Social Indicators Research
https://doi.org/10.1007/s11205-020-02537-y

ORIGINAL RESEARCH

A Tourist Segmentation Based on Motivation, Satisfaction and Prior Knowledge with a Socio-Economic Profiling: A Clustering Approach with Mixed Information

Pierpaolo D'Urso[1] · Livia De Giovanni[2] · Marta Disegna[3] · Riccardo Massari[1] · Vincenzina Vitale[1]

Accepted: 28 October 2020

Abstract
The popularity of the cluster analysis in the tourism field has massively grown in the last decades. However, accordingly to our review, researchers are often not aware of the characteristics and limitations of the clustering algorithms adopted. An important gap in the literature emerged from our review regards the adoption of an adequate clustering algorithm for mixed data. The main purpose of this article is to overcome this gap describing, both theoretically and empirically, a suitable clustering algorithm for mixed data. Furthermore, this article contributes to the literature presenting a method to include the "Don't know" answers in the cluster analysis. Concluding, the main issues related to cluster analysis are highlighted offering some suggestions and recommendations for future analysis.

Keywords Fuzzy clustering · Mixed data · "Don't know" answers · Visitors

✉ Pierpaolo D'Urso
pierpaolo.durso@uniroma1.it

Livia De Giovanni
ldegiovanni@luiss.it

Marta Disegna
disegnam@bournemouth.ac.uk

Riccardo Massari
riccardo.massari@uniroma1.it

Vincenzina Vitale
vincenzina.vitale@uniroma1.it

[1] Department of Social Sciences and Economics, Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy

[2] Department of Political Sciences, Luiss University, Rome, Italy

[3] Department of Accounting, Finance and Economics, The Business School, Bournemouth University, 89 Holdenhurst Road, Bournemouth BH8 8EB, UK

Published online: 19 November 2020

1 Introduction

Cluster analysis is an exploratory description of a multidimensional dataset that aims to identify homogeneous groups of units, as similar as possible within groups and as different as possible among groups (Hennig et al. 2016). Market segmentation has been introduced in the early '50s and since then the number of clustering algorithms is massively grown. Being a multivariate descriptive tool, each cluster analysis will return a different picture of the data accordingly to the clustering algorithm used for the analysis. In this respect, it is worth to bear in mind that there are neither true or false results, but only suitable or unsuitable clustering algorithms. In fact, each clustering algorithm has its own characteristics and the researchers should adopt the one that returns the more meaningful result and that better suits the kind of data involved. For instance, if we believe that the observed units should belong to all clusters simultaneously rather than to be constraint to a sole cluster, then we should adopt an overlapping (fuzzy) clustering algorithm instead of a non-overlapping (crisp) clustering algorithm. To give another example, if categorical variables are used as segmentation variables, the Euclidean distance is not the best way to define distances between each pair of units. Instead, researchers should use a suitable distance or dissimilarity measure for categorical variables, such as the Jaccard similarity index or the Simple Matching coefficient (D'Urso and Massari 2019).

In this study, a review of the clustering algorithms adopted in the tourism studies published in four leading International Tourism Research Journals between 2015 and 2019 has been conducted. Results suggest that researchers in this field often choose a clustering algorithm (and the distance) without, or wrongly, motivating their choice. For instance, researchers in tourism have so far adopted the SPSS TwoStep clustering algorithm considering it a suitable way to identify homogeneous groups based on a set of variables of different nature/kind, i.e. mixed data (Zheng et al. 2019; Ritchie et al. 2017; Tkaczynski et al. 2015). However, Bacher et al. (2004) demonstrated that this algorithm doesn't work well with mixed data and they suggested to adopt other clustering algorithms instead. Nonetheless, in the tourism literature no clustering algorithms for mixed data, different from the SPSS TwoStep clustering algorithm, appeared to have been discussed so far. Therefore, the main purpose of this study is to describe, both theoretically and empirically, a novel clustering algorithm, recently published by D'Urso and Massari (2019), suitable to identify clusters of units based on mixed data. This clustering algorithm is so flexible that it works with any kind of data in input. Therefore, the second important contribution of this study is to discuss and present, for the first time in the tourism literature, a suitable way to include the "Don't know" answer in the cluster analysis. As highlighted by Dolnicar (2013), the "Don't know" answer is frequently included in surveys but, as our review has revealed, this information has never been included in a cluster analysis conducted in the tourism field.

The paper is structured as follows. Firstly, the results of the review of the clustering algorithms, as well as of their main characteristics, adopted in the articles recently published in four leading International Tourism Research Journals are presented (Sect. 2). In the following section (Sect. 3), the clustering algorithm for mixed data is theoretically presented along with a discussion on how to include imprecise data and "Don't know" answers in the analysis, how to validate, label and profile the final clusters. In Sect. 4, the results of the clustering algorithm for mixed data applied to the case study of the GEOPARC Bletterback (South-Tyrol, Northern Italy), a UNESCO World Heritage site, are presented and discussed. Finally, Sect. 5 concludes the paper highlighting the main contributions of this study and providing directions for future analyses.

2 Review of Cluster Analysis in Tourism

An updated review of recent studies performing a cluster analysis in travel and tourism has been conducted in this paper to highlight: (1) the kind of segmentation variables used; (2) the clustering methods and distances adopted and (3) the inclusion of the "Don't know" information in the analysis. Following the review undertaken by Dolnicar and Grün (2008), we considered the 45 studies published between 2015 and 2019 in the three leading International Tourism Research Journals (i.e. *Annals of Tourism Research*, *Tourism Management* and *Journal of Travel Research*) as well as the *Journal of Travel & Tourism Marketing* for its long tradition in publishing segmentation studies. Table 1 summarises the key information of the analysed papers. Note that only the articles in which static data (i.e. neither time or space were involved) analysed through an overlapping (Fuzzy) or non-overlapping (Crisp) clustering algorithm have been reviewed in this study. All published articles adopting model-based clustering algorithms, or any other segmentation algorithms (such as Network analysis), have not been considered in this review.

2.1 Clustering Algorithms

In the majority of the articles reviewed, a non-overlapping clustering algorithm (47%) has been adopted while overlapping algorithms are still barely used (7%). Among the non-overlapping clustering algorithm, the hierarchical algorithms (mainly using the Ward's method) are the most popular (48%) followed by sequential combinations of hierarchical and non-hierarchical algorithms (among which the Bagged Clustering algorithm), in which the hierarchical algorithm has often been used to determine the number of clusters to impose in the non-hierarchical algorithm. Researchers in tourism often use the result of Ward's clustering as a starting point for the *k*-means cluster analysis believing that this procedure will reduce algorithm randomness (Ernst and Dolnicar 2018). However, it has been demonstrated that this procedure works well only when the true number of clusters is known (Ernst and Dolnicar 2018).

The factor-cluster approach is a very popular procedure used in tourism (33%) confirming what has been found in the past (Ernst and Dolnicar 2018; Dolnicar and Grün 2008; Dolnicar 2003, 2002; Frochot and Morrison 2000). In specific, Dolnicar (2003), analysing the 234 publications collected by Baumann (2000), found that 27% of the articles published before 2000 in the area of Business administration were using this approach. In travel and tourism, when considering the period 1980-2000, the proportion of articles using this approach increases to 64% and 45%, accordingly to the reviews conducted by Frochot and Morrison (2000) and Dolnicar (2002), respectively. Between 2000 and 2005 the proportion rose again (58% accordingly to Dolnicar and Grün 2008), while considering a more recent period (2010–2016), Ernst and Dolnicar (2018) found a substantial reduction in popularity of this approach (23%). Accordingly to our review, the main motivation for the adoption of this approach is to reduce the multidimensionality of the original dataset to fewer factors usable as segmentation variables in the cluster analysis, confirming previous results. Among the disadvantages in using the factor analysis as a pre-processing technique, it is important to bear in mind (1) the important loss of information occurring when poor factors (low proportion of total variance explained) are used and (2) the fact that clusters are extracted from a fictitious space, i.e. factors rather than original variables (Dolnicar et al. 2012; Dolnicar and Grün 2008; Dolnicar 2003; Arabie and Hubert 1994). While

Table 1 Published cluster analysis studies in tourism analysed in this study

Author(s)	Algorithm	Segmentation variables	Data recoding	Distance
Antonakakis et al. (2019)	Non-H (K-means)	Continuous	No	NA
Derek et al. (2019)	H (Ward)	5 scale points	No	Euclidean
Fernández-Morales and Cisneros-Martínez (2019)	Bagged cluster	Continuous	No	Euclidean
Chung et al. (2019)	KliqueFinder	Continuous	No	NA
Khoo-Lattimore et al. (2019)	Fuzzy C-medoids for fuzzy data	5 scale points	Fuzzy	Fuzzy
Pan (2019)	H (Ward)	Discrete	No	Euclidean
Wang et al. (2019)	Factor-Cluster (NA)	7 scale points	FA	NA
Wassler et al. (2019)	H (NA) + Non-H (NA)	7 scale points	No	NA
Zheng et al. (2019)	TwoStep	Mixed data	No	NA
Disegna et al. (2018)	Fuzzy C-means for fuzzy data	10 scale points	Fuzzy	Fuzzy
Ferrante et al. (2018)	H (Ward)	Continuous	No	Euclidean
Gu et al. (2018)	Factor-Cluster (K-means)	7 scale points	FA	NA
Guttentag et al. (2018)	H (Ward) + Non-H (K-means)	6 scale points	No	Euclidean
Kim and Kim (2018)	Factor-Cluster (K-means)	5 scale points	FA	NA
Kim et al. (2018)	H (Ward)	5 scale points	Standardisation	Euclidean
Kline et al. (2018)	Factor-Cluster (H + K-means)	5 scale points	FA	NA
Murdy et al. (2018)	Factor-Cluster (K-means)	7 scale points	FA	NA
Rezaei et al. (2018)	TwoStep	Continuous	No	NA
Salvatore et al. (2018)	H (Ward) + Non-H (NA)	Continuous	No	NA
Stylidis et al. (2018)	H (Ward) + Non-H (K-means)	5 scale points	NA	Euclidean
Weaver et al. (2018)	H (NA)	5 scale points	No	NA
Bhati and Pearce (2017)	H (Ward)	Discrete	No	NA
Dryglas and Salamaga (2017)	Factor-Cluster (Ward + K-means)	5 scale points	FA	Euclidean
Hall et al. (2017)	Factor-Cluster (Ward + K-means)	9 scale points	FA	Euclidean
Kruger et al. (2017)	H (Ward)	Categorical	NA	City-blick
León-Borges and Lizardi-Jiménez (2017)	Non-H (K-means)	Continuous	No	Euclidean
Ritchie et al. (2017)	TwoStep	Mixed data	No	NA

Table 1 (continued)

Author(s)	Algorithm	Segmentation variables	Data recoding	Distance
Williams et al. (2017)	Non-H (K-means)	Discrete	No	NA
Zhao and Timothy (2017)	Factor-Cluster (K-means)	5 scale points	FA	NA
Abrate and Viglia (2016)	H (NA)	Continuous	Regression	NA
Amaro et al. (2016)	H (Ward, Complete, Average) + Non-H (K-means)	5 scale points	No	Euclidean
Choi et al. (2016)	Factor-Cluster (K-means)	7 scale points	FA	NA
D'Urso et al. (2016)	Fuzzy C-means for fuzzy data	10 scale points	Fuzzy	Fuzzy
Font et al. (2016)	TwoStep	Binary	No	NA
Liang and Hui (2016)	Non-H (K-means)	Continuous	Regression	NA
Liu et al. (2016)	Factor-Cluster (K-means)	5 scale points	FA	NA
Nawijn et al. (2016)	Factor-Cluster (NA)	7 scale points	FA	NA
Paker and Vural (2016)	Factor-Cluster (Ward + K-means)	5 scale points	FA	Euclidean
Alexander et al. (2015)	Factor-Cluster (K-means)	5 scale points	FA	NA
Denizci Guillet et al. (2015)	H (Ward)	Continuous	Conjoint Analysis	Euclidean
Pesonen (2015)	H (Ward)	Binary	NA	Euclidean
Prayag et al. (2015)	Bagged cluster	7 scale points	No	Euclidean
Priporas et al. (2015)	Factor-Cluster (Ward + K-means)	5 scale points	FA	Euclidean
Tkaczynski et al. (2015)	TwoStep	Mixed data	No	NA
Vila et al. (2015)	Factor-Cluster (Ward + K-means)	5 scale points	FA	Euclidean

FA factor analysis; *H* hierarchical algorithm; *Non-H* non-hierarchical algorithm; *TwoStep* SPSS TwoStep clustering; *NA* not specified

Dolnicar (2003) advised to use factor analysis before the cluster analysis only when factors were able to explain an high proportion of total variance and to interpret the results bearing in mind that the clusters were made in a transformed space, Dolnicar et al. (2012) recently demonstrated that the factor-cluster approach never performs better than a cluster analysis run on raw data directly. Therefore, the final suggestion, recently remarked in Dolnicar (2019), was to use other methods to reduce the number of variables involved in the cluster analysis (Dolnicar et al. 2012; Dolnicar and Grün 2008). Consequently, it was surprising to discover that a large proportion of studies recently published in travel and tourism still adopted the factor-cluster approach and it was even more surprising to note that the average total variance explained by the factors used in the cluster analysis was quite low (64.89%).

Finally, from our analysis it emerges that 11% of the studies adopted the SPSS TwoStep clustering procedure mainly because this algorithm is able to identify clusters based on mixed data (Zheng et al. 2019; Ritchie et al. 2017; Tkaczynski et al. 2015). However, accordingly to a simulation study performed by Bacher et al. (2004), this algorithm performs well if all the variables are continuous (as in Rezaei et al. 2018) while the results are less satisfactory when the segmentation variables are of mixed type because different combinations of categorical variables can determine the final partitions. Therefore, when mixed data are used in the cluster analysis, it is not recommended to use the SPSS TwoStep clustering procedure and different clustering algorithms should be adopted (Bacher et al. 2004).

In terms of distance adopted in the clustering algorithm, 53% of the reviewed studies didn't specify it while in the remaining studies the Euclidean distance was the most commonly used.

2.2 Segmentation Variables and "Don't Know" Answer

The majority (58%) of the studies reviewed in this paper used ordinal segmentation variables, mainly five and seven points Likert-type scales, confirming what has been already found by Frochot and Morrison (2000), Dolnicar (2002) and Dolnicar (2003) reviewing studies published before 2000. Likert-type scales, firstly introduced by Likert (1932), are ordinal scales since they are made up by a set of items, formulated in terms of linguistic expressions usually recoded into integers, characterised by a rank order. While Likert (Likert 1932) suggested that the distance between two consecutive response categories in a 5-points scale were equal, nowadays many researchers in different fields believe that the distance between scale points can't be defined (Dolnicar 2019) and the intervals between two consecutive response categories can't be presumed equal (Jamieson 2004). The discussion on the possibility to compute or not the distance between scale points is still open (Harpe 2015). However, it is important to note that this is a fundamental choice that any researchers have to take before to analyse Likert-type scale responses or before to use these information in more advanced statistical techniques, such as cluster analysis. In fact, following Likert idea, Likert-type scale can be considered as an interval scale and the responses can therefore be analysed using any arithmetic computations (e.g. summation, mean, standard deviation and Pearson's correlation) and any parametric tests (e.g. ANOVA test or *t*-test). However, if the more recent view (i.e. undefined distance between scale points) is embraced, mean should not be used as a measure of centrality to describe Likert-type scale responses and median should be computed instead. To reinforce this concept, Jamieson (2004) pointed out that the average value between "good" and "fair" is not "fair-and-a-half", and this is true even when the linguistic expressions are coded into integers.

Furthermore, non-parametric tests should be used instead of parametric tests and statistical methods defined on a metric space, such as factor analysis and item response theory, should not be adopted with this kind of data (Dolnicar 2002; Arimond and Elfessi 2001). Unfortunately, our review reveals that academics in travel and tourism using Likert-type scale responses as segmentation variables commonly ignored these criticisms frequently adopting the factor-cluster analysis (58%), non-overlapping clustering algorithm with Euclidean distance computed on raw data (19%) or non-overlapping clustering algorithm with Euclidean distance computed on standardised data (such as in Kim et al. 2018). While the suggestion of Dolnicar (2019) is not to use ordinal data (such as Likert-type scales) for segmentation purposes, our advice is to use suitable metrics and techniques when Likert-type responses are involved. For instance, one can recode the Likert-type responses into fuzzy data before the adoption of a clustering algorithm in which a distance for fuzzy data is used (Khoo-Lattimore et al. 2019; Disegna et al. 2018; D'Urso et al. 2016).

Continuing the review of segmentation variables used in travel and tourism, we found that 22% of the studies adopted continuous variables, 7% discrete variables, 7% mixed data, and in the remaining studies either binary or categorical data have been used.

Finally, accordingly to Dolnicar (2013), the "Don't know" option is frequently offered in surveys to prevent respondents from guessing when they do not know the answer. However, our review revealed that this option has never been included in a cluster analysis. We have been able to verify that this option was not included in the survey of seven studies in which the questionnaire was freely available. Unfortunately, when the questionnaire was not included in the study, we were unable to verify if this option was included, but treated as a missing information in the cluster analysis, or not in the original questionnaire.

3 Methodology

3.1 Mixed Data and Dissimilarity Measures

The majority of clustering algorithms are usually able to deal with either quantitative or qualitative (usually coded into dummies) variables and only a small proportion of literature is devoted to clustering algorithms of mixed data, i.e. data with mixed attributes.

Two main approaches are usually adopted in the literature to deal with the problem of using mixed data as segmentation variables in a cluster analysis. The first approach has been developed by Guha et al. (1999) who suggested to pre-convert all the variables to one type, i.e. either all numerical or all categorical, before the adoption of a cluster algorithm. This approach can be followed in the case the set of segmentation variables is made up by a combination of categorical and quantitative variables but it is not suitable when variables of different kinds (e.g. time series, space-time data, imprecisely observed data, textual data) are considered. Furthermore, this approach suffers of several drawbacks, as highlighted in D'Urso and Massari (2019) and Foss et al. (2016).

The second approach, developed by Gower (1971), is based on the definition of a suitable dissimilarity measure in which attributes are ideally weighted in order to define their relevance in the identification of the final partition. Let's assume that $\mathbf{X}$ is the matrix of P segmentation variables of different types (i.e. mixed data) observed on n units. Let's also assume that the P variables are arranged in S blocks of data of the same kind, i.e. the first p_1 variables are of the same kind (for instance, quantitative), the second p_2 variables are of the same kind but different from the first block of variables p_1 (for instance, categorical), and so on, such as $\sum_{s=1}^{S} p_s = P$. Therefore, the matrix $\mathbf{X}$ can be represented as follows:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x_{1,p_1} & \dots & x_{1,p_1+\ldots+p_{s-1}+1} & \dots & x_{1,p_1+\ldots+p_s} & \dots & x_{1,p_1+\ldots+p_{S-1}+1} & \dots & x_{1,p_1+\ldots+p_S} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ x_{i,1} & \dots & x_{i,p_1} & \dots & x_{i,p_1+\ldots+p_{s-1}+1} & \dots & x_{i,p_1+\ldots+p_s} & \dots & x_{i,p_1+\ldots+p_{S-1}+1} & \dots & x_{i,p_1+\ldots+p_S} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p_1} & \dots & x_{n,p_1+\ldots+p_{s-1}+1} & \dots & x_{n,p_1+\ldots+p_s} & \dots & x_{n,p_1+\ldots+p_{S-1}+1} & \dots & x_{n,p_1+\ldots+p_S} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}'_{11} & \dots & \mathbf{x}'_{1s} & \dots & \mathbf{x}'_{1S} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{x}'_{i1} & \dots & \mathbf{x}'_{is} & \dots & \mathbf{x}'_{iS} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{x}'_{n1} & \dots & \mathbf{x}'_{ns} & \dots & \mathbf{x}'_{nS} \end{bmatrix}$$

where the vector $\mathbf{x}'_{is}$ represents the set of observed values of the p_s variables of the s-th type for the i-th unit.

In such situation, Gower (1971) suggested to compute the squared distance between units i and j for each s-th block of variables using an adequate distance measure (for a list of distance measures and dissimilarity indices that can be used for different kind of variables see D'Urso and Massari 2019).

The final squared distance to use as input in a cluster algorithm will then be computed as a weighted sum of the S distances as follows:

$$d_{ij}^2 = \sum_{s=1}^{S} (w_s \cdot {}_s d_{ij})^2 = \sum_{s=1}^{S} \left[w_s \cdot d(\mathbf{x}_{is}, \mathbf{x}_{js}) \right]^2 \tag{1}$$

where w_s is the weight of the s-th distance matrix, i.e. the weight of the p_s variables, in the calculation of the final distance.

As it is easy to understand, this kind of dissimilarity measure is able to analyse simultaneously different kinds of data (e.g. numeric, categorical, time-varying, spatial, binary, imprecise, symbolic, sequential).

In this study, we consider two kinds of data, i.e. imprecise and binary data. Therefore, in the following discussion, we introduce and describe two suitable kinds of dissimilarity measures that can be used for the computation of d_{ij}^2 in Eq. 1.

3.1.1 Dissimilarity Measures for Imprecise Data

To make use of imprecise information, such as those collected through Likert-type scales, for clustering purposes it is necessary to pre-treat the data to limit the imprecision and vagueness that characterised them. A popular *a posteriori* correction mechanism used in the literature is to recode the imprecise information into fuzzy variables (Disegna et al. 2018). In the specific case of Likert-type scales variables, each individual score or expression is recoded into a range of possible values, i.e. into a fuzzy data. A general class of fuzzy data is the LR (Left and Right) fuzzy data (Dubois and Prade 1988). Specifically, let us assume that $\tilde{x}_{ik}$ is the k-th LR fuzzy variable ($k = 1, \dots, K$) observed on the i-th unit ($i = 1, \dots, N$), usually denoted as $\tilde{x}_{ik} = (m_{1ik}, m_{2ik}, l_{ik}, r_{ik})_{LR}$. Thus, the membership function, $\mu_{\tilde{x}_{ik}}(a_{ik})$, is defined as:

$$\mu_{\tilde{x}_{ik}}(a_{ik}) = \begin{cases} L\left(\frac{m_{1ik}-a_{ik}}{l_{ik}}\right) & a_{ik} \leq m_{1ik}\,(l_{ik} > 0) \\ 1 & m_{1ik} \leq a_{ik} \leq m_{2ik} \\ R\left(\frac{a_{ik}-m_{2ik}}{r_{ik}}\right) & a_{ik} \geq m_{2ik}\,(r_{ik} > 0) \end{cases} \tag{2}$$

where both L and R are decreasing "shape" functions defined in [0, 1]; m_{1ik} and m_{2ik} (with $m_{2ik} > m_{1ik}$) are respectively the left and right centres and the interval $[m_{1ik}, m_{2ik}]$ is usually called the "core" of the fuzzy number; l_{ik} and r_{ik} represent the left and right spreads, i.e. the vagueness of the observation. When both L and R are linear functions, the trapezoidal fuzzy number is defined. A trapezoidal fuzzy number characterised by $m_{1ik} = m_{2ik}$, i.e. only one centre is defined, is called triangular fuzzy number.

Table 2 reports a list of papers in which different fuzzy numbers have been suggested to recode Likert-type scales with different number of items. Note that some Likert-type scales have not been used in a fuzzy framework so far and no fuzzy recoding has been found for them.

The fuzzy recoding not only allows to cope with imprecise information but it also represents a tool to distinguish the "Don't know" answers from the missing answers (i.e. "Don't reply") that otherwise will be treated in the same way, i.e. deleted from the dataset. As suggested by Coppi et al. (2006), the answer "Don't know" can be recoded assuming a uniform distribution. For the best of our knowledge, no studies in the tourism literature have treated the "Don't know" answer and this is a first attempt to deal with this kind of information.

Once the Likert-type variables have been recoded into fuzzy data, an adequate distance measure for fuzzy data, such as the one suggested by Yang and Ko (1996), has to be used in the clustering analysis.

Following D'Urso (2007), the multidimensional version of the distance measure for trapezoidal (Tl) fuzzy numbers suggested by Yang and Ko (1996) between the i-th and j-th units, with $i \neq j$, is formalised as follows:

$$\begin{aligned} {}_{Tl}d_{ij} = (&\|\mathbf{m}_{1i} - \mathbf{m}_{1j}\|^2 + \|\mathbf{m}_{2i} - \mathbf{m}_{2j}\|^2 + \|(\mathbf{m}_{1i} - \lambda\mathbf{l}_i) - (\mathbf{m}_{1j} - \lambda\mathbf{l}_i)\|^2 \\ &+ \|(\mathbf{m}_{2i} + \beta\mathbf{r}_i) - (\mathbf{m}_{2j} + \beta\mathbf{r}_i)\|^2)^{\frac{1}{2}} \end{aligned} \tag{3}$$

where $\mathbf{m}_{1i}$ and $\mathbf{m}_{2i}$ are the vectors of the left and right centres, respectively; $\mathbf{l}_i$ and $\mathbf{r}_i$ are the vectors of the left and right spreads, respectively; $\|\cdot\|^2$ is the squared Euclidean distances; λ and β are parameters that summarise the shape of the membership function (D'Urso 2007).

When dealing with triangular (T) fuzzy data, i.e. only one center is defined, the Yang-Ko squared fuzzy distance between the i-th and j-th units, with $i \neq j$, will be:

$${}_{T}d^2_{ij} = \|\mathbf{m}_i - \mathbf{m}_j\|^2 + \|(\mathbf{m}_i - \lambda\mathbf{l}_i) - (\mathbf{m}_j - \lambda\mathbf{l}_i)\|^2 + \|(\mathbf{m}_i + \beta\mathbf{r}_i) - (\mathbf{m}_j + \beta\mathbf{r}_i)\|^2. \tag{4}$$

As suggested by D'Urso (2007) and Yang and Ko (1996), both L and R can be assumed to be linear, i.e. $\lambda = \beta = 1/2$, in both Eqs. 3 and 4.

3.1.2 Dissimilarity Measures for Binary Data

Different dissimilarity measures for binary data have been suggested in the literature so far (Everitt et al. 2011a; Eskin et al. 2002; Ng et al. 2007). A well-known dissimilarity measure is the simple matching coefficient (Sokal 1958). The simple matching coefficient between the i-th and j-th generic units is computed as follows:

Table 2 Recode Likert-type scale variables into fuzzy data

Items	Trapezoidal fuzzy number	Triangular fuzzy number
3	Andria et al. (2019) and Hadjimichael (2009)	–
4	Coppi et al. (2006) and Sii et al. (2001)	–
5	Andria et al. (2019), Coppi et al. (2006) and Sii et al. (2001)	Li (2013), D'Urso et al. (2013), Chan et al. (2012), Hung et al. (2010), Chou et al. (2008), D'Urso (2007), Hsu and Lin (2006), Hung and Yang (2005) and Lalla et al. (2005)
6	Sii et al. (2001)	ElMousalami (2019) and Verkuilen (2005)
7	Herrera et al. (1997)	Izadikhah et al. (2017), D'Urso (2015) and Kazemifard et al. (2011)
8	–	–
9	Coppi et al. (2006) and Herrera et al. (1997)	Martín et al. (2018), Lin et al. (2011) and Hsu et al. (2009)
10	–	Disegna et al. (2018), D'Urso et al. (2016) and Hadjimichael (2009)
11	–	Martín and Viñán (2017)

$$ {}_{SM}d_{ij} = \frac{a+d}{a+b+c+d} \tag{5} $$

where a is the number of variables equal to one for both units (common "presences"), d is the number of variables equal to zero for both units (common "absences"), $a+b$ is the number of variables equal to one for the i-th unit, $a+c$ is the number of variables equal to one for the j-th unit.

3.2 Clustering Algorithm for Mixed Data

For a detailed review of the clustering algorithms of mixed data suggested in the literature see Ahmad and Khan (2019) and D'Urso and Massari (2019).

In this study, we suggest to use the clustering algorithm of mixed data developed by D'Urso and Massari (2019) in a fuzzy framework, i.e. the Fuzzy C-Medoids Clustering of Mixed Data model (FCMd-MD model), based on the Gower's approach. Specifically, the fuzzy clustering approach has been preferred over the crisp clustering approach mainly because it is generally more efficient (dramatic changes in the value of cluster membership are less likely to occur in estimation procedures), it is less affected by both local optima and convergence problems, and it allows to relax the constraint that each unit can belong to a sole cluster (Disegna et al. 2018; Everitt et al. 2011b; Hwang et al. 2007). The C-Medoids clustering has been chosen over the C-Means clustering since the latter is not suitable when mixed data are used as segmentation variables (D'Urso and Massari 2019). Furthermore, from a practical point of view, the possibility to represent each final cluster by means of a real unit (the medoid) instead of a virtual one (the weighted mean computed over all units in a cluster) is appealing for policy makers and practitioners, in general. Since the FCMd-MD clustering model belongs to the class of procedures for partitioning around medoids, it attempts to alleviate the negative effects of presence of outliers in the dataset; thus it can be considered more robust than its possible C-means version in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. However, the FCMd-MD provides only a timid robustification.

The FCMd-MD model allows to discover homogeneous groups of units based on mixed data while measuring the relevance of each block of variables of the same kind towards the clustering process. In particular, thanks to the weighting system used for the calculation of the pairwise distances, the FCMd-MD model is able to rank the attribute types, i.e. groups of variables of the same kind, on the basis of their relevance in the computation of the final partition. Consequently, this algorithm is also able to identify blocks of variables that can eventually be removed from the cluster analysis causing little, if any, differences in the final result.

The FCMd-MD objective function that has to be minimised is as follows (D'Urso and Massari 2019):

$$ \begin{cases} \min : & \sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^{m} d_{ic}^{2} = \sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^{m} \sum_{s=1}^{S} \left[w_s \cdot d(\mathbf{x}_{is}, \widetilde{\mathbf{x}}_{cs}) \right]^2 \\ (\text{s.t.}) & \sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \geq 0 \\ & \sum_{s=1}^{S} w_s = 1,\ w_s \geq 0 \end{cases} \tag{6} $$

where u_{ic} indicates the membership degree of the i-th unit to the c-th cluster; $m > 1$ is a weighting exponent that controls the fuzziness of the final partition; $\widetilde{\mathbf{x}}_{cs}$ is the vector of values observed for the c-th medoid on the s-th variable type; $d_{ic}^2 = \sum_{s=1}^{S} [w_s \cdot d(\mathbf{x}_{is}, \widetilde{\mathbf{x}}_{cs})]^2$

is the overall weighted squared distance between the i-th unit and the c-th medoid (as described in Eq. 1); w_s is the weight associated to the s-th block of homogeneous variables and, hence, to the s-th distance ($s = 1, \dots, S$). For comparison reasons, the S distances have been normalised to vary in the range [0, 1] before the computation of the overall weighted squared distance. Finally, it is important to note that the weights w_s are automatically estimated within the clustering algorithm by solving a Lagrangian optimisation problem with two constraints, one for the membership degrees and one for the weights (for more details on the Lagragian problem see D'Urso and Massari 2019).

3.3 Fuzzy Partition Validity

To validate the final partition, the Fuzzy Silhouette (FS) index (Campello and Hruschka 2006) can be adopted. This index represents the weighted average of individual silhouettes width, λ_i, with weights derived from the fuzzy membership matrix $\mathbf{U} = \{u_{ic} : \ i = 1, \dots, n; c = 1, \dots, C\}$:

$$FS = \frac{\sum_{i=1}^{n} (u_{ip} - u_{iq})^{\alpha} \cdot \lambda_i}{\sum_{i=1}^{n} (u_{ip} - u_{iq})^{\alpha}}, \qquad \lambda_i = \frac{(b_i - a_i)}{\max\{b_i, a_i\}} \tag{7}$$

where a_i is the average distance between the i-th unit and the units belonging to the cluster p ($p = 1,...,C$) with which i is associated with the highest membership degree; b_i is the minimum (over clusters) average distance of the i-th unit to all units belonging to the cluster q with $q \neq p$; $(u_{ip} - u_{iq})^{\alpha}$ is the weight of each λ_i calculated upon $\mathbf{U}$, where p and q are, respectively, the first and second best clusters (accordingly to the membership degree) to which the i-th unit is associated; $\alpha \geq 0$ is an optional user defined weighting coefficient. The traditional Silhouette coefficient is obtained by setting $\alpha = 0$. The higher the value of FS, the better the assignment of the units to the clusters simultaneously obtaining the minimisation of the intra-cluster distance and the maximisation of the inter-cluster distance.

In this study, the FS has been properly modified to implement the squared distance as described in Eq. 1.

3.4 Cluster's Results, Labelling and Profiling

As per any fuzzy clustering algorithm, the FCMd-MD model results return C medoids, one representative per each cluster, and one ($N \times C$) matrix $\mathbf{U}$ containing the level of membership of the i-th unit to the c-th cluster. Since the higher the membership degree, the higher the strength of the association between each unit and each cluster, it is reasonable to use the $\mathbf{U}$ matrix to both label and profile the final clusters instead of using a "defuzzification" procedure (see, for instance, Khoo-Lattimore et al. 2019; Disegna et al. 2018). As described in D'Urso and Massari (2019), the weighted average of a generic quantitative variable $X = \{x_1, \dots, x_n\}$ in the c-th cluster (μ_{X_c}) can be computed as follows:

$$\mu_{X_c} = \frac{\sum_{i=1}^{n} u_{ic} x_i}{\sum_{i=1}^{n} u_{ic}}. \tag{8}$$

Similarly, the weighted proportion of the generic l-th modality of the $Y = \{y_1, \dots, y_n\}$ categorical variable, characterised by L ($L \geq 2$) categories, in the c-th cluster, ($w_{Y_{lc}}$), can be computed as follows:

$$w_{Y_{lc}} = \frac{\sum_{i=1}^{n} y_{il} u_{ic}}{\sum_{i=1}^{n} u_{ic}} \tag{9}$$

where y_{il} is the value of the l-th modality of the Y variable observed for the i-th unit, which is equal to 1 if the category is observed on the i-th unit and 0 otherwise. It is straightforward to verify that the greater the membership degree of unit i to cluster c, the greater the contribution of x_i and y_{il} to the weighted average and the weighted proportion respectively. The concept of weighted averages and weighted proportions can be easily extended to other attribute types.

Moreover, as described in Khoo-Lattimore et al. (2019), one can estimate the fractional multinomial logit to further profile the final clusters and identify the main factors influencing the membership to each cluster. Please note that the dependent variables of this model are the membership degrees contained in the **U** matrix.

Finally, the FCMd-MD model also returns the S weights w_s associated with each kind of variable included in the algorithm. It is important to note that the weights are comparable among them since they only depend on the normalised distance matrices. Therefore, they allows to identify (1) which kind of variables is more important in the identification of the final clusters and (2) which kind of variables can eventually be removed from the analysis because irrelevant.

4 A Segmentation of the Visitors in the GEOPARC Bletterbach Park of the South-Tyrol Region

The GEOPARC Bletterbach is a geological park located in the South-Tyrol region, Northern Italy (see Fig. 1). This site has been listed in the UNESCO World Heritage sites, together with other eight mountainous systems making up the Dolomites, since June 2009. The Bletterbach gorge offers to the visitors an enthusiastic trip inside the mountains, to the discovery of 40 million years of geological history of the Dolomites area.

As far as the protection of landscape and cultural heritage is concerned, it is worth noting that Bolzano is one of the most virtuous Italian provinces. For more than one indicator included by Istat in the BES[1] domain of *landscape and cultural heritage*, Bolzano takes the first place, even over the time, as we can see by looking at the trend of the expenditure by municipalities for protection and valorisation of cultural properties and activities in euro per capita (see Fig. 2) and that of density of farmhouses per 100 km^2 (see Fig. 3).

We argue that 57 euros of municipal expenditure for landscape protection per capita of Bolzano, in the 2017, is three times higher than the Italian average (18.8 euros per capita)

[1] BES is the acronym for *Equitable and Sustainable Well-being*; for details on Istat BES project, see Istat (2020).

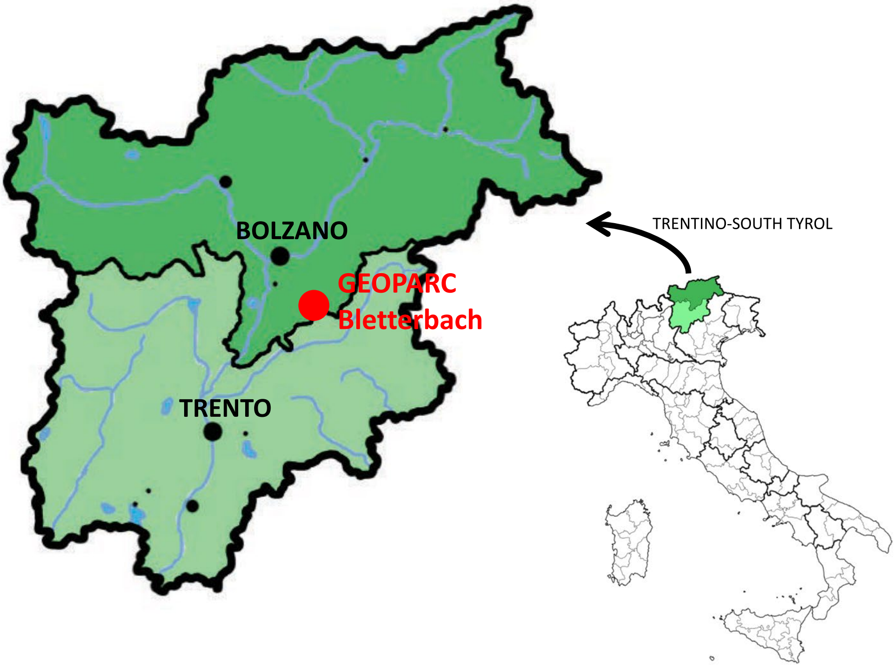

Fig. 1 Location of the GEOPARC Bletterback

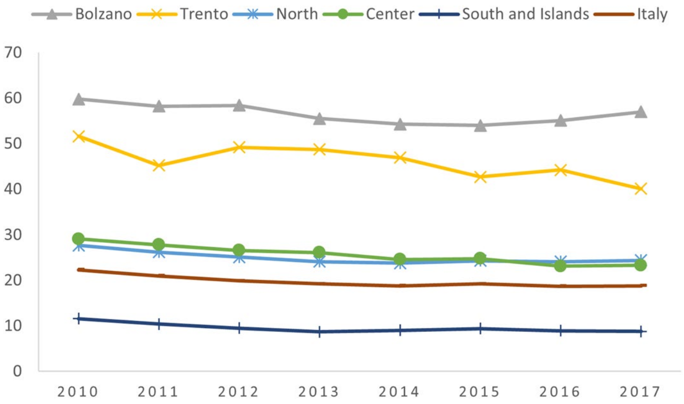

Fig. 2 BES indicator *Current expenditure of municipalities for culture*. Time series 2010–2017

and so far from the lowest value registered in Campania (4.6 euros per capita). It is worth highlithing that the second place is reached by the province of Trento with a value of about 2 times the Italian average.

In regard to the density of farms that practice agritourism, the gap becomes more evident. In the 2018, Bolzano shows the extreme value of 43 farms per 100 km^2, about 5 times the Italian average (7.8 per 100 km^2), followed by Tuscany and Umbria with 20.1 and 16.6

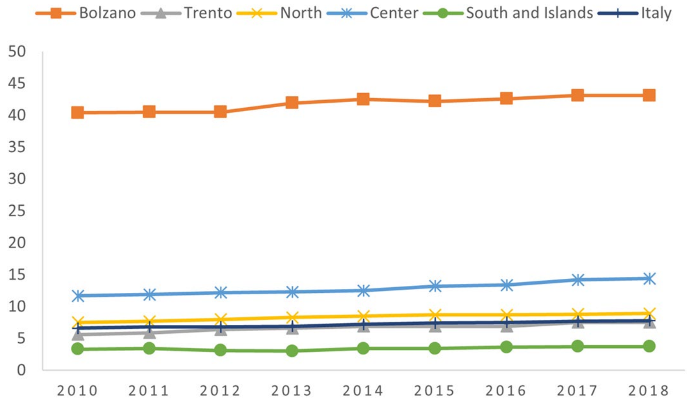

Fig. 3 BES indicator *Spread of rural tourism facilities—number of farmhouses per 100 km*^{2}. Time series 2010–2018

farms per 100 km^2, respectively. Basilicata and Aosta Valley, instead, are the two regions with the lowest values, under 2 farms per 100 km^2.

Bolzano is also one of the regions with the lowest levels of illegal building activities and, together with Trento, that with the lowest proportion of people dissatisfied for the landscape deterioration of the place where they live.

Data used in this study have been collected in 2016 and 2017 on different days of the opening period, going from the beginning of May to the end of October (subjected to weather conditions). A non-probability sampling technique has been adopted, as is commonly used in this type of research (Finn et al. 2000). GEOPARC Bletterbach staff were given clear instructions on ensuring randomness when recruiting visitors to participate in the survey. For example, only one respondent from a household or group could participate in the survey. The self administered surveys were completed at the end of the visit ensuring that visitors had a full personal opinion about the GEOPARC Bletterbach before participating in the research. A total of 686 questionnaires were collected, however, the final number of usable questionnaires is 443 due to the presence of missing information in either the segmentation or the profiling variables that have invalidated a huge amount of questionnaires.

The questionnaire has been designed together with the President and the Director of the GEOPARC Bletterbach, ensuring the inclusion of relevant information for the design of future managerial and marketing strategies. Characteristics of the trip, of the GEOPARC Bletterbach and of the respondents have been collected. In particular, respondents were asked to evaluate on a 6-point Likert-type scale both the importance of six motivation items (the percentage distribution of these items in the sample is reported in the top left graph of Fig. 5) and the satisfaction with seven items (the percentage distribution of these items in the sample is reported in the top left graph of Fig. 6). It is important to note that each of the 6-point Likert-type scale variables (both for the motivational items and the satisfaction items) was accompanied by the “Don’t know” option. The percentages of the “Don’t know” for the whole sample are represented in Figs. 8 and 9 and for the motivation items and the satisfaction items respectively. Furthermore, three questions have been set up to evaluate visitors prior knowledge about the site. In particular, it has been investigated if the respondent had already visited the GEOPARC Bletterbach before, if he/she knew that

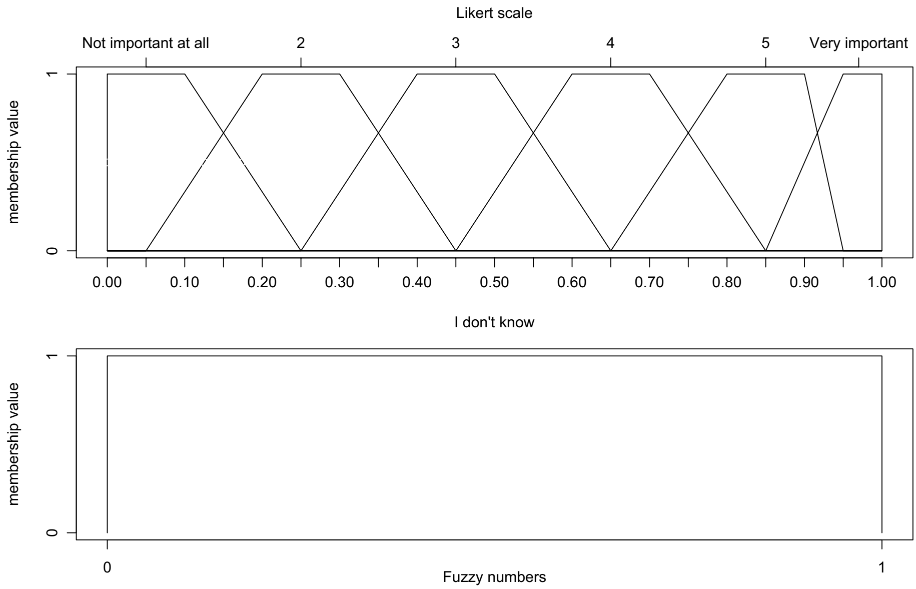

Fig. 4 Likert-type scale terms recoded into fuzzy numbers

the site has been listed in the UNESCO World Heritage site since 2009, and if the respondent had visited any other mountainous systems making up the Dolomites before the current visit at the GEOPARC Bletterbach. The top left graph of Fig. 7 represents the percentage distribution of these questions in the sample. The final part of the questionnaire contains socio-economic and demographic information about the respondent.

4.1 Clusters' Results and Labelling

If the aim is to identify homogeneous groups of visitors accordingly to their motivations, satisfactions and prior knowledges about the site, it is necessary to use a clustering algorithm for mixed data, as the FCMd-MD model presented in Sect. 3.2, since these variables are of different kinds. Please note that despite both motivation and satisfaction items are measured using a 6-point Likert type scale, the two scales measure different things, i.e. importance level and satisfaction level, therefore they have to be considered as two different kinds of variables.

It has been necessary to recode the Likert variables into fuzzy data, as explained in Sect. 3.1.1. In particular, the trapezoidal fuzzy number has been used in this study to recode the linguistic expression of the 6-point Likert-type scale while the "Don't know" answers have been recoded assuming the uniform distribution. The fuzzy recoding is displayed in Fig. 4.

As suggested in Sect. 3.1.2, the simple matching coefficient is adopted as dissimilarity measure for the prior knowledge variables while the distance for trapezoidal fuzzy numbers, expressed in Eq. 3, has been used for the fuzzy recoded data regarding both motivation and satisfaction variables.

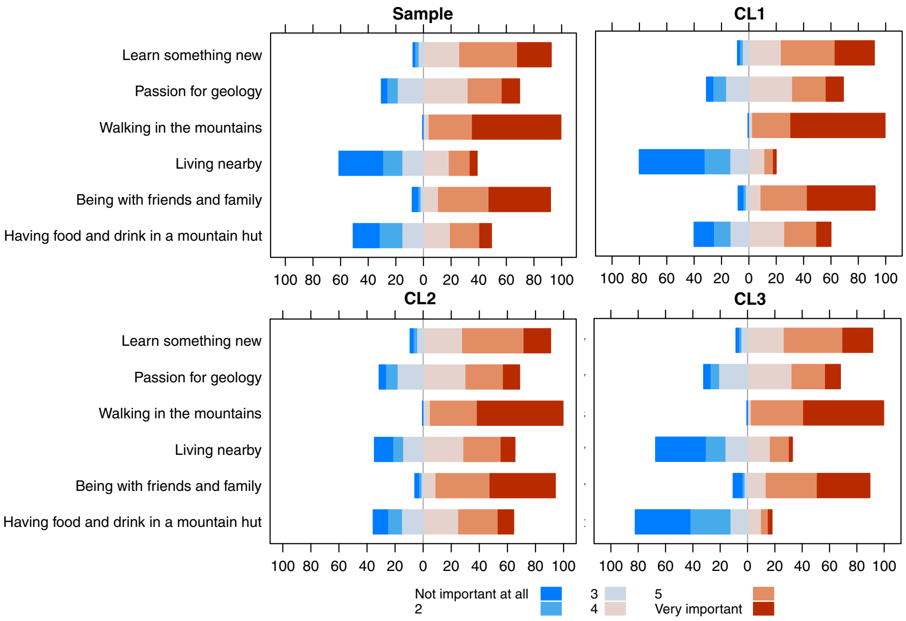

Fig. 5 Motivation variables: percentage distribution in the sample and weighted percentage distribution in each cluster

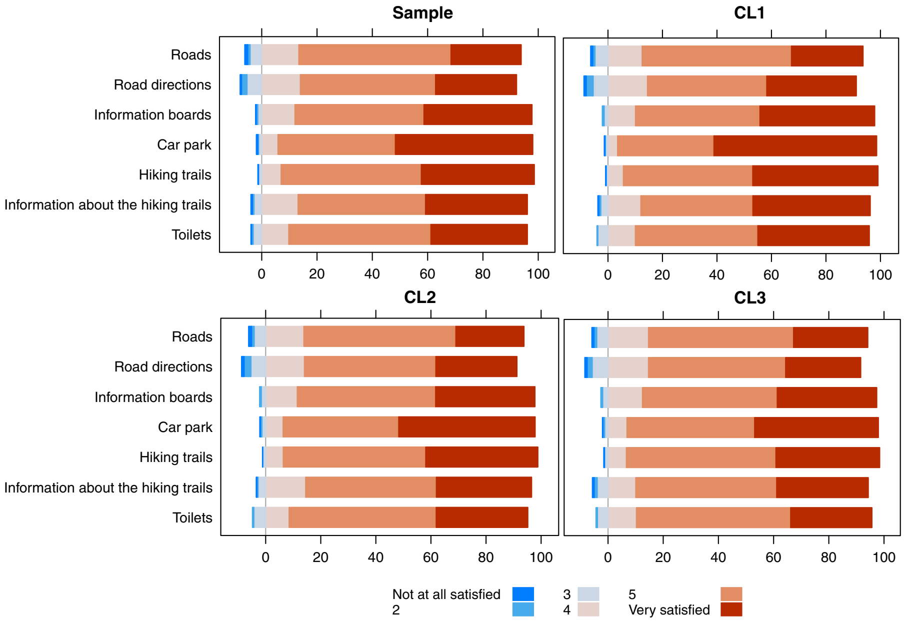

Fig. 6 Satisfaction variables: percentage distribution in the sample and weighted percentage distribution in each cluster

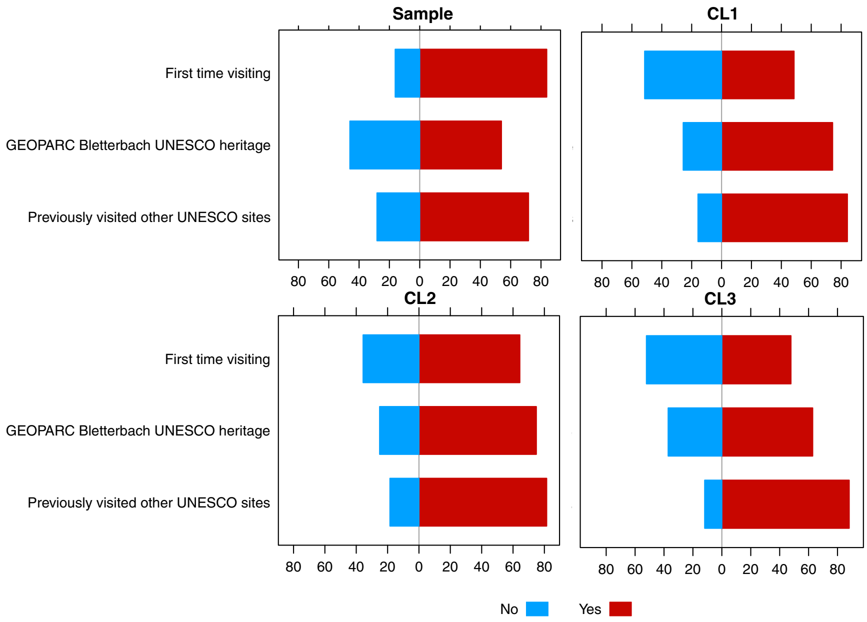

Fig. 7 Past knowledge variables: percentage distribution in the sample and weighted percentage distribution in each cluster

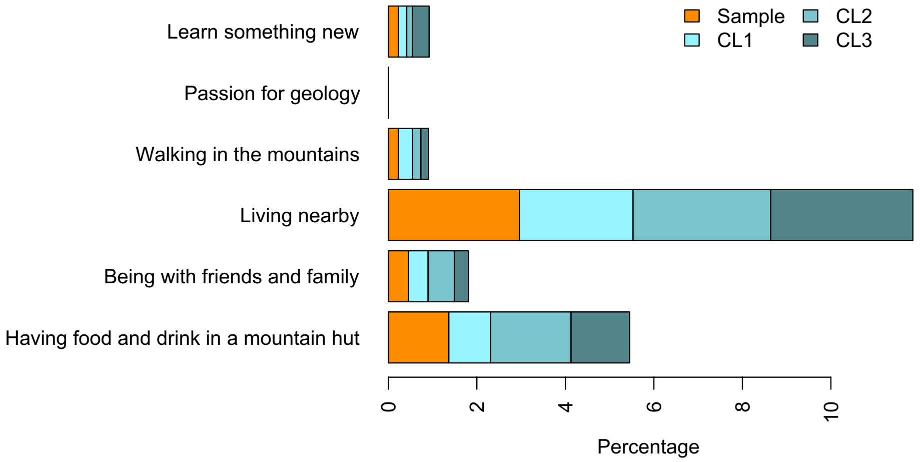

Fig. 8 Motivation variables: (weighted) percentage distribution of "I don't know" answer in the (clusters) sample

The Fuzzy Silhouette (FS) validity index (as described in Sect. 3.3) has been used to identify both the best fuzziness weight m (see Eq. 6) and the best partition, i.e. the final number of cluster C. The higher value of FS, which identify the best assignment of the units to the clusters, has been found setting $m = 1.3$ and $C = 3$.

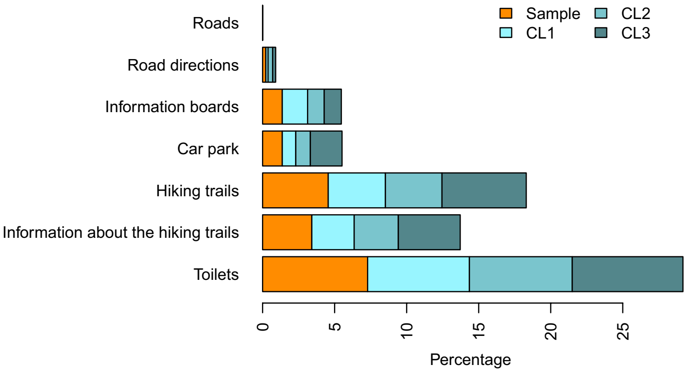

Fig. 9 Satisfaction variables: (weighted) percentage distribution of "I don't know" answer in the (clusters) sample

The weights w_s (see Eq. 6) are equal to 0.44, 0.47 and 0.09 for the three groups of variables, i.e. motivation, satisfaction and prior knowledge variables respectively. This result informs us that the prior knowledge variables play a small role in the identification of the final clusters and that this group of variables can eventually be removed from the cluster analysis. Moreover, satisfaction items play a slightly more important role in the identification of the final clusters compared to motivation items.

The medoids of the three clusters are units 101, 75 and 179. They are the representatives of the clusters. The labelling of the clusters is based on the weighted frequency distributions of the variables in each cluster as described in Sect. 3.4. Figures 5, 6 and 7 represent the weighted percentages of motivation, satisfaction and past knowledge variables respectively. Figures 8 and 9 show the weighted percentages of "Don't know" answers for motivation and satisfaction items respectively. Analysing these results, it emerges that cluster 1 (CL1) is made up by visitors who considered "Walking in the mountains" very important for the choice to visit the park and who were the most satisfied with all the elements of the GEOPARC Bletterbach but for the "roads", for which the weighted distribution is almost the same across clusters. Therefore CL1 can be labelled the "Mountains enthusiasts". Cluster 2 (CL2) presents the highest proportion of visitors for which "Living nearby" and "Having food and drink in a mountain hut" are important or very important elements for the choice to visit the GEOPARC Bletterbach, so this cluster can be labelled the "Locals". Finally, in cluster 3 (CL3) are grouped visitors less interested in "Being with friends and family" and "Having food and drink in a mountain hut" during their visit at the park. Moreover, CL3 visitors are also less satisfied with the GEOPARC Bletterbach elements analysed, especially with the "car park" and the "Hiking trails". In fact, even if this group collects the highest proportion of visitors who didn't know that the GEOPARC Bletterbach was a UNESCO heritage site since 2009, it also collects the highest proportion of visitors who "previously visited other UNESCO sites". Most probably, before to answer to the survey, these visitors compared the GEOPARC Bletterbach with other sites. Therefore, this cluster can be labelled the "Experienced visitors". Finally, it is important to note the "Don't know" answers are almost equally distributed in the three clusters with only minor differences.

4.2 Profiling

As described in Sect. 3.4, the final clusters can be profiled using additional information (i.e. data not used in the identification of the clusters) computing both weighted proportions (for qualitative variables) and weighted means (for quantitative variables). In this study, socio-economic and demographic variables collected in the survey have been used to profile the clusters.

A descriptive analysis of the whole sample (unweighted), as well as of each cluster (weighted), is reported in Table 3 of the "Appendix". As we can observe, the weighted sizes of the clusters are quite similar to each other indicating the absence of niche clusters. The chi-square test (for qualitative data) and the repeated analysis of variance test (for quantitative data) have been calculated but no significant difference among groups have been found on the basis of the profiling variables used.

To further describe the cluster, the membership degrees have been used as dependent variable of the Fractional Multinomial Logit (FML) model. The interpretation of the FML model results is very similar to the interpretation of the more common multinomial logit model results. Table 4 in the "Appendix" shows the estimated coefficients for the "Mountains enthusiasts" (CL1) and the "Locals" (CL2) clusters. The "Experienced visitors" (CL3) cluster has been chosen as reference category toward which the results have to be compared. Accordingly to the results, being married increases the probability to belong to the "Mountains enthusiasts", being between 30 and 50 years old increases the probability of being a "Locals" and having a high level of education (University degree or more) increases the probability to belong to the "Experienced visitors".

5 Conclusions

In this paper, a review of the clustering algorithms adopted in travel and tourism articles published in the four leading International Tourism Research Journals, in the last 5 years, has been conducted. From this review, it has emerged that only few studies have been conducted performing a cluster analysis using mixed data as segmentation variables and, in all these studies, the clustering algorithm adopted (i.e. the SPSS TwoStep clustering algorithm) was not appropriate. Therefore, the first important contribution of this paper is to present, both theoretically and by means of an application in the tourism field, a novel clustering algorithm, the FCMd-MD model proposed by D'Urso and Massari (2019), suitable to discover groups of homogeneous units characterised by a set of mixed data. In terms of segmentation variables, the review conducted in this study confirmed what has been found in the past, i.e. ordinal scales (such as Likert-type scale) are the most frequently used kind of variables. However, in the reviewed studies, we didn't find a critical and rigorous explanation of the approach followed for both the analysis and further use of these kind of data. As highlighted in Sect. 2.2, in the literature the debate regarding the classification, from a quantitative point of view, of Likert-type scale responses is still open and two main strands of thought have emerged so far: a first one that follows Likert's idea of equidistance between two consecutive response categories; a second one in which researchers believe that it is not possible, or not meaningful, to compute the distance between two consecutive response categories. The decision about which strand of thought the researcher wants to follow in the treatment of the Likert-type scale responses is fundamental since the choice among further statistical analysis depends on it. In particular, following Likert's point of view, Likert-type scale responses can be analysed through any statistical indices or methods defined in a metric space. However, following the second approach, Likert-type scales are ordinal, not interval, scales and the use of indices or methods designed in a metric space are not suitable. In travel and tourism only few researchers (Dolnicar 2019; Disegna et al. 2018; D'Urso et al. 2016) rose some concerns about the quantitative nature of the Likert-type scale responses and their use in a cluster analysis. However, while they all reached the same conclusion, i.e. Likert-type scales are ordinal scales for which the distance between two consecutive response categories can't be computed, Dolnicar (2019) suggested not to use this kind of data in a cluster analysis and Disegna et al. (2018) suggested to analyse these data in a fuzzy metric space, pre-recoding conveniently the Likert-type scale variables into fuzzy variables. Following Disegna et al. (2018)'s suggestion, in this paper we present how to include imprecise information in the cluster analysis. The second important methodological contribution of this paper is to present how to include the "Don't know" answers in a cluster analysis by means of a fuzzy metric space. Even if the "Don't know" answer is frequently included in travel and tourism surveys (Dolnicar 2013), accordingly to our review, this kind of information has never been used in a cluster analysis so far.

The main limitation of the clustering algorithm for mixed data suggested in this study is its inability to perform well in presence of outliers (as described in Sect. 3.2) and future studies will be devoted to the creation of a robust version of the FCMd-MD model, as stated by D'Urso and Massari (2019). Furthermore, in this study we assumed that each respondent has the same uncertainty/vagueness function for both motivation and satisfaction items. Therefore, the same fuzzy recoding has been applied per each respondent and each item. To be more accurate, the fuzzy recoding should be personalised to both respondent and information level. As already mentioned in Disegna et al. (2018), this is not an easy task and further studies should be directed to suggest suitable solutions for this issue.

Finally, we would like to make some suggestions to researchers in travel and tourism who want to perform a cluster analysis in the future. Firstly, accordingly to our review, the majority of researchers in this field didn't state the distance used in the clustering algorithm. This information is vital for the replication of the analysis using different data, so we would like to encourage researchers to provide such important information. Secondly, as already remarked by Dolnicar (2019), we don't encourage the use of the factor-cluster analysis with the purpose of reduce the number of segmentation variables since the cluster analysis is performed on a distort space. Furthermore, researchers should pay particular attention to the use of both factor analysis and Euclidean distance in cluster algorithms when Likert-type scale variables are involved in the analysis since these tools are designed under a metric space and their use in this context, for what already discussed above, should be carefully motivated.

Acknowledgements The authors thank the Editor and the referees for their useful comments and suggestions which helped to improve the quality and presentation of this manuscript.

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Appendix

Table 3 Profiling of clusters by sociodemographic characteristics

	Sample	Mountains enthusiasts (CL1)	Locals (CL2)	Experienced visitors (CL3)
Weighted proportions	100	33.51	34.52	31.97
Age (average)	44.03	44.21	44.19	43.68
Aged between 30 and 50 years old (%)	60.82	60.26	64.38	57.56
Married (%)	82.92	85.79	81.97	80.93
High level of education (%)	51.94	50.24	47.95	58.02
Employee (%)	59.91	60.29	59.39	60.06
Household size (average)	3.1	3.02	3.11	3.1
3 or more household members (%)	63.55	62.61	64.55	63.47
Italy (%)	29.16	31.79	26.35	29.42
Income easily satisfied household needs (%)	34.4	36.18	33.59	33.39
Income satisfaction: Don't know (%)	24.37	21.06	27.41	24.57

Weighted percentages and weighted means are reported

Table 4 Fractional multinomial logit results

Socio-demographic characteristics	Mountains enthusiasts (CL1)	Locals (CL2)
Aged between 30 and 50 years old	0.089 (0.12)	0.361 (0.13)***
Married	0.332 (0.15)**	0.022 (0.17)
High level of education	– 0.355 (0.12)***	– 0.434 (0.13)***
Employee	0.006 (0.12)	– 0.052 (0.13)
3 or more household members	– 0.065 (0.12)	– 0.017 (0.13)
Italy	0.113 (0.13)	– 0.184 (0.15)
Income easily satisfied household needs	0.092 (0.14)	0.087 (0.15)
Income satisfaction: Don't know	– 0.139 (0.14)	0.168 (0.16)
Constant	– 0.088 (0.2)	0.090 (0.21)

Coefficients are interpreted relative to the omitted category of "Experienced visitors" (CL3) cluster. All test results are not significant unless indicated otherwise. Robust standard errors are in parentheses. N = 439; Wald $\chi^2(14)$=35.24; $p > \chi^2 = 0.0014$. ***Significant at $p \leqslant 0.01$, **significant at $p \leqslant 0.05$, *significant at $p \leqslant 0.1$

References

Abrate, G., & Viglia, G. (2016). Strategic and tactical price decisions in hotel revenue management. *Tourism Management*, *55*, 123–132.

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, *7*, 31883–31902.

Alexander, A., Kim, S.-B., & Kim, D.-Y. (2015). Segmenting volunteers by motivation in the 2012 London Olympic games. *Tourism Management*, *47*, 1–10.

Amaro, S., Duarte, P., & Henriques, C. (2016). Travelers' use of social media: A clustering approach. *Annals of Tourism Research*, *59*, 1–15.

Andria, J., di Tollo, G., & Pesenti, R. (2019). A fuzzy evaluation of tourism sustainability. In *Business and consumer analytics: New ideas* (pp. 911–932). Springer.

Antonakakis, N., Dragouni, M., Eeckels, B., & Filis, G. (2019). The tourism and economic growth enigma: Examining an ambiguous relationship through multiple prisms. *Journal of Travel Research*, *58*, 3–24.

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In R. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160–189). Oxford: Blackwell.

Arimond, G., & Elfessi, A. (2001). A clustering method for categorical data in tourism market segmentation research. *Journal of Travel Research*, *39*, 391–397.

Bacher, J., Wenzig, K., & Vogler, M. (2004). *SPSS TwoStep Cluster: A first evaluation* (Vol. 2004-2).

Baumann, R. (2000). *Marktsegmentierung in den Sozial- und Wirtschaftswissenschaften: eine Metaanalyse der Zielsetzungen und Zugänge*. Diploma thesis at Vienna University of Economics and Management Science. Vienna.

Bhati, A., & Pearce, P. (2017). Tourist attractions in Bangkok and Singapore; linking vandalism and setting characteristics. *Tourism Management*, *63*, 15–30.

Campello, R. J., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, *157*, 2858–2875.

Chan, K. Y., Kwong, C., & Hu, B. Q. (2012). Market segmentation and ideal point identification for new product design using fuzzy data compression and fuzzy clustering methods. *Applied Soft Computing*, *12*, 1371–1378.

Choi, M. J., Heo, C. Y., & Law, R. (2016). Developing a typology of Chinese shopping tourists: An application of the Schwartz model of universal human values. *Journal of Travel & Tourism Marketing*, *33*, 141–161.

Chou, T.-Y., Hsu, C.-L., & Chen, M.-C. (2008). A fuzzy multi-criteria decision model for international tourist hotels location selection. *International Journal of Hospitality Management*, *27*, 293–301.

Chung, M. G., Herzberger, A., Frank, K. A., & Liu, J. (2019). International tourism dynamics in a globalized world: A social network analysis approach. *Journal of Travel Research*, *59*, 1–17.

Coppi, R., Giordani, P., & D'Urso, P. (2006). Component models for fuzzy data. *Psychometrika*, *71*, 733–761.

Denizci Guillet, B., Guo, Y., & Law, R. (2015). Segmenting hotel customers based on rate fences through conjoint and cluster analysis. *Journal of Travel & Tourism Marketing*, *32*, 835–851.

Derek, M., Woźniak, E., & Kulczyk, S. (2019). Clustering nature-based tourists by activity. Social, economic and spatial dimensions. *Tourism Management*, *75*, 509–521.

Disegna, M., D'Urso, P., & Massari, R. (2018). Analysing cluster evolution using repeated cross-sectional ordinal data. *Tourism Management*, *69*, 524–536.

Dolnicar, S. (2002). A review of data-driven market segmentation in tourism. *Journal of Travel & Tourism Marketing*, *12*, 1–22.

Dolnicar, S. (2003). Using cluster analysis for market segmentation: Typical misconceptions, established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research*, *11*, 5–12.

Dolnicar, S. (2013). Asking good survey questions. *Journal of Travel Research*, *52*, 551–574.

Dolnicar, S. (2019). Market segmentation analysis in tourism: A perspective paper. *Tourism Review*. ahead–of–print.

Dolnicar, S., & Grün, B. (2008). Challenging "factor-cluster segmentation". *Journal of Travel Research*, *47*, 63–71.

Dolnicar, S., Kaiser, S., Lazarevski, K., & Leisch, F. (2012). Biclustering: Overcoming data dimensionality problems in market segmentation. *Journal of Travel Research*, *51*, 41–49.

Dryglas, D., & Salamaga, M. (2017). Applying destination attribute segmentation to health tourists: A case study of polish spa resorts. *Journal of Travel & Tourism Marketing*, *34*, 503–514.

Dubois, D., & Prade, H. (1988). *Possibility theory*. New York: Plenum press.

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

D'Urso, P. (2015). Fuzzy clustering. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 545–573). London: Chapman and Hall.

D'Urso, P., De Giovanni, L., Disegna, M., & Massari, R. (2013). Bagged Clustering and its application to tourism market segmentation. *Expert Systems with Applications*, *40*, 4944–4956. https://doi.org/10.1016/j.eswa.2013.03.005.

D'Urso, P., Disegna, M., Massari, R., & Osti, L. (2016). Fuzzy segmentation of postmodern tourists. *Tourism Management*, *55*, 297–308.

D'Urso, P., & Massari, R. (2019). Fuzzy clustering of mixed data. *Information Sciences*, *505*, 513–534.

ElMousalami, H. (2019). Fuzzy logic for preconstruction project planning index. *MOJ Civil Engineering*, *5*, 5–19.

Ernst, D., & Dolnicar, S. (2018). How to avoid random market segmentation solutions. *Journal of Travel Research*, *57*, 69–82.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In Barbará D., Jajodia S. (eds.), *Applications of data mining in computer security* (pp. 77–101). Springer.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011a). *Cluster analysis* (5th ed.). London: Wiley.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.)., Wiley series in probability and statistics Hoboken: Wiley.

Fernández-Morales, A., & Cisneros-Martínez, J. D. (2019). Seasonal concentration decomposition of cruise tourism demand in Southern Europe. *Journal of Travel Research*, *58*, 1389–1407.

Ferrante, M., Magno, G. L. L., & De Cantis, S. (2018). Measuring tourism seasonality across European countries. *Tourism Management*, *68*, 220–235.

Finn, M., Walton, M., & Elliott-White, M. (2000). *Tourism and leisure research methods: Data collection, analysis, and interpretation*. London: Pearson Education.

Font, X., Garay, L., & Jones, S. (2016). A social cognitive theory of sustainability empathy. *Annals of Tourism Research*, *58*, 65–80.

Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, *105*, 419–458.

Frochot, I., & Morrison, A. M. (2000). Benefit segmentation: A review of its applications to travel and tourism research. *Journal of Travel & Tourism Marketing*, *9*, 21–45.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*, 857–871.

Gu, Q., Qiu Zhang, H., King, B., & Huang, S. (2018). Wine tourism involvement: A segmentation of Chinese tourists. *Journal of Travel & Tourism Marketing*, *35*, 633–648.

Guha, S., Rastogi, R., & Shim, K. (1999). ROCK: A robust clustering algorithm for categorical attributes. In*1999. Proceedings., 15th international conference on data engineering* (pp. 512–521). IEEE.

Guttentag, D., Smith, S., Potwarka, L., & Havitz, M. (2018). Why tourists choose Airbnb: A motivation-based segmentation study. *Journal of Travel Research*, *57*, 342–359.

Hadjimichael, M. (2009). A fuzzy expert system for aviation risk assessment. *Expert Systems with Applications*, *36*, 6512–6519.

Hall, J., O'Mahony, B., & Gayler, J. (2017). Modelling the relationship between attribute satisfaction, overall satisfaction, and behavioural intentions in Australian Ski resorts. *Journal of Travel & Tourism Marketing*, *34*, 764–778.

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*, 836–850.

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2016). *Handbook of cluster analysis*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/b19706.

Herrera, F., Herrera-Viedma, E., & Verdegay, J. L. (1997). Linguistic measures based on fuzzy coincidence for reaching consensus in group decision making. *International Journal of Approximate Reasoning*, *16*, 309–334.

Hsu, T.-H., & Lin, L.-Z. (2006). Using fuzzy set theoretic techniques to analyze travel risk: An empirical study. *Tourism Management*, *27*, 968–981.

Hsu, T.-H., Tsai, T.-N., & Chiang, P.-L. (2009). Selection of the optimum promotion mix by integrating a fuzzy linguistic decision model with genetic algorithms. *Information Sciences*, *179*, 41–52.

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

Hung, W.-L., Yang, M.-S., & Lee, E. S. (2010). A robust clustering procedure for fuzzy data. *Computers & Mathematics with Applications*, *60*, 151–165.

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

Istat,. (2020). *Rapporto BES 2019: il benessere equo e sostenibile in Italia*. Roma: Istituto nazionale di statistica.

Izadikhah, M., Saen, R. F., & Ahmadi, K. (2017). How to assess sustainability of suppliers in volume discount context? A new data envelopment analysis approach. *Transportation Research Part D: Transport and Environment*, *51*, 102–121.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical education*, *38*, 1212–1218.

Kazemifard, M., Zaeri, A., Ghasem-Aghaee, N., Nematbakhsh, M. A., & Mardukhi, F. (2011). Fuzzy emotional COCOMO II software cost estimation (FECSCE) using multi-agent systems. *Applied Soft Computing*, *11*, 2260–2270.

Khoo-Lattimore, C., Prayag, G., & Disegna, M. (2019). Me, my girls, and the ideal hotel: segmenting motivations of the girlfriend getaway market using fuzzy c-medoids for fuzzy data. *Journal of Travel Research*, *58*, 774–792.

Kim, E., Fredline, L., & Cuskelly, G. (2018). Heterogeneity of sport event volunteer motivations: A segmentation approach. *Tourism Management*, *68*, 375–386.

Kim, S., & Kim, S. (2018). Segmentation of potential film tourists by film nostalgia and preferred film tourism program. *Journal of Travel & Tourism Marketing*, *35*, 285–305.

Kline, C., Lee, S. J., & Knollenberg, W. (2018). Segmenting foodies for a foodie destination. *Journal of Travel & Tourism Marketing*, *35*, 1234–1245.

Kruger, M., Viljoen, A., & Saayman, M. (2017). Who visits the Kruger national park, and why? Identifying target markets. *Journal of Travel & Tourism Marketing*, *34*, 312–340.

Lalla, M., Facchinetti, G., & Mastroleo, G. (2005). Ordinal scales and fuzzy set systems to measure agreement: An application to the evaluation of teaching activity. *Quality and Quantity*, *38*, 577–601.

León-Borges, J.-A., & Lizardi-Jiménez, M. (2017). Hydrocarbon pollution in underwater sinkholes of the Mexican Caribbean caused by tourism and asphalt: Historical data series and cluster analysis. *Tourism Management*, *63*, 179–186.

Li, Q. (2013). A novel Likert scale based on fuzzy sets theory. *Expert Systems with Applications*, *40*, 1609–1618.

Liang, Z.-X., & Hui, T.-K. (2016). Residents' quality of life and attitudes toward tourism development in china. *Tourism Management*, *57*, 56–67.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 5–55.

Lin, L.-Z., Chen, W.-C., & Chang, T.-J. (2011). Using FQFD to analyze island accommodation management in fuzzy linguistic preferences. *Expert Systems with Applications*, *38*, 7738–7745.

Liu, Z., Huang, S. S., Hallak, R., & Liang, M. (2016). Chinese consumers' brand personality perceptions of tourism real estate firms. *Tourism Management*, *52*, 310–326.

Martín, J. C., Román, C., & Gonzaga, C. (2018). How different n-point Likert scales affect the measurement of satisfaction in academic conferences. *International Journal for Quality Research*, *12*, 421–440.

Martín, J. C., & Viñán, C. S. (2017). Fuzzy logic methods to evaluate the quality of life in the regions of ecuador. *Quality Innovation Prosperity*, *21*, 61–80.

Murdy, S., Alexander, M., & Bryce, D. (2018). What pulls ancestral tourists "home"? An analysis of ancestral tourist motivations. *Tourism Management*, *64*, 13–19.

Nawijn, J., Isaac, R. K., Liempt, A., v., Gridnevskiy, K., et al. (2016). Emotion clusters for concentration camp memorials. *Annals of Tourism Research*, *61*, 244–247.

Ng, M. K., Li, M. J., Huang, J. Z., & He, Z. (2007). On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *29*, 503–507.

Paker, N., & Vural, C. A. (2016). Customer segmentation for marinas: Evaluating marinas as destinations. *Tourism Management*, *56*, 156–171.

Pan, S. (2019). Tourism slogans-towards a conceptual framework. *Tourism Management*, *72*, 180–191.

Pesonen, J. A. (2015). Targeting rural tourists in the internet: Comparing travel motivation and activity-based segments. *Journal of Travel & Tourism Marketing*, *32*, 211–226.

Prayag, G., Disegna, M., Cohen, S. A., & Yan, H. (2015). Segmenting markets by bagged clustering: Young Chinese travelers to western Europe. *Journal of Travel Research*, *54*, 234–250.

Priporas, C.-V., Vassiliadis, C. A., Bellou, V., & Andronikidis, A. (2015). Exploring the constraint profile of winter sports resort tourist segments. *Journal of Travel Research*, *54*, 659–671.

Rezaei, J., Kothadiya, O., Tavasszy, L., & Kroesen, M. (2018). Quality assessment of airline baggage handling systems using SERVQUAL and BWM. *Tourism Management*, *66*, 85–93.

Ritchie, B. W., Chen, P. M., & Sharifpour, M. (2017). Segmentation by travel related risks: An integrated approach. *Journal of Travel & Tourism Marketing*, *34*, 274–289.

Salvatore, R., Chiodo, E., & Fantini, A. (2018). Tourism transition in peripheral rural areas: Theories, issues and strategies. *Annals of Tourism Research*, *68*, 41–51.

Sii, H. S., Ruxton, T., & Wang, J. (2001). A fuzzy-logic-based approach to qualitative safety modelling for marine systems. *Reliability Engineering & System Safety*, *73*, 19–34.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationship. *University of Kansas Science Bulletin*, *28*, 1409–1438.

Stylidis, D., Kokho Sit, J., & Biran, A. (2018). Residents' place image: A meaningful psychographic variable for tourism segmentation? *Journal of Travel & Tourism Marketing*, *35*, 715–725.

Tkaczynski, A., Rundle-Thiele, S. R., & Prebensen, N. K. (2015). Segmenting potential nature-based tourists based on temporal factors: The case of Norway. *Journal of Travel Research*, *54*, 251–265.

Verkuilen, J. (2005). Assigning membership in a fuzzy set analysis. *Sociological Methods & Research*, *33*, 462–496.

Vila, T. D., Darcy, S., & González, E. A. (2015). Competing for the disability tourism market-a comparative exploration of the factors of accessible tourism competitiveness in Spain and Australia. *Tourism Management*, *47*, 261–272.

Wang, J., Liu-Lastres, B., Ritchie, B. W., & Pan, D.-Z. (2019). Risk reduction and adventure tourism safety: An extension of the risk perception attitude framework (rpaf). *Tourism Management*, *74*, 247–257.

Wassler, P., Nguyen, T. H. H., Schuckert, M., et al. (2019). Social representations and resident attitudes: A multiple-mixed-method approach. *Annals of Tourism Research*, *78*, 102740.

Weaver, D., Tang, C., Shi, F., Huang, M.-F., Burns, K., & Sheng, A. (2018). Dark tourism, emotions, and postexperience visitor effects in a sensitive geopolitical context: A Chinese case study. *Journal of Travel Research*, *57*, 824–838.

Williams, N. L., Inversini, A., Ferdinand, N., & Buhalis, D. (2017). Destination eWOM: A macro and meso network approach? *Annals of tourism research*, *64*, 87–101.

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

Zhao, S. N., & Timothy, D. J. (2017). Tourists' consumption and perceptions of red heritage. *Annals of Tourism Research*, *63*, 97–111.

Zheng, D., Ritchie, B. W., & Benckendorff, P. J. (2019). Segmenting residents based on emotional reactions to tourism performing arts development. *Journal of Travel & Tourism Marketing*, *36*, 877–887.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.