Bagged Clustering and its application to tourism market segmentation

Pierpaolo D'Urso[a], Livia De Giovanni[b], Marta Disegna[c,*], Riccardo Massari[a]

[a]*La Sapienza, Roma, Italy*
[b]*LUISS, Roma, Italy*
[c]*School of Economics and Management, Free University of Bolzano, Italy*

Abstract

Aim of the paper is to propose a segmentation technique based on the Bagged Clustering (BC) method. In the partitioning step of the BC method, B bootstrap samples with replacement are generated by drawing from the original sample. The Fuzzy C–Medoids Clustering (FCMdC) method is run on each bootstrap sample, obtaining $(B \times C)$ medoids and the membership degrees of each unit to the different clusters. The second step consists in running a hierarchical clustering algorithm on the $(B \times C)$ medoids. The best partition of the medoids is obtained investigating properly the dendrogram. Then each unit is assigned to each cluster based on the membership degrees observed in the partitioning step. The effectiveness of the suggested procedure has been shown analyzing a suggestive tourism segmentation problem. We analyze two sample of tourists, each one attending a different cultural attraction, enlightening differences among clusters in socio-economic characteristics and in the motivational reasons behind visit behavior.

Keywords: Bagged Clustering, Fuzzy C–medoids, Dissimilarity measures for quantitative and qualitative data, Tourism market segmentation, Normalized weighted Shannon entropy.

*Corresponding author
Email addresses: pierpaolo.durso@uniroma1.it (Pierpaolo D'Urso), ldegiovanni@luiss.it (Livia De Giovanni), marta.disegna@unibz.it (Marta Disegna), riccardo.massari@uniroma1.it (Riccardo Massari)

1 Introduction

For a long time visitors of cultural attractions were treated as a homogeneous group of people. The tendency of the recent tourism literature is instead to consider them as heterogeneous groups with different characteristics, perceptions and needs (Hughes, 2002). Brida et al. (2012) showed that visitors of Christmas Markets in Northern Italy clustered into three groups, according to a set of motivational factors that affect the visit behavior. Other studies showed that tourists who visited art museums presented different socio-demographic characteristics (in particular regarding the level of education, income and occupation) than those who engaged in festivals, musical activities, theme parks, amusements parks, local fairs, and events (Bennett & Council, 1994; Kim et al., 2007; Schuster, 1991).

Market segmentation is a process used in order to discover homogeneous subgroups in the market, according to specific characteristics of customers. In tourism market segmentation, tourists grouped in the same segment are similar to each other (and different from those in other segments) in the way they react to internal stimulus, as desires or emotions, and/or external stimulus, as promotions or advertising.

Understanding the characteristics and the behavior of tourists can be crucial for marketing success. This information allows the marketers to direct marketing efforts toward the groups of tourists more economically significant, improving the overall survival and profitability of cultural attractions, businesses, firms, and destinations in a market that is more and more competitive.

Since the introduction of market segmentation in the late 1950s, the number and type of approaches for segmentation has grown enormously (Dolnicar & Leisch, 2004; Liao et al., 2012). Boone & Roehm (2002) pointed out that there are over 50 methods that can be applied to deal with market segmentation problems. Since each method conducts a multivariate description of the data, grouping units based on their similarity, "different methods present different views of data" (Leisch, 2006). Unfortunately, as emphasized by many researchers, no absolutely "correct" method to segment exists in the literature (Beane & Ennis, 1987; Dolnicar et al., 2008; Kotler et al., 2010; Tkaczynski & Rundle-Thiele, 2011), since the underlying relationships among units have different structures, depending on data at hand, and the researcher must find the best segmentation method to capture this hidden structure. In addition, the researcher intervenes in different moments of the estimation process, "creating" an ever increasing number of new segmentation methods and giving subjective interpretations of the final results.

Segmentation techniques can be classified into two groups, namely supervised and unsupervised classification techniques. Supervision means that "membership of data points which can illustrate the general structure of the group is required in order to derive the classification rules" (Budayan et al., 2009). Unsupervision

implies that there is no rule for the initiation of classification and that the empirical distribution and characteristics of the data will determine the segments' membership.

Cluster analysis methods represent the most used unsupervised market segmentation techniques in the literature and comprise a set of different techniques, which can be broadly divided into partitioning and hierarchical methods (Saarenvirta, 1998). Given a set of selected segmentation variables, these methodologies aggregate units in groups, in such a way that each group contains the most similar units and, at the same time, is the most dissimilar from the remainder groups.

Beyond more traditional methods, non-linear techniques, such as Neural Networks (NNs) algorithms and Kohonen Map (Kohonen, 1989), or Self–Organizing Map (SOM), have also been used in tourism research. Mazanec (1992) is one of the first scholars to use NNs, applying this technique to a market segmentation analysis of Austrian tourists in the "Euro-Sports Region". Dolnicar (1997) used the SOM to identify the characteristics of summer tourists visiting Austria. The latter method was also used, for example: to identify strategic groups of UK hotels (Curry et al., 2001); to segment senior travelers in Western Australia (Kim et al., 2003); to segment the international tourist market in Cape Town, South Africa (Bloom, 2005); to segment the visitors of a particular cultural attraction in the Northern Italy, the Christmas Markets (Brida et al., 2012).

More recently, the Bagged Clustering (BC) algorithm, based on the Bagging ("bootstrap aggregating") procedure (Breiman, 1996), has been introduced in the tourism market segmentation (Dolnicar & Leisch, 2000, 2003, 2004; Leisch, 1999).

BC combines sequentially partitioning and hierarchical clustering methods, to overcome some limitations of both the two procedures. In its initial formulation, first a partitioning method, namely the classic k–means algorithm, is applied to B bootstrap samples generated from the data set. Then a hierarchical method is applied to the results of the partitioning steps. This procedure presents two main advantages, with respect to more traditional clustering techniques (Leisch, 1999): i) it is not necessary to impose the number of clusters in advance; ii) the final solution is less dependent on the initialization of the algorithm.

The aim of this paper is to propose a novel segmentation technique based on the BC algorithm. The main difference is that in the partitioning step the Fuzzy C-Medoids Clustering (FCMdC) algorithm (Krishnapuram et al., 2001) is adopted. FCMdC inherits both the benefits of the partitioning around medoids-based clustering approach and the flexibility and other benefits of the fuzzy approach (see, e.g., D'Urso et al., 2013). By means of this approach, units are classified in homogeneous classes characterized by prototypal observed units (the medoids), which synthesize the structural information of each cluster. Units are assigned to different clusters with fuzzy membership degrees, representing an uncertainty measure

in the assignment process. Conversely to crisp clustering in which the membership degrees can assume values 0 or 1, in the fuzzy clustering the membership degrees assume values between 0 and 1. This approach has the advantage to allow capturing the vague (fuzzy) behavior of particular units. This is reasonable in the market segmentation, when some customers may share some characteristics to more than one segment and, hence, assigning one of them to only one cluster entail a loss of information.

In order to show the effectiveness of the procedure, an empirical analysis on tourism data is finally provided and discussed. The analysis is carried out on two different surveys. The first considers tourists that visit the Museum of Modern and Contemporary Art of Trento and Rovereto, Italy, during the summer season of 2011. In the second survey, Italian visitors of the Christmas Market held in Merano, Italy, in 2011 have been interviewed. Both surveys collected a set of socio-economic characteristics of the tourists and information about the trip. In addition, questions aimed at detecting the motivational factors affecting their visit behavior have been submitted to respondents. In both applications, the segmentation variables are the items regarding the motivational factors, while the socio-economic characteristics and other additional information serve for the ex-post analysis of the obtained clusters.

The paper is organized as follows: in Section 2 we first proceed by overviewing the clustering technique proposed; in Section 3 samples and questionnaires employed in the empirical application are presented, and the clustering results are discussed; in Section 4 final considerations and remarks for future researches are discussed.

2 Methodology

Two main issues in market segmentation based on cluster analysis are deserved to be mentioned. First, the detection of the appropriate number of market segments (clusters) in the dataset and, secondly, the allocation of customers to these clusters, assessing the accuracy of cluster assignments for each unit.

The Bagging ("bootstrap aggregating") procedure (Breiman, 1996) is a resampling method applied in the field of supervised and unsupervised learning to improve the accuracy of prediction. Based on this approach, Leisch (1999) proposed the Bagged Clustering (BC) method, which combines partitioning and hierarchical clustering procedures to deal with the two methodological issues mentioned above. A partitioning method, the classic *k*–means algorithm, is applied to bootstrap samples drawn from the original dataset and then a hierarchical clustering is performed on the resulting cluster centroids. Then, each unit is assigned to the cluster in which falls the closest centroid. Leisch (1999) heuristically showed that

this procedure outperforms existing partitioning method.

In this paper, we propose a Bagged Clustering method, by adopting a fuzzy approach in the partitioning step. In particular, we consider the Fuzzy *C*–Medoids Clustering Algorithm (FCMdC) (Krishnapuram et al., 2001).

Let $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N\}$ be a set of N units (data matrix) and let indicate with $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_i, \cdots, \tilde{\mathbf{x}}_C\}$ a sub-set of $\mathbf{X}$ with cardinality C. $\tilde{\mathbf{X}}$ is the set of the medoids. The medoids can be seen as prototypal units belonging to the considered data matrix, which synthesize the structural information of each cluster.

The FCMdC clustering method can be formalized as follows:

$$\begin{cases} \min: & \sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic}^{p} d_{ic}^{2} \\ \sum_{c=1}^{C} u_{ic} & = 1, \qquad u_{ic} \geq 0, \end{cases} \tag{1}$$

where u_{ic} denotes the membership degree of the i-th unit to the c-th cluster, d_{ic} indicates the distance measure between the i-th unit and the medoid of the c-th cluster and $p > 1$ is a parameter that controls the fuzziness of the partition. The fuzziness parameter should be chosen in advance. Values too close to 1 will result in a partition with all memberships close to 0 or 1. Large values will lead to membership degrees close to $1/C$ (Wedel & Steenkamp, 1989). In the case of FCMdC, a value between 1 and 1.5 is recommended (Kamdar & Joshi, 2000). In all the applications described in section 3 we set $p = 1.5$.

Solving the constrained optimization problem (1) by means of the Lagrangian multiplier method (Krishnapuram et al., 2001) the local optimal solutions are:

$$u_{ic} = \frac{1}{\sum_{c'=1}^{C} \left(\frac{d_{ic}}{d_{ic'}}\right)^{\frac{2}{p-1}}}. \tag{2}$$

Fuzzy clustering approach presents many advantages with respect to standard clustering (Hwang et al., 2007). First, the detected groups in data could overlap, allowing units to belong partially to multiple clusters. In real–world situations, it is often difficult to draw a clear–cut boundary between clusters. McBratney & Moore (1985) observed that the "soft" classification of fuzzy clustering could be more suitable than the deterministic classification of non–overlapping clustering methods such as *k*–means. Finally Heiser & Groenen (1997) showed that fuzzy clustering is less affected by local optima problems.

In addition, the partitioning around medoids strategy (Kaufman & Rousseeuw, 2005) allows to identify prototypal units that summarize the main features of each clusters, instead of "virtual" units such as centroids.

The clustering procedure could be summarized as follows.

1. Construct B boostrap samples of N units, $\mathbf{X}^1,\cdots,\mathbf{X}^b,\cdots,\mathbf{X}^B$, where $\mathbf{X}^b$ is a data matrix obtained by drawing with replacement from the original data matrix $\mathbf{X}$.

2. Run the FCMdC algorithm using an appropriate distance measure, on each bootstrap sample.

   From this procedure we obtain $(B\times C)$ medoids: $\{\tilde{\mathbf{x}}_1^1,\cdots,\tilde{\mathbf{x}}_c^1,\cdots,\tilde{\mathbf{x}}_C^1\},\cdots,$ $\{\tilde{\mathbf{x}}_1^b,\cdots,\tilde{\mathbf{x}}_c^b,\cdots,\tilde{\mathbf{x}}_C^b\},\cdots,\{\tilde{\mathbf{x}}_1^B,\cdots,\tilde{\mathbf{x}}_c^B,\cdots,\tilde{\mathbf{x}}_C^B\},,$

   where C is the number of medoids detected in the partitioning clustering method and $\tilde{\mathbf{x}}_c^b$ is the c-th medoid of the b-th boostrap sample $\mathbf{X}^b$ $(b=1,\cdots,B;\, c=1,\cdots,C)$.

3. Arrange all the medoids in a new dataset $\mathbf{M}_{B\times C}$.

4. Run a hierarchical cluster algorithm on $\mathbf{M}_{B\times C}$, in order to produce a family of partitions of the medoids. The result is represented with a dendrogram and the best partition of H final clusters is obtained investigating the graphic.

5. The membership degree of unit i to each final cluster $h\,(h=1,\ldots,H)$ is obtained selecting the maximum membership degree of the unit to all the medoids in the cluster. Let $\tilde{\mathbf{x}}_{1[h]},\ldots,\tilde{\mathbf{x}}_{C^h[h]}$ be the C^h medoids classified in the h-th final cluster $(\sum_{h=1}^{H}C^h=B\times C)$, and let $u_{i1[h]},\ldots,u_{iC^h[h]}$ be the membership degrees of the unit i to the C^h medoids. Then the membership degree of the i-th unit to the h-th final cluster is defined as $\hat{u}_{ih}=\max\{u_{i1[h]},\ldots,u_{iC^h[h]}\}$, $h=1,\ldots,H$.

Figure 1 schematically shows the steps of the algorithm.

Remark 1. Distance measures.
The distance measure adopted at both step of the clustering procedure is the same, and it depends on the type of variables.

Let $\mathbf{X}=\{\mathbf{x}_1,\cdots,\mathbf{x}_i,\cdots\mathbf{x}_N\}$ be the data matrix, where $\mathbf{x}_i=(x_{i1},\cdots,x_{ik},\cdots,x_{iK})$ represents the value observed for the i-th unit on the k-th variable $(i=1,\cdots,N;\, k=1,\cdots,K)$.

When the variable are continuous or discrete, one can use the Euclidean distance between units $\mathbf{x}_i$ and $\mathbf{x}_{i'}$:

$$d_{ii'}=\|\mathbf{x}_i-\mathbf{x}_{i'}\|=\left[\sum_{k=1}^{K}(x_{ik}-x_{i'k})^2\right]^{\frac{1}{2}} \tag{3}$$

or, more in general, a distance from the Minkowski class.

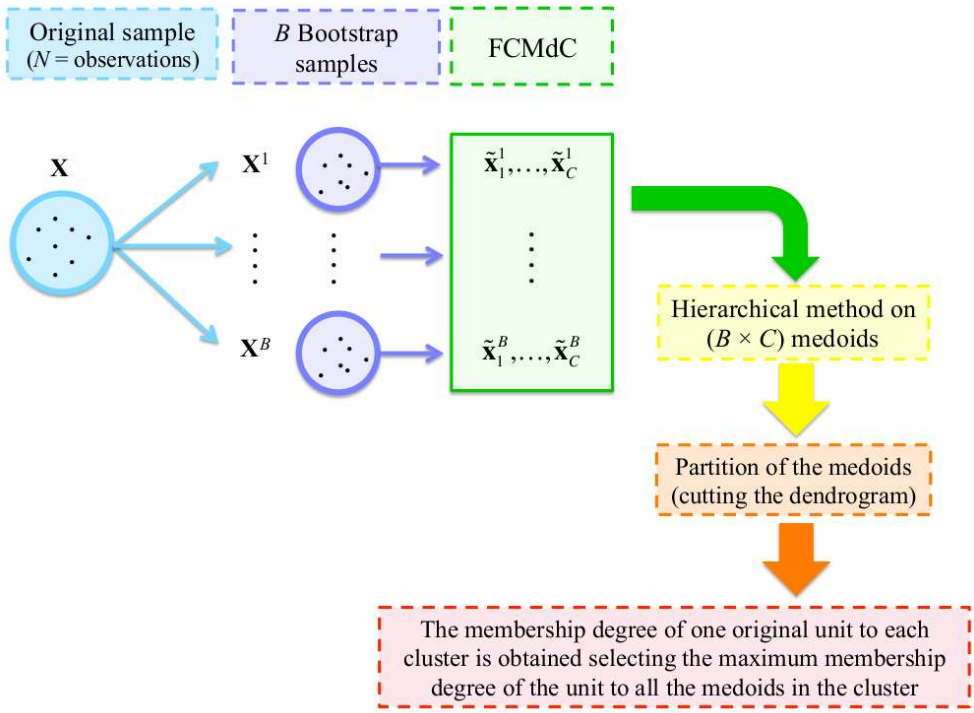

Figure 1: The Bagged Clustering algorithm.

When dealing with binary variables, assuming values equal to 1 or 0, it is possible to make use of the Jaccard dissimilarity index between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$:

$$d_{ii'} = 1 - \frac{a}{a+b+c}, \tag{4}$$

where a is the total number of variables being 1 for both unit $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, b is the total number of variables being 1 for $\mathbf{x}_i$ and 0 for $\mathbf{x}_{i'}$, and c the total number of variables being 0 for $\mathbf{x}_i$ and 1 for $\mathbf{x}_{i'}$.

When the variables are ordinal (e.g., questions detected using Likert scale), the items are often represented by linguistic expressions. Coppi & D'Urso (2002) observed that subjective evaluations are best represented in a fuzzy framework, reflecting the uncertainty and the heterogeneity in individual evaluation (see also Hung & Yang, 2005). Hence, to take into account the vagueness of the variables, the values can be fuzzified (D'Urso, 2007) and the Yang–Ko distance (Yang & Ko, 1996) for fuzzy data can be adopted.

Let define the fuzzy data matrix as follows:

$$\mathbf{X} \equiv \{x_{ik} = (c_{ik}, l_{ik}, r_{ik}) : i = 1, \cdots, N, k = 1, \cdots = K\}, \tag{5}$$

where $x_{ik} = (c_{ik}, l_{ik}, r_{ik})$ represents the k-th fuzzy variable observed on the i-th unit, c_{ik} denotes the center, l_{ik} and r_{ik} the left and right spread respectively (D'Urso, 2007), with the following membership function

$$\mu_{x_{ik}}(u_{ik}) = \begin{cases} L\left(\frac{c_{ik}-u_{ik}}{l_{ik}}\right) & u_{ik} \leq c_{ij}\,(l_{ik} > 0) \\ R\left(\frac{u_{ik}-c_{ik}}{r_{ik}}\right) & u_{ik} \geq c_{ik}\,(r_{ik} > 0) \end{cases} \tag{6}$$

where L (and R) is a decreasing "shape" function from $\Re^+$ to $[0,1]$ with $L(0) = 1$; $L(z_{ik}) < 1$ for all $z_{ik} > 0, \forall i, j$; $L(z_{ik}) > 0$ for all $z_{ik} < 1, \forall i, j$; $L(1) = 0$ (or $L(z_{ik}) > 0$ for all z_{ik} and $L(+\infty) = 0$). The fuzzy number $x_{ik} = (c_{ik}, l_{ik}, r_{ik})_{LR}$, $i = 1, \ldots, I$; $j = 1, \ldots, J$, consists of an interval which runs from $c_{ik} - l_{ik}$ to $c_{ik} + r_{ik}$ and the membership functions serve to assign differenti weights to the values in the interval.

If L and R are of the form:

$$L(z) = R(z) = \begin{cases} 1 - z & 0 \leq z \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

then $\mathbf{X}$ is a "triangular" fuzzy data matrix with the following membership function:

$$\mu_{x_{ik}}(u_{ik}) = \begin{cases} 1 - \frac{c_{ik}-u_{ik}}{l_{ik}} & u_{ik} \leq c_{ik}\,(l_{ik} > 0) \\ 1 - \frac{u_{ik}-c_{ik}}{r_{ik}} & u_{ik} \geq c_{ik}\,(r_{ik} > 0). \end{cases} \tag{8}$$

For each component of the fuzzy data matrix (5) we can define the following data matrices:

$$\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_i, \cdots \mathbf{c}_N\}$$
$$\mathbf{L} = \{\mathbf{l}_1, \cdots, \mathbf{l}_i, \cdots \mathbf{l}_N\}$$
$$\mathbf{R} = \{\mathbf{r}_1, \cdots, \mathbf{r}_i, \cdots \mathbf{r}_N\}$$

where $\mathbf{c}_i, \mathbf{l}_i$ and $\mathbf{r}_i$ are the vectors of the i-th observations for the centers, the left and the right spreads respectively.

Then, in a multivariate framework the Yang–Ko distance can be formulated as follows (D'Urso, 2007):

$$d_{ii'} = (\|\mathbf{c}_i - \mathbf{c}_{i'}\|^2 + \|(\mathbf{c}_i - \alpha\mathbf{l}_i) - (\mathbf{c}_{i'} - \alpha\mathbf{l}_{i'})\|^2 + \|(\mathbf{c}_i + \beta\mathbf{r}_i) - (\mathbf{c}_{i'} + \beta\mathbf{r}_{i'})\|^2)^{\frac{1}{2}} \quad (9)$$

where the parameters α and β reduces the weight of the spreads with respect to the centers (D'Urso, 2007). Given the triangular membership function (8), in this work we assume $\alpha = \beta = 0.5$.

Remark 2. Membership degrees.
As described above, the membership degree $\hat{u}_{ih}$ of the unit i to the final cluster $h\,(h = 1, \cdots, H)$, obtained with the hierarchical procedure, is derived considering the maximum membership degree of i to all the medoids that are classified in cluster h. If $\tilde{i}$ is a medoid classified in cluster h, then $\hat{u}_{\tilde{i}h} = 1$. However, $\hat{u}_{\tilde{i}h'}\,(h' \neq h)$ is not constrained to be equal to zero. More generally, the sum of the membership degrees of an unit to all H clusters is not necessarily equal to one.

Remark 3. Clusters evaluation.
Finally it is important to evaluate the internal variability of each variable in each cluster. When using quantitative or variables, the box-plot can be used in order to analyze more in depth the distribution of each variable. Otherwise, when considering binary or categorical variables, an entropy index can be adopted.

Hence, we make use of the weighted Shannon Entropy index, normalized to vary in the range $[0, 1]$.

Let consider a categorical variable $\mathbf{x} = \{v_1, \ldots, v_m, \ldots, v_M\}$ which takes $M\,(M \geq 2)$ values. In particular, when the variable is binary $M = 2$. The weighted frequency of the m-th category of $\mathbf{x}$ is:

$$wf_m = \frac{\sum_{i=1}^{N} \mathrm{I}(m)_i w_i}{\sum_{i=1}^{N} w_i} \quad (10)$$

where w_i is the sample weight of the i-th unit and $\mathrm{I}(m)_i$ is an indicator function which is equal to 1 if the i-th unit assumes the m-th category of $\mathbf{x}$ and 0 otherwise. We could write:

$$\mathrm{I}(m)_i w_i = w_{mi}$$

where $w_{mi} = w_i$ if $\mathbf{x}_i = v_m$ and 0 otherwise. Then (10) becomes:

$$wf_m = \frac{\sum_{i=1}^{N} w_{mi}}{\sum_{i=1}^{N} w_i} \tag{11}$$

The normalized weighted Shannon Entropy index (WS^*) can be written as:

$$WS^* = \frac{-\sum_{m=1}^{M} wf_m \log(wf_m)}{\log(M)} \tag{12}$$

In our context, we make use of the membership degrees $\hat{u}_{ih}$ as weights in the formulae (11)-(12). In particular, with reference to the final hierarchical partition, $w_{imh} = \hat{u}_{ih}$, i.e. the membership degree of the i-th unit to the h-th cluster is considered as the sample weight of the unit i which takes the m-th category of $\mathbf{x}$.

Then the expressions (11) and (12) become respectively:

$$wf_{mh} = \frac{\sum_{i=1}^{N} w_{mih}}{\sum_{i=1}^{N} w_{ih}} \tag{13}$$

$$WS^*_h = \frac{-\sum_{m=1}^{M} wf_{mh} \log(wf_{mh})}{\log(M)}. \tag{14}$$

Note that in this way we implicitly take into account the fuzziness of the partition, incorporating the vagueness of the assignment of each unit to a cluster in the computation of indices (13)-(14).

3 Tourism market segmentation: a case study

In this section we provide an application of the proposed clustering method in the field of tourism market. Profiling visitors can be of crucial importance for local policymakers, managers and marketing analysts. Identifying homogeneous

clusters of consumers-visitors can be in fact an essential step both for planning and developing appropriate strategies and for directing the political and economical actions towards the groups economically more relevant.

Two datasets drawn from surveys conducted among the visitors of two different cultural attractions held in the Trentino-South Tyrol region, Northern Italy, are considered in the following empirical analysis. The objective was to find "motivations clusters", that is homogeneous groups of visitors from the point of view of their motivations in visiting the particular cultural attraction analyzed.

The first survey regards the visitors of the Museum of Modern and Contemporary Art (shortened to MART) of Trento and Rovereto, Italy, the two main cities of the Trentino region (Figure 2). A total of 591 interviews were successfully collected during the summer season (from June to September) of 2011. Most visitors come from Italy (92.72%), mainly from neighboring regions, covering on average about 260 Km (standard deviation 690 Km): 29.78% from Veneto, 29.10% from Trentino-South Tyrol (of which the majority – 79.07% – come from the province of Trento), and 11.51% from Lombardia.

The MART museum is divided into three buildings. The main building (where the interviews were being held) is located in Rovereto, the hometown of the futurist artist Fortunato Depero. This main building was designed by the Swiss architect Mario Botta in the late 1980s. The museum hosts a permanent collection of modern art, where works are displayed on a rotating basis, and temporary exhibitions. It holds the most important collections in Italy concerning different artistic genres of modern and contemporary art, in particular futurism. Interviews were held just after the visit of the museum, in selected working and weekend days of the four months considered, and in different time periods of the day.

The second survey was conducted among Italian visitors of one of the most important Christmas Market (CM) of the South Tyrol region, held in Merano, Italy (Figure 2). CMs have a very long tradition and draw their origins from German speaking countries. The first and foremost CM was held in Berlin in the 18th century and its primary function was to create a place where families could purchase children's presents. Soon CMs became a symbol of the German culture and the South Tyrol region, which was part of Austria till 1919 and it is populated even today by a majority of German language, has inherited this tradition. The CM has become more and more important for many cities in the South Tyrol region due to its capacity of attract significant flows of visitors, mainly from Italian region. The first city in Italy that hosted a CM was Bolzano, the main city of the South Tyrol region, in 1990. At the moment also Merano, Bressanone, Brunico, and Vipiteno host this event creating a sort of "circuit" of South Tyrol CM. The stands at the CM sell typical local products including food and beverage, Christmas decorations, small gifts and presents, and local artifacts. Although the term "market" might recall trade and shopping, these events attract flows of visitors

mainly interested in the Christmas atmosphere that they can experience during the visit (Brida et al., 2013).

Figure 2: Map of theTrentino-South Tyrol region, Northern Italy.

An overall number of 797 Italian visitors of the Merano CM were interviewed, both stay–over and same day–visitors, whereas local residents were excluded. The survey was conducted in 2011 during the four weeks of advent (from 30 November to 24 December). The majority (62%) of the visitors were interviewed at the end or during their visit. Interviewees came mainly from neighboring regions of Northern Italy covering on average 340 Km (standard deviation 215 Km) to reach Merano: 25.1% from Lombardia, 19.7% from Veneto, 14.8% from Emilia-Romagna, and 9.73% from Trentino-South Tyrol. Interviews were held in the most visited parts of the CM, mainly (86%) during the week–end days (from Friday to Sunday) of the four weeks. Most of the interviews (52%) were conducted in the afternoon–evening and all of them were collected during period of good weather.

The questionnaires of both survey were anonymous and self-administered, and were written in three languages (Italian, German and English). The two questionnaires have a similar structure and they are divided into three sections: 1. information regarding the cultural attraction visited (repeated visiting, factors that stimulated the visit, authenticity perception, shopping expenditure); 2. information on the trip (the number of nights and the type of accommodation, expenditure

per night for accommodation, expenditure per day for food and drink, motives of the trip); 3. socio-demographic and economic characteristics of the interviewees and their families.

The convenience sampling method (Cochran, 2007) was adopted, as there was no sufficient information on the characteristics of visitors of the museums and of the CM in order to apply a probabilistic design. A research team member was present in order to solve questions or doubts that emerged among interviewees. Interviewers selected only one person per household, or travel party, that passed through a previously selected spot. In order to encourage cooperative behavior, respondents were informed that the research had exclusively scientific aims, and that impartiality in the data analysis was guaranteed. Furthermore, a pilot survey was carried out to test the questionnaire before conducting the full survey, in order to avoid bias related to its structure and wording.

In order to find homogeneous groups of visitors according to their reasons of visiting each of these particular cultural attractions, we carried out two separate application, running the proposed clustering method on data collected from both surveys.

In both applications we consider $C = 10$ medoids and $B = 50$ bootstrap samples, resulting in a total of 500 $(B \times C)$ medoids. Several iterations are performed in order to avoid local optima solutions.

3.1 The case of the MART museum

The questionnaire used in the MART museum asked the respondents if they agreed or not (dichotomous answer) with a set of push factors that motivated the visit. The set of factors included: satisfying a curiosity ("curiosity"), resting/relaxing ("relax"), a specific interest in such an attraction ("interest"), accompanying friend/family member with a specific interest in such an attraction ("friend"), learning something new ("learn"), telling friends about the visit ("tell"), doing something that one ought to do ("do"), contributing to preserving this attraction for future generation ("future"), revisiting this museum ("revisit"), showing the museum to friends or relatives ("show"), professional or academic reasons ("work"), doing something worthwhile ("worthwhile"), occupying some leisure time ("leisure"), visiting temporary exhibition ("temporary"), seeing the building ("building").

Since the segmentation variables are dichotomous, the matrix of centers was hierarchically clustered using the Jaccard dissimilarity index, and Ward's linkage method.

Results are reported in Figure 3.

The top panel of Figure 3 displays the dendrogram derived from the BC procedure. The plot under the dendrogram shows the standardized heights at which

each cluster is aggregated (solid line), and the first differences of these heights (dashed line). Local peaks in the dashed line drive the selection of the number of clusters. Indeed, local peaks correspond to the longest distances between two consecutive aggregations in the dendrogram.

The peaks in the first differences line suggest that the visitors of the MART museum can be divided into two groups, of which one contain only a small part (27.24%) of the total sample of visitors interviewed, or five groups. In order to better understanding particular visitors' behavior, the five clusters solution was taken into consideration in the following analysis.

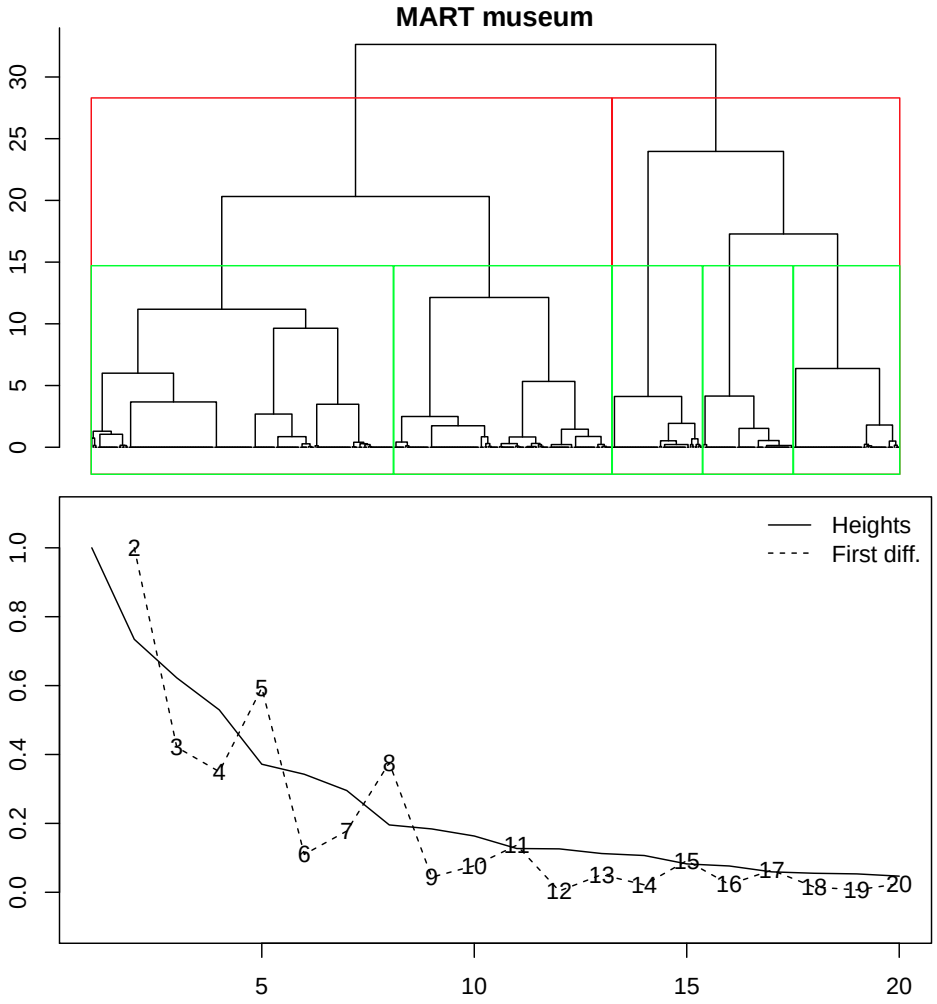

Figure 3: Dendrogram [top panel] and relative heights of aggregation (solid line) and first differences of relative heights (dashed line) [bottom panel].

In order to better analyze and describe the characteristics of the segments, the weighted frequency (13) and the normalized weighted Shannon entropy (14) calculated for each *k*-th segmentation variable and for each *h*-th cluster, are used (see Table 1). The weights used in our study are the degree of membership of the tourist *i* to the *h*-th cluster, derived as explained in Section 2 (see Remarks 2 and 3).

Table 1: Normalized weighted Shannon entropy, weighted frequency (%) of value 1 in brackets.

Variables	CL1 (N=299)		CL2 (N=54)		CL3 (N=57)		CL4 (N=50)		CL5 (N=131)	
curiosity	0.419	(8.49)	0.789	(23.63)	0.988	(56.33)	0.986	(43.04)	1.000	(49.94)
relax	0.293	(5.15)	0.509	(11.32)	0.713	(19.56)	0.999	(51.80)	0.700	(18.94)
interest	0.601	(85.36)	0.968	(60.47)	0.858	(28.20)	0.996	(53.86)	0.834	(73.52)
friend	0.324	(5.91)	0.999	(48.38)	0.518	(11.60)	0.528	(11.96)	0.696	(18.73)
learn	0.652	(16.74)	0.662	(17.19)	0.984	(42.47)	0.665	(17.34)	0.924	(33.90)
tell	0.045	(0.49)	0.092	(1.17)	0.082	(1.02)	0.096	(1.23)	0.113	(1.51)
do	0.578	(13.77)	0.573	(13.57)	0.568	(13.39)	0.480	(10.36)	0.634	(15.98)
future	0.149	(2.13)	0.210	(3.33)	0.227	(3.68)	0.204	(3.19)	0.281	(4.87)
revisit	0.190	(2.91)	0.738	(20.82)	0.269	(4.60)	0.318	(5.77)	0.369	(7.09)
show	0.142	(2.01)	0.477	(10.25)	0.229	(3.72)	0.303	(5.39)	0.243	(4.01)
work	0.179	(2.70)	0.279	(4.84)	0.292	(5.13)	0.306	(5.46)	0.376	(7.28)
worthwhile	0.700	(18.91)	0.604	(14.78)	0.555	(12.91)	0.594	(14.36)	0.873	(29.34)
leisure	0.468	(9.96)	0.436	(8.99)	0.468	(9.96)	0.538	(12.32)	0.517	(11.58)
temporary	0.195	(96.99)	0.361	(93.12)	0.357	(93.23)	0.404	(91.94)	0.414	(91.67)
building	0.422	(8.56)	0.726	(20.19)	0.607	(14.88)	0.668	(17.47)	0.978	(41.33)

Results reported in Table 1 reveal that telling friends about the visit at the museum (“tell”) is homogeneously considered not as an important push factor in any cluster (the normalized Shannon entropy index is near to 0 and the weighted frequencies of value 1 are very low). Therefore this factor does not allow us to characterize the segments.

Visitors of cluster 1, the largest group identified, visited the museum mainly because they are attracted by a specific interest in such an attraction (“interest”) that probably is the temporary exhibition (“temporary”). Therefore, these visitors have been named “Interested in the exhibition”.

Visitors of cluster 2 could be named “Family” since they visited the museum to accompany a friend/family member with a specific interest in such an attraction (“friend”), therefore they want to show the museum to them (“show”), and because they have already seen it but they want to revisit (“revisit”).

Visitors of cluster 3 visited the museum mainly for the sake of satisfying a curiosity (“curiosity”) and learning something new (“learn”). Therefore, these visitors have been named “Knowledge seeker”.

Cluster 4 regards visitors that mainly want to rest/relax (‘relax”) occupying some leisure time (“leisure”) so that these visitors have been named “Not inter-

ested".

Finally, cluster 5 is composed by visitors who made the visit at the museum mainly because they are attracted by a specific interest in such an attraction ("interest") that probably is the building ("building"). Furthermore, they visited MART because it is something one ought to do ("do") and something worthwhile ("worthwhile"), probably because the highest percentage of people who came for professional or academic reasons are concentrated in this cluster ("work"). Therefore, these visitors have been named "Interested in the building".

Applying the "classic" BC procedure proposed by Leisch (1999) to the same dataset, Brida et al. (2012a) identified 3 clusters. Comparing the results obtained using the two methodologies, we can observe that the "Knowledge seeker" segment is identified as a niche segment with both algorithms. By adopting the methodology proposed in this paper, the "Interested" group identified by Brida et al. (2012a) is now divided into two groups, with different motivational needs, the "Interested in the exhibition" and the "Interested in the building". The third segment identified with the classic BC procedure grouped all the people who were "not" motivated by one of the factors proposed. Using the algorithm proposed in this paper, it is possible to split this heterogeneous group into two groups with peculiar characteristics, the "Family" and the "Not interested".

3.2 The case of the Christmas Market

Items concerning motivations of the visit at the CM asked how strongly each respondent agreed with a set of push factors. Each item used a 5-points Likert scale (qualitative answer in which 1 means "not at all" and 5 means "very important"). The considered set of items included "shopping", socialising with friends and relatives ("socialise"), "relax", meeting new people ("meet"), doing something different and original ("do"), bringing my partner/family ("family"), supporting a local community initiative ("support"), tasting local products–food and beverages ("taste"), staying in a unique Christmas atmosphere ("atmosphere"), visiting the town centre in Merano ("town"), "merry-go-round" for children, "train" for children, "pastry shop" for children, "ice-skating" rink, band–musical group ("band").

The segmentation variables were transformed into triangular fuzzy numbers (see Remark 1). As mentioned above, the considered variables express agreement degree that are vague, since they are the results of subjective judgements. Fuzzy set theory captures this vagueness and can suitably measure the imprecision and errors that can be present in the analysis of visitors evaluations (Benítez et al., 2007).

To this end, the 5-points Likert scale is transformed as illustrated in Table 2, where we adopt the recoding proposed by Hung & Yang (2005).

Table 2: 5–points Likert scale values and their corresponding fuzzy numbers (center, left spread, right spread).

Value	**Fuzzy number**		
x	c	l	r
5	1	0.25	0
4	0.75	0.25	0.25
3	0.5	0.25	0.25
2	0.25	0.25	0.25
1	0	0	0.25

The 500 final centres were been hierarchically clustered using the Yang–Ko distance (9) and Complete's linkage method.

Results are displayed in Figures 4 and 5.

Figure 4 is similar to Figure 3 and the peaks suggest that the visitors of the CM can be divided into two groups, one of which is smaller than the other containing only 31.24% of the total Italian visitors interviewed, or four. In order to better understanding particular visitors' behavior, also in this case the "second best" solution, with four cluster, is considered. Cluster 1 is the bigger and grouped 51.57% of the total visitors; cluster 2 and 4 are nearly of the same size (respectively 17.69% and 17.19% of the total visitors); cluster 3 is a niche segment grouping only 13.55% of the whole sample.

The box-plots in Figure 5 allow us to investigate the weighted distribution of the segmentation variables per each cluster identified. As for the case of MART results, the weights are the membership degrees of each tourist to each cluster. These graphs allow a better and easier interpretation of the results. In Figures 6 and 7 are reported the boxplots for the left spread and the right spread, respectively, of the segmentation fuzzy variables, i.e. $\mathbf{m}_k - \mathbf{l}_k, \mathbf{m}_k + \mathbf{r}_k$.

For the sake of interpretation, it is important to emphasize that the higher the height of the box (i.e. the Interquartile Range), the smaller the homogeneity of the segment with respect to the variable considered. This implies that segments are better characterized by those variables presenting low dispersion, and that a strong dispersion within a variable indicates non homogeneity of the tourists of the segment with respect to that characteristic.

Staying in a unique Christmas atmosphere ("atmosphere") was homogeneously recognized as a very important push motivation in visiting the Merano CM, with the exception of cluster 1 that grouped also people who did not consider very important this factor. Other very important factors for the majority of the sample were tasting local food and beverages ("taste") and visiting the town centre of Merano ("town"). Cluster 1 grouped those that were less interested than the others in "relax" and doing something different and original ("do").

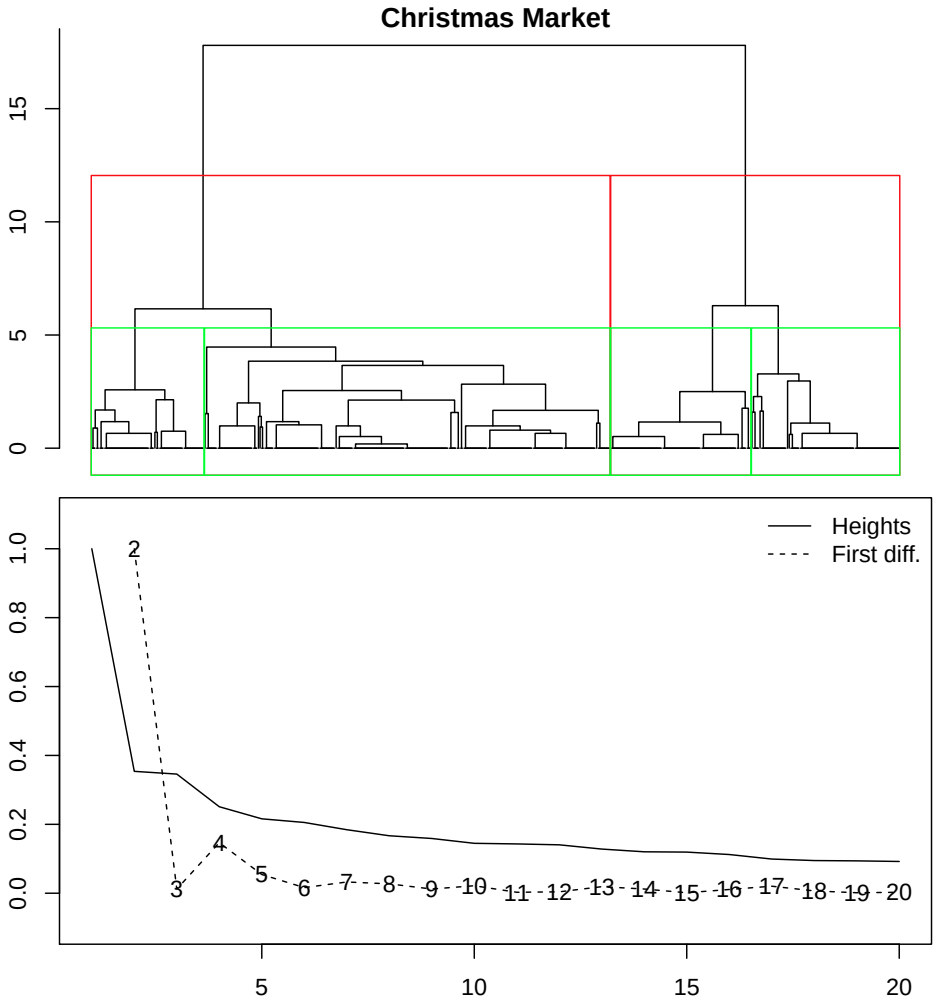

Figure 4: Dendrogram [top panel] and relative heights of aggregation (solid line) and first differences of relative heights (dashed line) [bottom panel].

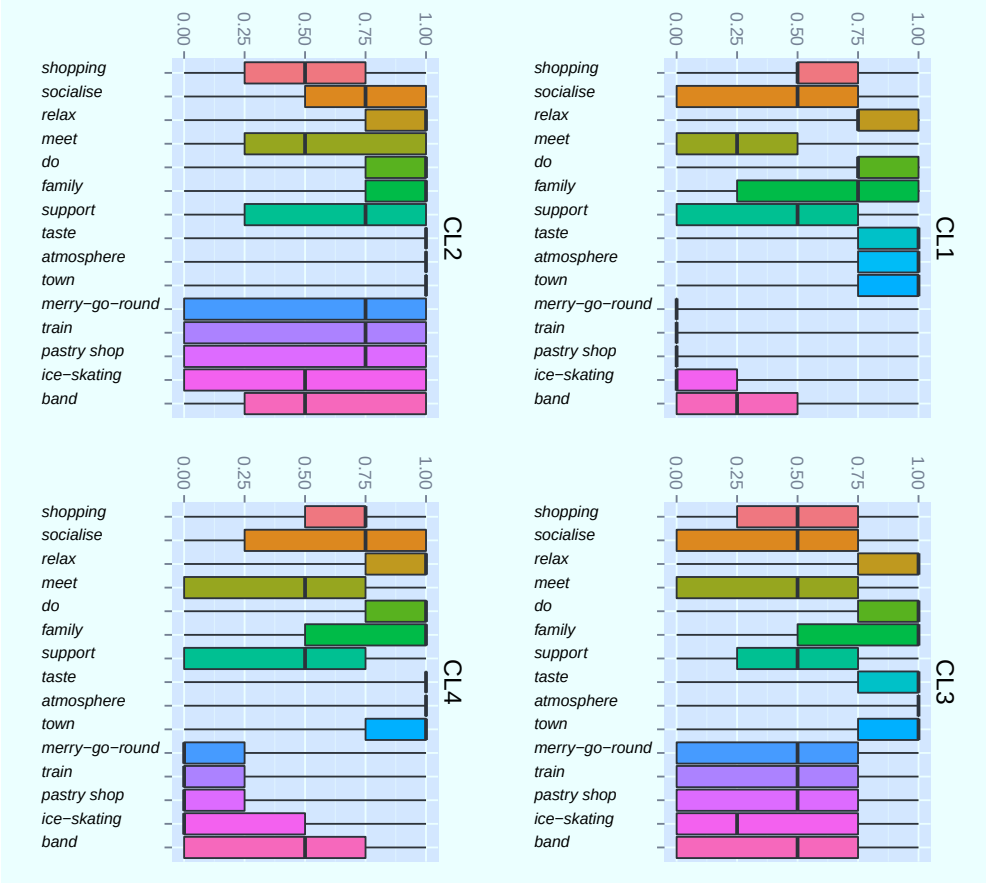

Figure 5: Box-plot for the four solution using the centers of the fuzzy data

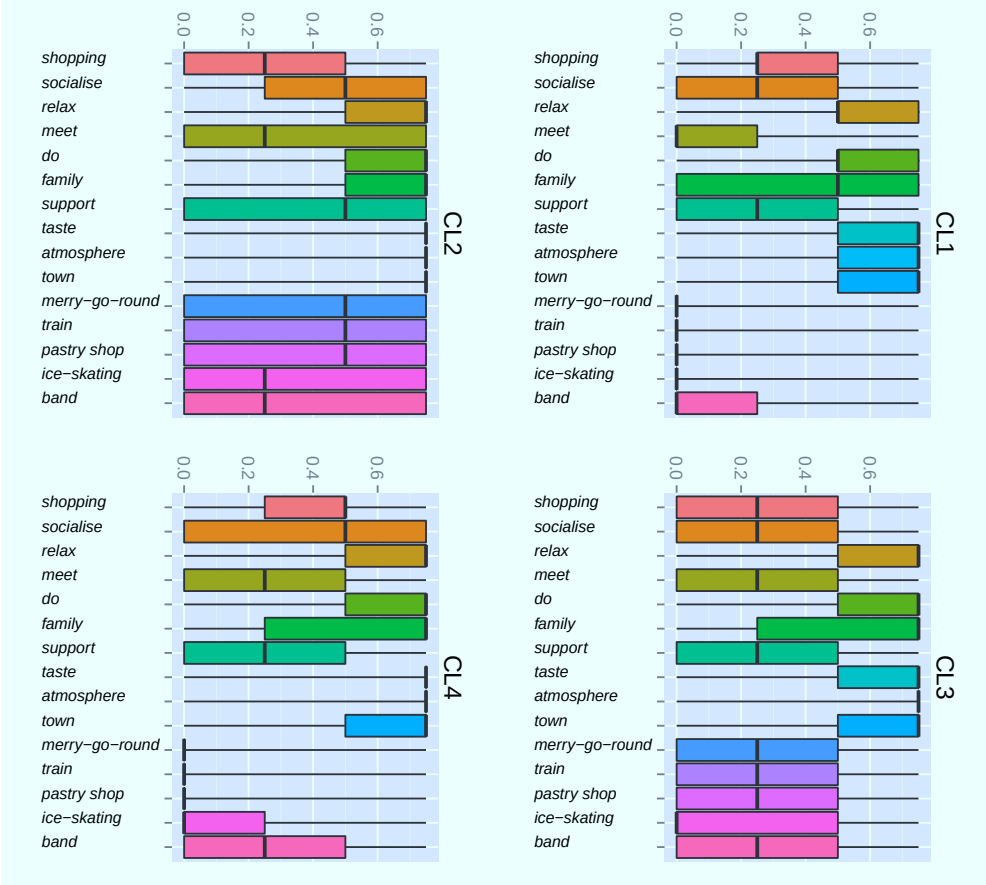

Figure 6: Box-plot for the four solution using the left spread of the fuzzy data

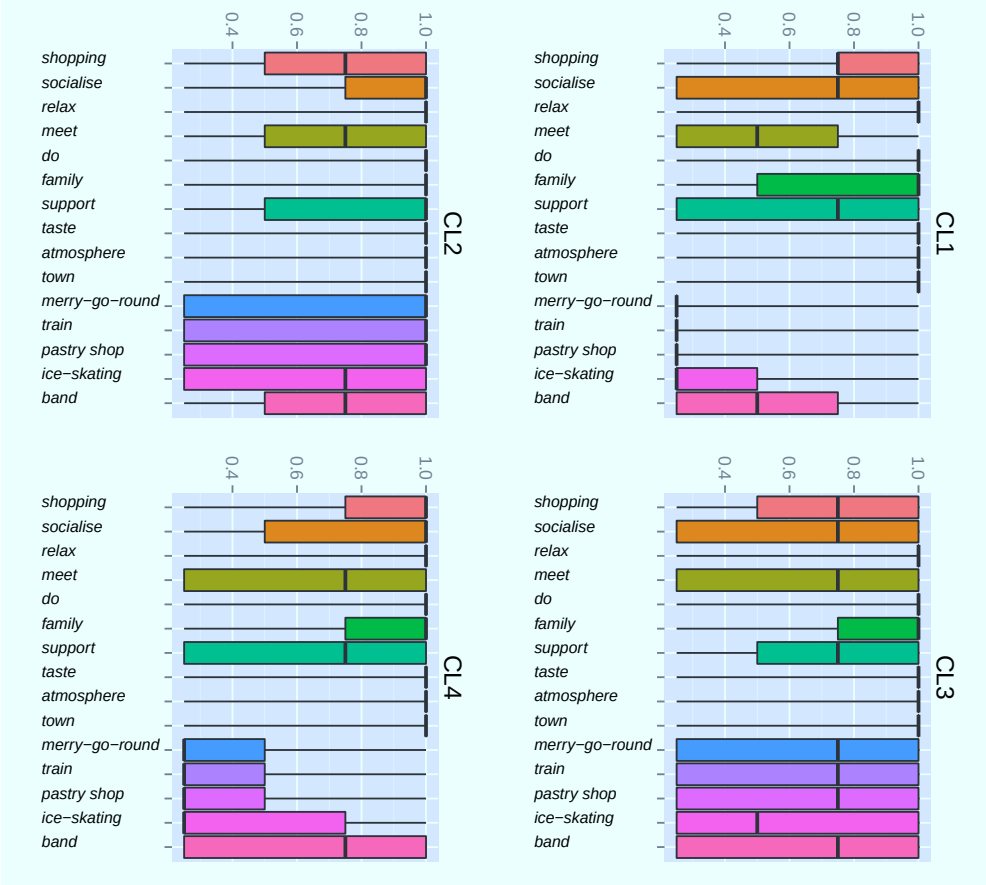

Figure 7: Box-plot for the four solution using the right spread of the fuzzy data

The overall score assigned to the children's attractions, i.e. the "merry–go–round", the "train", and the "pastry shop", allows us to distinguish the clusters into two types: that in which the travel group included children (clusters 2 and 3), and that in which the children were more or less absent (the remainder). As regards to other attractions proposed by the organizers of the CM, i.e. the "ice–skating" and "band", we can see that cluster 2 grouped the visitors more interested in these attractions, while cluster 1 included those who are less interested. Therefore, it is not surprising that visitors of clusters 2 and 3 considered staying with their partner/family ("family") as important when they visit this type of cultural event. Also the members of cluster 4 gave importance to the family motivation, but they probably did the visit only with their partner, due to the low score that they gave to children's attractions.

The rate assigned to supporting a local community initiative ("support") and to the opportunity to "socialise" with friends and relatives are in general very similar and a positive correlation seems to be. Only the visitors of cluster 4 considered socialising as an important factor for the visit unlike supporting a local initiative. Finally, visitors of cluster 4 are the most interested in "shopping", followed by visitors of cluster 1, while visitors of cluster 1 are the less interested in visiting the CM in order to meet new people ("meet").

Summarizing the results, cluster 1 (the bigger) can be named "No CM", because it grouped visitors who paid less attention than others to stay in a unique Christmas atmosphere, relaxing, doing other particular activities during the visit, meeting new people, using the attractions proposed by the organizers of the CM. In fact, it seems that these visitors come for a short and scheduled visit in order to buy some particular products that, maybe, they have already seen before.

Cluster 2, named "Pro activities", is probably mainly composed by families with children enthusiastic in all the activities linked to the Christmas atmosphere and in all the attractions and opportunities for socialisation. Cluster 3 (the niche), named "Basic", is probably composed by families with children interested in visit the CM and the city without a particular interest or disinterest in one activity. Cluster 4, named "Shopping", is mainly composed by couples who visit the CM for shopping, i.e. mainly for a commercial reason.

Applying the classic BC to the same dataset, 6 clusters are identified (see Brida et al., 2012b). Comparing the results obtained using the two methodologies, we can observe that in both cases the sample is mainly split depending on the presence/absence of children into the travel group. The two main segments, the "No CM" and the "Pro activities", identified with the classic BC algorithm are identified also with our procedure. Therefore, we can conclude that the most important and well distinguished segments of visitors of the Merano CM are identified with both methodologies and that, additionally, the procedure proposed in this paper allow us to better outline the remaining group of people identifying two more

characterized groups in place of four.

3.3 Visitors profile

The additional information collected through the surveys were used to characterize the clusters identified in each procedure in terms of socio–demographic (gender, age, level of education, origin, occupation) and economic (household income, total expenditure per person per night, and expenditure for shopping per person) variables. Table 3 reports the complete list of these profiling variables with a brief description of them.

Table 3: Description of variables used in the econometric model.

Independent variables	**Descriptions**
First–time visitors	1 = the interviewee visits for the first–time the cultural attraction; 0 = otherwise.
Very satisfied	1 = the interviewee is overall very satisfied about the visit; 0 = otherwise.
Socio–demographic and economic characteristics	
Male	1= male; 0= female.
Age	Age of the respondent (continuous).
Age^2	Squared age of the respondent (continuous).
University	1 = Education level is university degree or postgraduate; 0 = otherwise.
Married	1 = Married; 0 = otherwise.
Italy	1= Italy; 0 = otherwise.
Km	Distance in kilometres between the city of residence and the city that hosts the cultural attraction visited (continuous).
Employed	1= Autonomous worker and employed (full-time or part-time); 0 = otherwise.
Income	Central value of each income category; 0 = if the respondent does not declare his/her income (continuous).
Missing income	1 = respondent does not declare his/her income; 0 = otherwise.
Expenditure	
Total expenditure	Individual expenditure for accommodation, food and beverage, shopping in the shops of the city, pharmacy, tour guide services, other expenditures linked to the visit (excluding expenditure for transportation) per night, in Euros (continuous).
Shopping at the museum	Individual expenditure at the shop of the museum, in Euros; 0 = respondent does not spend or does not visit the shop (continuous).
Shop not visited	1 = respondent does not visit the shop of the museum; 0 = otherwise.
Shopping at the CM	Individual expenditure at the stands of the CM, in Euros (continuous).

In Tables 4 and 5 are reported the main characteristics of the clusters observed, i.e. the mean values of the continuous variables and the percentages of the categorical variables. Note that income is treated as categorical, recoding its value into four classes, namely: "Missing income", "$0 \dashv 25,000$", "$25,000 \dashv 50,000$", "$> 50,000$". Then, the percentage of visitors in each income class for each cluster is reported.

For each variable we verify if there are significant differences between clusters with the Chi-square test for the qualitative variables, income included, and with the ANOVA test for the continuous variable. In the last column we report the *p*-values of the tests.

Some statistically significant dependency emerged between clusters and the profiling variables for each case study.

Among the visitors of MART (Table 4), the "Interested in the exhibition" segment (CL1) was on average older (47 years old), whereas the "Knowledge seeker" segment (CL3) was of younger age (39 years old on average). The "Not interested" (CL4) and the "Knowledge seeker" (CL3) reported respectively the highest and the lowest percentage of visitors with a University degree or more and, probably linked to this, the highest and the lowest percentage of visitors with more than €50,000 net family income per year. This result, strange and contradictory at first sight, seems to be a clear signal that confirm the enculturation role assumed by this museum.

The majority of the "Knowledge seeker" (CL3) and "Interested in the building" (CL5) segments are first–time visitors, whereas the "Family" segment (CL2) is composed, more than the other segments, by "repeat visitors". Finally, the "Interested in the exhibition" (CL1) and the "Knowledge seeker" (CL3) segments seemed to be, respectively, the more and less satisfied about the visit. This result suggests that the temporary showrooms are attractive and well exposed but, at the same time, the museum should pay more attention to the explanation of its objets d'art in order to make easier their understanding.

Among the visitors of the Merano CM (Table 5), the "Pro activities" (CL2) and the "Shopping" (CL4) segments reported respectively the lowest and the highest percentage of visitors with a University degree or more. As expected, the "Pro activities" (CL2), followed by the "Basic" (CL3), is composed by the highest percentage of married people, while the "No CM" (CL1) contains the lowest percentage of this kind of visitors.

Visitors grouped in the "No CM" (CL1) segment cover on average less miles to reach the Merano CM, reinforcing the idea that these visitors are less interested than the others in this event. Regarding the distribution of the household annual net income, we can note that the "Pro activities" (CL2) present the lowest percentage of families in the lowest class (up to €25,000) and the highest percentage of visitors who do not state their family net income ("Missing income"). Finally, visitors of the "Pro activities" (CL2) and "Shopping" (CL4) segments are the most satisfied about the CM, while the "No CM" (CL1) are not so satisfied. Therefore, the Merano CM seems to be a success both regarding the attractions proposed by the organizers and the quality of stands. Obviously, visitors who are not interested in the amusement attractions proposed, shopping, and Christmas atmosphere cannot be very satisfied about this kind of event.

Table 4: Socio–demographic characteristics of the MART visitors and characteristics of the visit.

Variables	CL1	CL2	CL3	CL4	CL5	*p*-value
First–time visitors	40.60	31.48	59.65	40.00	58.02	***
Very satisfied	55.85	51.85	35.09	44.00	50.77	**
Male (%)	41.14	51.85	40.35	44.00	50.00	0.338
Age (*mean*)	46.83	44.94	38.96	42.64	41.12	***
University (%)	83.28	85.19	64.91	90.00	85.38	***
Married (%)	58.53	55.56	47.37	54.00	50.77	0.432
Italy (%)	94.30	94.44	96.49	92.00	88.46	0.183
Km (*mean*)	201.96	164.28	342.42	384.76	336.07	0.126
Employed (%)	67.22	59.26	61.40	76.00	60.77	0.245
Income (%)						*
$0 \dashv 25,000$	18.39	20.37	24.56	26.00	17.69	
$25,000 \dashv 50,000$	41.14	35.19	42.11	28.00	40.00	
$> 50,000$	16.72	25.93	7.02	32.00	22.31	
Missing income	23.75	18.52	26.32	14.00	20.00	
Total expenditure (*mean*)	14.95	14.82	11.96	11.33	22.14	0.131
Shopping at the museum (*mean*)[a]	7.69	10.44	6.26	6.59	4.56	0.524
Shop not visited (%)	33.22	33.33	35.09	42.00	41.22	0.475

Notes:
Chi-square test was used for qualitative variables and continuous variables recoded in classes.
ANOVA test was used in order to test whether the mean value of the quantitative variables significantly differ among the clusters identified.
All test results are not significant unless indicated otherwise:
***Significant at $p \leq 0.01$, **Significant at $p \leq 0.05$, *Significant at $p \leq 0.1$.

[a] The sub–group of interviewees who have visited the shop of the museum are taken into account.

Table 5: Socio–demographic characteristics of the CM visitors and characteristics of the visit.

Variables	CL1	CL2	CL3	CL4	*p*-value
First–time visitors	64.39	62.41	61.11	72.26	0.224
Very satisfied	58.64	71.43	68.22	71.32	***
Male (%)	44.04	35.46	41.67	42.34	0.364
Age (*mean*)	37.75	39.24	38.81	38.43	0.448
University (%)	30.66	20.57	22.22	32.85	**
Married (%)	56.10	85.82	82.41	61.31	***
Italy (%)[a]	–	–	–	–	–
Km (*mean*)	315.35	397.38	335.14	361.24	***
Employed (%)	87.59	89.36	84.26	83.21	0.369
Income (%)					*
$0 \dashv 25,000$	28.71	14.18	21.30	27.01	
$25,000 \dashv 50,000$	34.06	43.26	43.52	37.23	
$> 50,000$	11.68	11.35	12.96	11.68	
Missing income	25.55	31.21	22.22	24.09	
Total expenditure (*mean*)	58.88	52.71	55.88	61.40	0.464
Shopping at the CM (*mean*)	38.00	53.56	34.24	41.12	0.294

Notes:
Chi-square test was used for qualitative variables and continuous variables recoded in classes.
ANOVA test was used in order to test whether the mean value of the quantitative variables significantly differ among the clusters identified.
All test results are not significant unless indicated otherwise:
***Significant at $p \leq 0.01$, **Significant at $p \leq 0.05$, *Significant at $p \leq 0.1$.

[a] All the interviewees are Italian.

The membership of each cluster identified per each survey was analyzed more in deep using the Multinomial Logit model. With this analysis we can find which socio-demographic and economic characteristics significantly influence the likelihood to be part of one of the groups with respect to a base, or baseline, group.

The "Interested in the exhibition" (CL1, MART results) and the "No CM" (CL1, CM results) were used as baseline respectively in the MART (Table 6) and the CM models (Table 7). The variables used for these models are described in Table 3.

MART Multinomial Logit result confirms some findings of the descriptive analysis discussed above. In particular, with respect to the baseline group, we can note that: the first–time visitors are more likely members of the "Knowledge seeker" (CL3) and of the "Interested in the building" (CL5) segment; the very satisfied visitors in the visit less likely are part of the "Knowledge seeker" (CL3) and of the "Not interested" (CL4) segment; visitors with a high level of education are less likely members of the "Knowledge seeker" (CL3); the higher the net family income, the higher the membership to the "Family" (CL2) group. The age of the visitor seems not to significantly influence the membership of one group instead of another, like the descriptive analysis suggested.

In addition, these empirical results leads to the conclusion that the "Family" (CL2) members are significantly discriminated from the "Interested in the exhibition" (CL1) group also because they are less likely autonomous workers or employed (full–time or part–time). "Knowledge seeker" (CL3) and "Not interested" (CL4) are more likely come from far away and the Italian visitors are more likely to be part of the "Knowledge seeker" group. Finally, men and visitors not interested in the shop of the museum more likely are part of the "Interested in the building" (CL5) group and, probably connected with these two variables, the higher the expenditure on shopping at the museum, the lower the likely to be a member of this segment.

Also the CM Multinomial Logit result, like the MART model result, confirms some findings of the descriptive analysis discussed above. As regards the socio-demographic and economic characteristics, we can note that the higher the level of education the lower the likely to be part of the "Pro activities" (CL2) group, while the higher the distance ("Km") the higher the likely to be part of this group. The positive relationship between the distance from the place of residence and interest in all the activities proposed by the organizers of the CM, and linked to the Christmas atmosphere, can be easily explain since visitors coming from far places made a long trip for something they considered worthwhile and precious.

Visitors who did not state their family income level are significantly more likely to be a member of the "Pro activities" (CL2) cluster. Married visitors are more probably grouped in the "Pro activities" (CL2) and "Basic" (CL3) groups, confirming that we have mentioned before (see paragraph 3.2). The higher the age

Table 6: Multinomial Logit coefficients for the MART survey.

Variables	CL2 Family		CL3 Knowledge seeker		CL4 Not interested		CL5 Interested in the building	
First–time visitors	-0.419	(0.31)	0.602	(0.33)*	-0.068	(0.34)	0.502	(0.23)**
Very satisfied	-0.241	(0.31)	-0.884	(0.32)***	-0.555	(0.32)*	-0.273	(0.22)
Socio–demographic and economic characteristics								
Male	0.487	(0.31)	0.088	(0.31)	0.163	(0.33)	0.502	(0.23)**
Age	0.029	(0.08)	-0.040	(0.08)	-0.131	(0.10)	-0.058	(0.06)
Age^2	-0.001	(< 0.01)	-0.001	(< 0.01)	0.001	(< 0.01)	0.001	(< 0.01)
University	0.084	(0.43)	-0.908	(0.37)**	0.677	(0.50)	0.307	(0.32)
Married	-0.020	(0.35)	0.127	(0.38)	-0.064	(0.37)	0.009	(0.27)
Km	-0.001	(< 0.01)	0.001	(< 0.01)**	0.001	(< 0.01)**	0.001	(< 0.01)
Italy	-0.613	(0.83)	1.822	(0.98)*	0.125	(0.84)	-0.358	(0.45)
Employed	-0.742	(0.42)*	-0.14	(0.41)	0.601	(0.52)	-0.408	(0.32)
Income	0.015	(0.01)**	-0.001	(0.01)	0.01	(0.01)	0.004	(0.01)
Missing income	0.227	(0.52)	-0.303	(0.59)	-0.307	(0.55)	-0.288	(0.35)
Expenditure								
Total expenditure	0.001	(0.01)	-0.003	(0.01)	-0.01	(0.01)	0.004	(< 0.01)
Shopping at the museum	-0.025	(0.03)	0.006	(0.03)	-0.012	(0.02)	-0.054	(0.02)**
Shop not visited	0.179	(0.35)	0.254	(0.36)	0.548	(0.34)	0.706	(0.25)***
Constant	-1.259	(2.01)	-0.936	(1.68)	0.19	(2.13)	0.798	(1.23)

Notes:
Base: CL1 = "Interested in the exhibition".
Robust Std. Err. in brackets. All test results are not significant unless indicated otherwise:
***Significant at $p \leq 0.01$, **Significant at $p \leq 0.05$, *Significant at $p \leq 0.1$
N = 587; Wald $\chi^2(60) = 115.32$; Prob $> \chi^2 = 0.00$; Pseudo $R^2 = 0.0769$; McFadden's $R^2 = 0.077$;
Cox & Snell $R^2 = 0.186$; Nagelkerke $R^2 = 0.199$.

of the visitors ("Age") the lower (less than proportional – "Age^2") the probability of being members of "No CM" (CL1) cluster. This means that the elders are more interested in the Christmas atmosphere and in shopping than the younger. Finally, autonomous and employed are significantly less likely to be a member of "No CM" cluster.

As regards the expenditure behavior, only the total expenditure level significantly affect the membership to one of the group identified. In particular, the higher the level of money spent for the trip, in general, the lower the likely to be part of the "Pro activities" (CL2) group. This negative relation reinforce the idea that the visitors belonging in this cluster are interested in doing and living fully the experience at the CM, tasting everything that is offered in it but, probably, quickly.

In conclusion, the results confirm that the "No CM" are the less satisfied in the visit at the CM, probably because they are not interested in it.

Table 7: Multinomial Logit coefficients for the CM survey.

Variables	CL2 Pro activities		CL3 Basic		CL4 Shopping	
First–time visitors	-0.277	(0.24)	-0.152	(0.26)	0.244	(0.23)
Very satisfied	0.468	(0.26)*	0.523	(0.26)**	0.463	(0.24)**
Socio–demographic and economic characteristics						
Male	-0.246	(0.25)	-0.052	(0.25)	-0.14	(0.23)
Age	0.799	(0.15)***	0.444	(0.12)***	0.131	(0.07)*
Age^2	-0.010	(<0.01)***	-0.005	(<0.01)***	-0.002	(<0.01)*
University	-0.463	(0.27)*	-0.348	(0.30)	0.181	(0.24)
Married	1.059	(0.29)***	0.924	(0.31)***	0.148	(0.26)
Km	0.002	(<0.01)***	0.001	(<0.01)	0.001	(<0.01)
Employed	-0.912	(0.40)**	-1.389	(0.38)***	-0.919	(0.35)***
Income	0.001	(<0.01)	0.001	(<0.01)	-0.001	(<0.01)
Missing income	0.712	(0.34)**	0.154	(0.37)	-0.081	(0.32)
Expenditure						
Total expenditure	-0.007	(<0.01)**	-0.002	(<0.01)	-0.001	(<0.01)
Shopping at the CM	0.001	(<0.01)	-0.003	(<0.01)	0.001	(<0.01)
Constant	-16.892	(2.94)***	-9.449	(2.23)***	-3.512	(1.32)***

Notes:
Base: CL1 = "No CM".
Robust Std. Err. in brackets. All test results are not significant unless indicated otherwise:
***Significant at $p \leq 0.01$, **Significant at $p \leq 0.05$, *Significant at $p \leq 0.1$
N = 587; Wald $\chi^2(60) = 115.32$; Prob $> \chi^2 = 0.00$; Pseudo $R^2 = 0.0769$; McFadden's $R^2 = 0.077$;
Cox & Snell $R^2 = 0.186$; Nagelkerke $R^2 = 0.199$.

4 Conclusions

In this paper we propose a clustering method based on the "bagging" (bootstrap aggregating) procedure. Bagging procedure is recognized to enhance stability of results and classification accuracy (Breiman, 1996).

Building on the BC method proposed by Leisch (1999), we make use of the Fuzzy *C*-Medoids Clustering (FCMdC) algorithm in the partitioning phase of the procedure. Once the hierarchical phase, which is carried out on the medoids identified, is completed, we attribute each unit to a cluster based on the maximum degree of membership to a particular medoid.

The proposed method inherits the properties of the original BC method:

1. the a priori definition of the number of clusters is not required;

2. classification results are more stable than those obtain by more traditional partitioning methods.

In addition, the fuzzy clustering approach allows for a more flexible allocation of units to each cluster. Indeed, some units can be fuzzy allocated to more than one cluster, if their characteristics are compatible with the profile of different clusters, a situation that cannot be detected with crisp clustering method.

Finally, the partitioning around medoids procedure allows to identify prototypes belonging to the considered dataset, that synthesize the structural information of each cluster (medoids). In many cases, dealing with observed units rather than with virtual units (centroids) could be suitable for the interpretation of the results.

To illustrate the main features of the proposed method, we carried out two applicative examples. The data collected through two surveys conducted among the visitors of two different cultural attractions located in the Trentino-South Tyrol region (Italy) were used. In both cases, the motivations of visit were used as segmentation variables. The first application is based on a survey conducted from June to September 2011 at the Museum of Modern and Contemporary Art (shortened to "MART") in Rovereto (Trento province, Italy), with face–to-face interviews submitted to 591 visitors. For this application the segmenting variables are dichotomous, therefore the dissimilarity measure adopted was the Jaccard dissimilarity index, and to analyze the empirical distribution of each segmenting variable among the clusters identified we make use of the normalized weighted Shannon entropy index.

The second applicative examples is carried out on data from a survey conducted during the Christmas Market ("CM") placed in Merano (Bolzano province, Italy) during the four weeks of Advent (from 30 November to 24 December), 2011. The face–to–face interviews submitted to 797 Italian visitors included

mainly Likert scale–based questions. To take into account the intrinsic vagueness of the answers, the variables were transformed into fuzzy variables. Then we make use of the Yang–Ko distance to detect the dissimilarity between units, and the distribution of the segmenting variables is represented via weighted boxplot.

In both cases, the weights are given by the membership degree of each unit to a specific cluster. In this way the fuzzy allocation of the units is explicitly taken into account.

The visitors of the MART museum were grouped into five clusters among which three could be considered as niche segments. The two bigger clusters identified are composed by visitors who are interested in the temporary exhibition proposed by the museum (CL1, "Interested in the exhibition") and in the building (CL5, "Interested in the building"). The three niche segments are composed by: visitors who made the visit to accompany a friend/family member (CL2, "Family"); visitors who want to learn something new and satisfy a curiosity (CL3, "Knowledge seeker"); visitors who are not interested in the visit but who want simply to rest/relax or occupying some leisure time (CL4, "Not interested"). The results suggest that MART museum must driving its promotional and marketing efforts to attract mainly the "Knowledge seeker" visitors. The members of this group are more probably first-time visitors, less satisfied in the visit than the other visitors, less educated and younger than the other visitors, and they came from far away, while remaining in Italy. So that, appropriate marketing strategies must be adopted in order to increase the visitors' satisfaction in the visit, encouraging these visitors to repeat the visit in the future.

The Italian visitors of the Merano CM were grouped into four clusters among which one could be considered as a niche segment (CL3, "Basic"). Also for this kind of cultural attraction, one group of visitors less interested in the visit emerged (CL1, "No CM"). The results suggest that the visitors interested in the activities proposed by the organisers, or in shopping, are more satisfied in the visit than the "No CM" members. Furthermore, the higher the age the lower the probability to being members of "No CM" group, i.e. the higher the age the higher the interest in the Christmas atmosphere and in shopping. Since the "No CM" is the bigger segment identify, grouping more than the half of the sample, new marketing strategies must be created in order to capture these visitors and to develop their loyalty to this kind of cultural event.

In future, we will investigate Bagged clustering-based segmentation methods for complex informational structures, i.e. for temporal and/or spatial information.

Acknowledgment

The authors thank the Professor Juan Gabriel Brida, Free University of Bolzano, for providing us data used in the empirical applications of this paper. The data collections were supported by the Autonomous Province of Bolzano project “Le attrazioni culturali e naturali come motore dello sviluppo turistico. Un’analisi del loro impatto economico, sociale e culturale”, Research Funds 2009.

References

Beane, T., & Ennis, D. (1987). Market segmentation: a review. *European Journal of Marketing*, *21*, 20–42.

Benítez, J., Martín, J., & Román, C. (2007). Using fuzzy number for measuring quality of service in the hotel industry. *Tourism Management*, *28*, 544–555.

Bennett, T., & Council, A. (1994). *The Reluctant Museum Visitor: A study of non-goers to history museums and art galleries*. Australia Council.

Bloom, J. (2005). Market segmentation: A neural network application. *Annals of Tourism Research*, *32*, 93–111.

Boone, D., & Roehm, M. (2002). Retail segmentation using artificial neural networks. *International journal of research in marketing*, *19*, 287–301.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140.

Brida, J. G., Disegna, M., & Osti, L. (2012). Segmenting visitors of cultural events by motivation: a sequential non-linear clustering analysis of Italian Christmas Market visitors. *Expert Systems with Applications*, *39*, 11349–11356.

Brida, J. G., Disegna, M., & Osti, L. (2013). Authenticity perception of cultural events: a host–tourist analysis. *Tourism, Culture & Communication, in press*.

Brida, J. G., Disegna, M., & Scuderi, R. (2012a). Visitors of two types of museums: a segmentation study. *Expert Systems with Applications, in press*. doi:10.1016/j.eswa.2012.10.039.

Brida, J. G., Disegna, M., & Scuderi, R. (2012b). Segmenting visitors of cultural event: the case of the Christmas Market visitors in Merano, Northern Italy. International conference on Sustainable Religious Tourism. Commandments, Obstacles & Challenges, 26–28 October 2012, Lecce–Tricase, Italy.

Budayan, C., Dikmen, I., & Birgonul, M. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy *c*-means method for strategic grouping. *Expert Systems with Applications*, *36*, 11772–11781.

Cochran, W. (2007). *Sampling techniques*. John Wiley & Sons.

Coppi, R., & D'Urso, P. (2002). Fuzzy k–mean clustering models for triangular fuzzy time trajectories. *Statistical Methods and Applications*, *11*, 21–24.

Curry, B., Davies, F., Phillips, P., Evans, M., & Moutinho, L. (2001). The Kohonen self-organizing map: an application to the study of strategic groups in the uk hotel industry. *Expert systems, 18*, 19–31.

Dolnicar, S. (1997). The use of neural networks in marketing: market segmentation with self organising feature maps. In *Proceedings of WSOM* (pp. 4–6). volume 97.

Dolnicar, S., Crouch, G., Devinney, T., Huybers, T., Louviere, J., & Oppewal, H. (2008). Tourism and discretionary income allocation. Heterogeneity among households. *Tourism Management, 29*, 44–52.

Dolnicar, S., & Leisch, F. (2000). *Getting more out of binary data. Segmenting markets by bagged clustering*. Working paper 71 SFB Adaptive Information Systems and Modelling in Economics and Management Science WU Vienna University of Economics and Business.

Dolnicar, S., & Leisch, F. (2003). Winter tourist segments in Austria: Identifying stable vacation styles using bagged clustering techniques. *Journal of Travel Research, 41*, 281–292.

Dolnicar, S., & Leisch, F. (2004). Segmenting markets by bagged clustering. *Australasian Marketing Journal (AMJ), 12*, 51–65.

D'Urso, P. (2007). Clustering of fuzzy data. In J. De Oliveira, & W. Pedrycz (Eds.), *Advances in fuzzy clustering and its applications* (pp. 155–192). J. Wiley and Sons.

D'Urso, P., Di Lallo, D., & Maharaj, E. (2013). Autoregressive model-based fuzzy clustering and its application for detecting information redundancy in air pollution monitoring networks. *Soft Computing-A Fusion of Foundations, Methodologies and Applications, 17*, 83–131.

Heiser, W., & Groenen, P. (1997). Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima. *Psychometrika, 62*, 63–83.

Hughes, H. (2002). Culture and tourism: a framework for further analysis. *Managing Leisure, 7*, 164–175.

Hung, W., & Yang, M. (2005). Fuzzy clustering on *lr*-type fuzzy numbers with an application in Taiwanese tea evaluation. *Fuzzy sets and systems, 150*, 561–577.

Hwang, H., Desarbo, W., & Takane, Y. (2007). Fuzzy clusterwise generalized structured component analysis. *Psychometrika, 72*, 181–198.

Kamdar, T., & Joshi, A. (2000). *On creating adaptive Web servers using Weblog Mining*. Technical Report TR-CS-00-05 Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County.

Kaufman, L., & Rousseeuw, P. (2005). *Finding groups in data: an introduction to cluster analysis*. Hoboken, NJ: Wiley.

Kim, H., Cheng, C., & O'Leary, J. (2007). Understanding participation patterns and trends in tourism cultural attractions. *Tourism Management*, *28*, 1366–1371.

Kim, J., Wei, S., & Ruys, H. (2003). Segmenting the market of West Australian senior tourists using an artificial neural network. *Tourism Management*, *24*, 25–34.

Kohonen, T. (1989). *Self-organizing and associative memory*. Berlin: Springer.

Kotler, P., Bowen, J., & Makens, J. (2010). *Marketing for Hospitality and Tourism, 5/e*. Upper Saddle River, New Jersey: Pearson Prentice Hall.

Krishnapuram, R., Joshi, A., Nasraoui, O., & Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Transactions on Fuzzy Systems*, *9*, 595–607.

Leisch, F. (1999). *Bagged clustering*. Working paper 51 SFB Adaptive Information Systems and Modelling in Economics and Management Science WU Vienna University of Economics and Business.

Leisch, F. (2006). A toolbox for *k*-centroids cluster analysis. *Computational statistics & data analysis*, *51*, 526–544.

Liao, S., Chu, P., & Hsiao, P. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, *36*, 11772–11781.

Mazanec, J. (1992). Classifying tourists into market segments. *Journal of Travel & Tourism Marketing*, *1*, 39–60.

McBratney, A., & Moore, A. (1985). Application of fuzzy sets to climatic classification. *Agricultural and Forest Meteorology*, *35*, 165–185.

Saarenvirta, G. (1998). Mining customer data: a step-by-step look at a powerful clustering and segmentation methodology. *DB2 Magazine online*, *3*, 10–20.

Schuster, M. (1991). *The audience for American art museums*. Technical Report 23 National Endowment for the Arts (NEA) Washington.

Tkaczynski, A., & Rundle-Thiele, S. (2011). Event segmentation: A review and research agenda. *Tourism Management*, *32*, 426–434.

Wedel, M., & Steenkamp, J. (1989). A fuzzy clusterwise regression approach to benefit segmentation. *International Journal of Research in Marketing*, *6*, 241–258.

Yang, M., & Ko, C. (1996). On a class of fuzzy *c*-numbers clustering procedures for fuzzy data. *Fuzzy Sets and Systems*, *84*, 49–60.