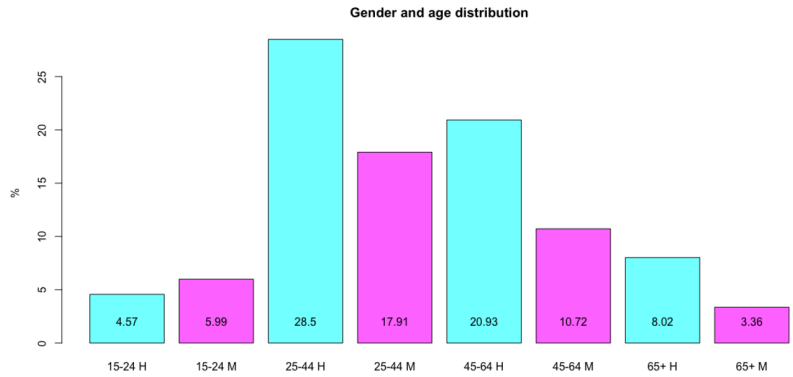

Figure 3: Gender and age distribution.

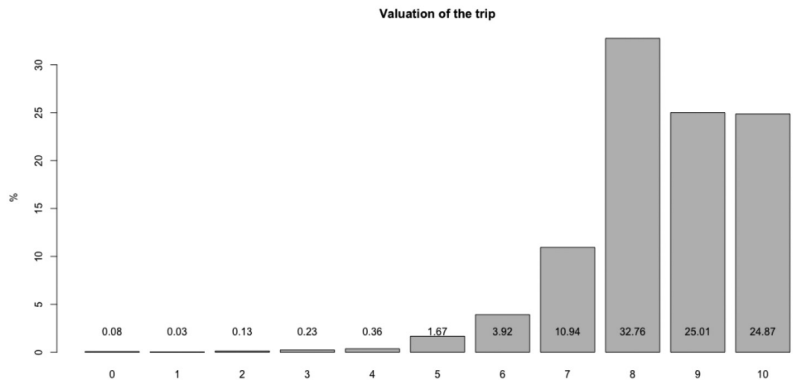

Figure 4: Valuation of the trip.

Table 1: Linear correlations between spending variables and those selected for the study.

	correlations					correlations			
variable	acomm spnd	transp spnd	leisu spnd	total spnd	acomm spnd	transp spnd	leisu spnd	total spnd	gasto Tot.
p004_2	0,030	-0,024	-0,003	-0,011	p268	-0,019	-0,145	0,042	-0,086
p010	0,009	0,173	0,062	0,146	G.p269_1	0,024	-0,043	0,378	0,129
p011	-0,032	-0,008	-0,006	-0,015	p280_1_1	-0,050	-0,051	-0,076	-0,077
p014_1	0,051	0,140	0,111	0,152	p280_1_2	-0,048	-0,053	-0,076	-0,078
p023	0,034	-0,217	0,015	-0,135	p280_2_1	-0,046	-0,063	-0,072	-0,083
p024	-0,063	-0,474	-0,079	-0,372	p280_2_2	-0,043	-0,049	-0,066	-0,071
p035	-0,012	0,160	-0,013	0,102	p280_3_1	-0,025	-0,045	-0,036	-0,052
p036	-0,029	0,103	-0,026	0,054	p280_3_2	-0,012	-0,036	-0,031	-0,040
p040_8	-0,140	-0,250	-0,181	-0,277	p280_4_1	0,021	-0,063	-0,048	-0,058
p041	-0,135	-0,141	-0,093	-0,165	p280_4_2	0,016	-0,029	0,015	-0,010
p044	0,028	0,046	0,039	0,054	p280_5_1	-0,052	-0,096	-0,096	-0,117
G.P084_2_1	0,033	0,009	0,049	0,034	p280_5_2	-0,011	-0,049	-0,033	-0,049
p093_1_1	-0,014	-0,089	-0,014	-0,070	p280_6_1	-0,070	-0,106	-0,125	-0,139
p093_1_2	-0,007	-0,014	-0,005	-0,013	p280_6_2	-0,070	-0,104	-0,123	-0,137
p095	0,284	0,076	0,036	0,130	p281	-0,010	-0,003	-0,031	-0,017
p096_1_1	0,347	-0,021	-0,008	0,060	p282	0,081	-0,010	0,043	0,029
p096_1_2	0,172	-0,013	-0,018	0,022	p283	-0,122	-0,113	-0,131	-0,158
p097	-0,006	-0,101	-0,070	-0,099	p285	-0,039	-0,014	-0,053	-0,040
p099	-0,234	-0,302	-0,378	-0,414	p286_2	0,014	-0,044	-0,025	-0,037
p099a	-0,195	-0,244	-0,316	-0,340	p286_2_1	0,011	-0,081	-0,023	-0,063
G.p100_1	0,007	0,013	0,045	0,029	p286_2_2	-0,233	-0,101	-0,028	-0,133
G.p137_1	-0,018	0,086	0,011	0,059	p286_3	0,002	-0,060	-0,015	-0,047
p139_06	0,224	0,214	0,110	0,242	p286_3_1	-0,004	-0,079	-0,017	-0,062
G.p139_07_1	0,169	0,343	0,188	0,350	p286_3_2	-0,145	-0,088	-0,021	-0,101
p144	-0,006	0,051	-0,012	0,029	p290	-0,042	-0,149	-0,025	-0,122
G.p145_1	-0,002	0,081	0,003	0,056	origen	0,068	0,027	0,027	0,045
p151	0,000	0,037	0,004	0,027	AM	0,014	0,003	0,011	0,010
G.p152_1	-0,002	0,033	0,005	0,024	NPernocta	-0,140	-0,250	-0,181	-0,277
p162	0,007	0,123	0,001	0,086	Aloja2	-0,386	-0,064	-0,107	-0,173
G.p163_1	0,021	0,176	0,040	0,141	Peso	-0,041	-0,139	-0,074	-0,134
p165	-0,056	0,109	-0,039	0,046	G.GastoTotal	0,200	0,101	0,270	0,224
G.p170_1	-0,024	0,179	0,011	0,122	p040_1	0,115	0,077	0,034	0,093
p181	0,036	0,008	0,029	0,025	p040_2	-0,504	-0,165	-0,109	-0,270
G.p182_1	0,013	0,021	0,031	0,030	p040_5	-0,133	-0,247	-0,178	-0,272
p184	0,142	0,067	0,071	0,107	p072	0,402	0,165	0,088	0,238
G.p185_1	0,051	0,044	0,064	0,067	p213	0,293	0,083	0,042	0,139
p195	0,003	-0,003	-0,047	-0,020	G.p214_1	0,347	0,034	0,051	0,122
G.p200_1	0,016	0,033	0,003	0,028	p220	0,479	0,239	0,088	0,307
p202	0,052	0,069	-0,018	0,051	G.p221_1	0,446	0,060	0,063	0,166
G.p203_1	0,047	0,065	0,030	0,067	p232	0,037	-0,069	-0,060	-0,064
p253	0,188	0,121	0,125	0,176	G.p233_1	0,093	-0,033	-0,013	-0,008
G.p254_1	0,075	-0,028	0,264	0,105	p235	0,101	-0,088	-0,030	-0,051
G.p256_1	-0,068	-0,116	0,080	-0,062	G.p240_1	0,149	-0,055	0,003	-0,003
p266	0,026	0,027	0,072	0,054	G.p246_1	-0,033	-0,014	0,029	-0,005
G.p267_1	0,020	0,077	0,209	0,142					

In order to decide objectively about the most suitable segmentation criteria and the variables that should intervene in each case, we carried out Fisher's linear correlation calculation of the survey variables, with respect to each level of daily spending per person. Although they may not be exactly what are subsequently used for the different segmentations, it gives an idea in advance about which variables are to be used.

So, based on the user typology as the most important issue, the segmentation of four types of spending is carried out:

1- Accommodation spending

2- Leisure spending

3- Transport spending
4- Total spending

All of the spending is given in euros, daily per person.

c. Optimal segmentation algorithms, based on optimal k-means clustering techniques - Grouping Models. Clustering techniques with variations to the K-Means algorithm [7]: This family of applied algorithms generates groups of responses (from the survey) which are similar, taking into account user, booking, establishment or zone typology. There are several variants and adaptations [8] from the original algorithm.

Optimal K-Means: the original segmentation method has been modified to find the optimal number of groups (clusters) in each case, maximizing the similarity between the members of each group.

The algorithm starts by assuming that the average similarity between members of the same group is decreased, and on entering a loop in search of the average similarity between the members of the groups it continues to increase. It leaves the loop when the similarity begins to decrease and the one that presents a maximum similarity between its members is chosen as optimum clustering. Next, the pseudocode for such a procedure is given.

```
BEGIN
Initialize DecreasingSimilarity=TRUE

While (DecreasingSimilarity)
      BEGIN
            clusters_center=change;
            k=2;
            While (clusters_center=change AND iterations <
      MAX_ITERATIONS)
            BEGIN
            (1) Initialize centers (k);
            (2) Redistribute items on clusters using minimum
                euclidean distance as classifier: clusters
                (k);
            (3) Calculate (similarity(k),
                DecreasingSimilarity);
            (4) Increase k;
            END
      END

k_opt = MAX {similarity(k)};

Return clusters (k_opt);
END
```

4 RESULTS

Below some of the significant segmentations obtained are presented for each of the families and the previously numbered types of spending. The procedure for selecting the variables included in each of the previously numbered segmentation families is a combination of

variables that are logically liable to intervene together with the linear correlations outlined above.

User typology is one of the most relevant aspects for public administration managers in terms of establishing different spending groups. So, it is this type of user that has been used in the segmentation models presented in this paper. Variables are used as a reason for travel or activity. These variables, highlighted within the mentioned typology, will be used within the model combined with the different expenses to see how they influence them.

To achieve the clustering models, spending measures are made by classifying the users according to the assessment they made of the trip and their gender. In addition, variables such as the travelers' country of origin combined with the different daily expenses and the gross domestic product of the country of origin will be used. The results of the commented models are then described by disaggregating the results according to the type of expenditure. The segments of most interest for public sector managers are shown below.

4.1 Accommodation spending

Figure 5 shows the average expenditure on accommodation related to trip motive, whose coding from 1 to 12 corresponds to the collection in the log design provided.

There are two interesting groups: one with high average accommodation costs and motivation both for work and personal (red group), and the other with low average accommodation costs with a work-only motivation (lower left corner).

4.2 Leisure spending

Within the segmentation made for average expenditure on leisure (Fig. 6), a combinatorial model of the variable of average expenditure on leisure is created according to the main motive for the trip. It can be observed that for trips with leisure motivation (7–13) the average expenditure in this aspect is between medium and high.

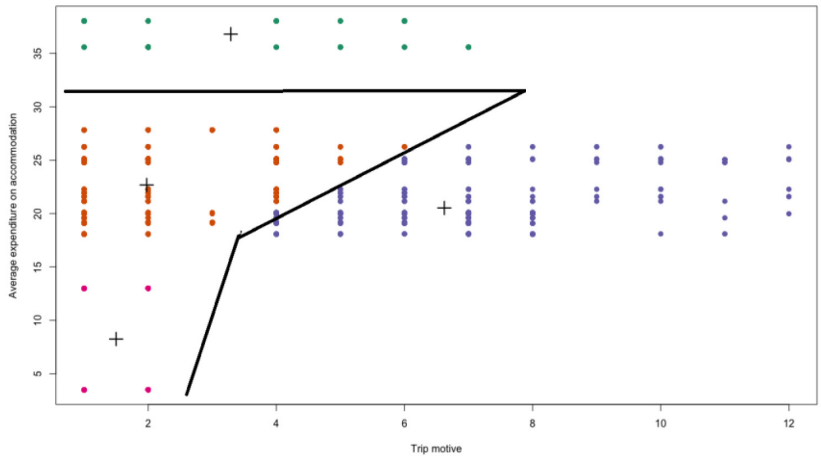

Figure 5: Accommodation spending optimal clustering

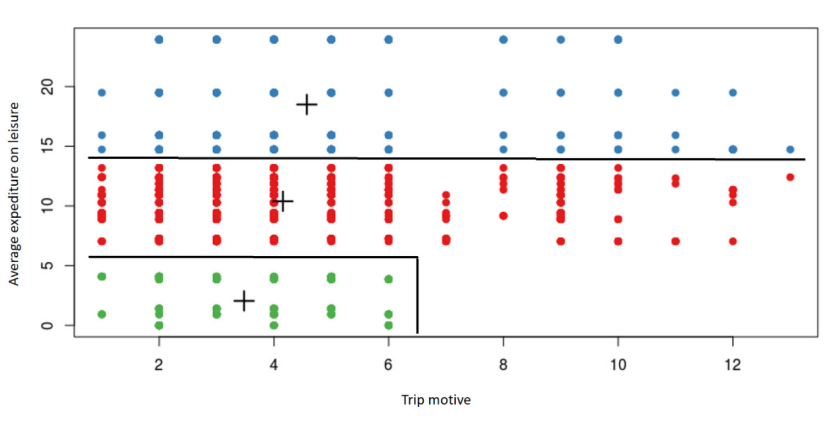

Figure 6: Leisure spending optimal clustering

4.3 Transport spending

Within the combination of the average cost in transport with the main motive of the trip, we can observe (Fig. 7) that three groups are created. For trips with leisure motivation, we have high expenses in transport, while for personal or work travel we have both low and high expenses.

4.4 Total spending (including Open Data sources)

This segmentation model has taken into account open sources of big data, more specifically the indicators from the World Bank, to extract the gross domestic product of the countries included in the sample.

The countries of origin are combined with the gross domestic product per capita (GDP) of the country of origin and the number of days that the trip lasts in addition to the total daily expenditure (Fig. 8).

Figure 8 shows the GDP per capita of the country of origin and the total expenditure per person per day. Four distinct groups are clearly seen in low income and low spending,

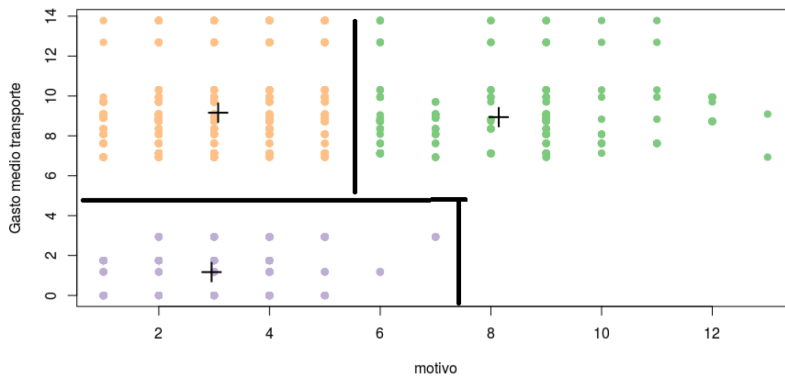

Figure 7: Transport spending optimal clustering

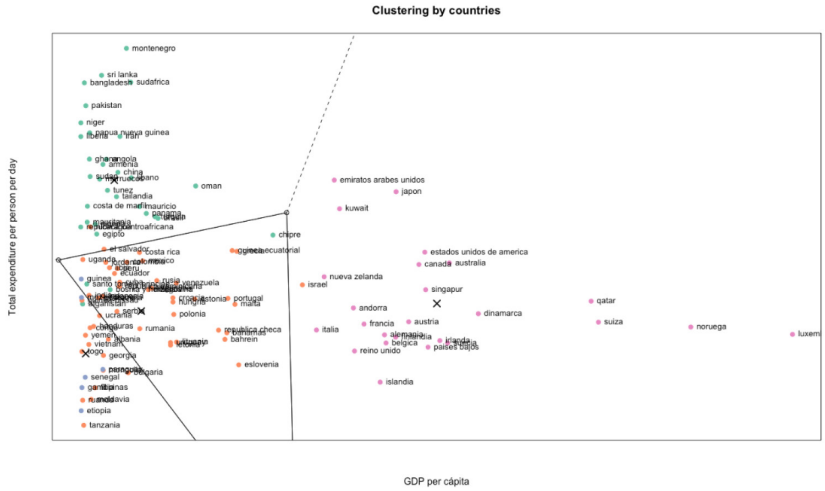

Figure 8: Total spending including GDP information

low-middle-income and low-middle-spending, as well as low-middle-income and high-spending, and high income and medium-high spending.

5 CONCLUSIONS AND FUTURE PROPOSALS

Segmentation is strictly and formally included within purely descriptive analytical models. The incorporation of variations that look for the optimum k number of groups in each case means the segments have a greater strategic value because they generate groups based on criteria of maximum differences.

On the other hand, the article highlights the importance of incorporating Open Data sources whenever possible. Some data are not collected by the surveys, but are inherent to the traveller, as is the GDP of the countries of origin.

A future line of research would be to continue to make segmentations of special strategic value for public tourism managers, based on other important parameters of travel, such as: the type of reservation and the typology of the establishment.

The segmentations performed show the optimal number of categories of users, according to their typologies and spending preferences. These segmentations can be used as a criterion of discretization of numerical variables to apply predictive classification models capable of handling only categorical variables as antecedents, such as algorithms from the ID3 or RBS family.

In addition, as a result of the challenges posed by e-Tourism [9], it is planned to carry out in-depth studies on how to implement, through Big Data techniques, predictive models on preferences in destinations and types of travel.

ACKNOWLEDGEMENTS

The authors are grateful to the support received from:

- Collaboration Agreement between the Valencian Agency of Tourism and the University Miguel Hernández de Elche, for the Promotion of Research and Innovation in the Valencian Tourist Sector (2016).
- Conselleria de Educación, Generalitat Valenciana Grant GVA/2016/053.

REFERENCES

[1] Solsona, J., *Tourism development in rural space, situation analysis and prospective. Study applied to the case of the Region of Valencia* (Doctoral thesis). Ed. Universitat Jaume I de Castellón, 2010.

[2] Tour Spain, Egatur Statistics, available at: http://estadisticas.tourspain.es, Accessed on: 20 March, 2017.

[3] Instituto Nacional de Estadística. http://www.ine.es/, Accessed on: 20 March, 2017.

[4] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, pp. 95–104, 1995.

[5] Hand, D., Mannila, H. & Smyth, P., *Principles of data mining*, Cambridge, MA: The MIT Press, pp. 84–102, 2001.

[6] Wasilewska, A & Menasalvas, E., Data preprocessing and data mining as generalization. Data mining: Foundations and practice. *Studies in Computational Intelligence*, **118**, pp. 469–484, 2008.
https://doi.org/10.1007/978-3-540-78488-3_27

[7] MacQueen, J.B., Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297, 1967.

[8] K Means, Based on a handout. Andrew Ng & Jordan, M. Standford University. available at: http://stanford.edu/~cpiech/cs221/handouts/kmeans.html. (Accessed on 15 March 2017.)

[9] Buhalis, D. & Law, R., Progress in information technology and tourism management: 20 years on and 10 years after the Internet the state of eTourism research. *Tourism Management*, **29**(4), pp. 609–623, 2008.
https://doi.org/10.1016/j.tourman.2008.01.005