Annals of Operations Research
https://doi.org/10.1007/s10479-024-05964-z

ORIGINAL RESEARCH

Cub model-based clustering of Likert-type data with a tourist satisfaction application

Nicolò Biasetton[1] · **Pierpaolo D'Urso**[2] · **Marta Disegna**[1] · **Luigi Salmaso**[1]

Received: 7 June 2023 / Accepted: 21 March 2024

Abstract
In investigating customer satisfaction with products or services, the most popular approach still relies on interviews or questionnaires to obtain consumers' opinions, and responses are usually measured by means of Likert-type scales. However, Likert-type data are inherently imprecise and uncertain. Thus, to obtain reliable analysis using such data, an a-posteriori correction must be adopted. The fuzzification procedure is the most common a-posteriori way to deal with uncertainty of Likert-type data. In this study, an alternative method to address the uncertainty of such data when used as input of a cluster analysis is proposed. The suggested method is based on the CUB model and the Fuzzy C-Medoids Clustering of Mixed Data algorithm and it is theoretically and empirically presented using real case study data. Advantages of the FCMd-CUB method are discussed in the conclusion section.

Keywords Likert-type variables · Fuzzy cluster analysis · CUB model · Mixed information

1 Introduction

Feelings, emotions, and preferences are complex psychological processes of interest in several disciplines (e.g. marketing, economics, and business) and are the basic information of any Customer Satisfaction (CS) analysis. These human emotions/feelings are typically captured in surveys that generally use ordinal scales, such as Likert-type scales. Since their introduction (Likert, 1932), Likert-type scales have been widely and commonly adopted in

✉ Nicolò Biasetton
nicolo.biasetton@unipd.it

Pierpaolo D'Urso
pierpaolo.durso@uniroma1.it

Marta Disegna
marta.disegna@unipd.it

Luigi Salmaso
luigi.salmaso@unipd.it

[1] Department of Management and Engineering, University of Padova, Stradella S. Nicola, 3, Vicenza 36100, Italy

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

Published online: 20 April 2024

both industry and academia, as they are user-friendly, easy to develop and easy to administer (see Gil et al., 2015). These scales are made up of a set of items, usually formulated in terms of linguistic expressions coded into natural numbers, characterised by a rank order. Despite their advantages, Likert-type scales can only return imprecise and vague information about the investigated respondents' feelings. First, people are asked to convert their thoughts into linguistic expressions and then into natural numbers, and these conversions can be inaccurate, causing loss of information, imprecision, and uncertainty (see D'Urso, 2007). Second, the meaning of each linguistic expression of the scale can be subjectively interpreted by the respondents due to their knowledge about the phenomenon under investigation, their culture, nationality, personal experiences and understanding of the question (see, Davidov et al., 2014). However, since most CS analyses rely on Likert-type scales to capture human thinking and personal feelings, it is fundamental to understand how to handle the uncertainty embedded in these scales to obtain reliable and accurate analysis and thus correctly inform practitioners. To the best of our knowledge, both a-priori alternatives and a-posteriori corrections have been suggested in the literature. Among a-priori alternatives, the use of different scales or methods to capture motivations, personal opinions/ judgements, or other psychographic variables has been discussed. In particular, la Rosa et al. (2015) and Shirahama et al. (2021) explored the use of simple visual analogue scales, while Gil and González-Rodríguez (2012) and Li (2013) suggested the use of fuzzy rating scales. The idea behind the a-priori alternatives is to avoid the use of Likert-type scales. Unfortunately, these alternative scales are less user-friendly, less popular, and more complex to both implement and analyse than the traditional Likert-type scale. Moreover, many researches work with secondary data, i.e., data obtained from existing sources and not originated by the researchers' own collection, and a-priori alternatives cannot be used. Because of these reasons, researchers have developed a-posteriori corrections to account for the uncertainty of the Likert-type data. Among a-posteriori corrections, two approaches emerge in the literature: the fuzzy sets theory-based approach; the CUB (Combination of discrete Uniform and shifted Binomial random variables) model-based approach. The approach based on fuzzy sets theory has been extensively used to consider the imprecision and vagueness inherent in both Likert-type variables and human thinking (see Coppi & D'Urso, 2002); Hu et al., 2005; Hung & Yang2010). Based on this approach, Likert-type data are recorded in fuzzy numbers prior to further analysis. Differently, the CUB model-based approach (introduced by Piccolo, 2003)aims to model the final answer as a mixture of two internal aspects, feeling and uncertainty.

As demonstrated by la Rosa et al. (2015), Likert-type data are among the most widely used segmentation variables in cluster analysis, but most studies do not take into account the uncertainty embedded in this type of data (Biasetton et al., 2023). As revealed in the literature review conducted by Biasetton et al. (2023), only 14% of the analysed articles explicitly address the issue of Likert-type data uncertainty in cluster analysis. This has been done either by transforming Likert-type scales into fuzzy numbers (Disegna et al., 2018; D'Urso et al., 2016, 2015; Khoo-Lattimore et al., 2019; D'Urso & Leski, 2020; D'Urso et al., 2014; D'Urso & De Giovanni, 2014) or by asking respondents to consider the distance between Likert-type scale points equal in order to reduce bias (Stavroulakis et al., 2020). Regarding the distance between scale points, while Likert (1932) suggested it could be considered equal, more recent researchers have introduced the idea that the distance between scale points cannot be defined (Dolnicar, 2019), and the intervals between two consecutive response categories cannot be presumed equal (Jamieson, 2004). This implies the impossibility of using any arithmetic computations to analyse Likert-type data. The discussion about the equality of distance between consecutive response categories is still ongoing (Harpe, 2015; D'Urso et al., 2021). To these literature, the work of Biasetton et al. (2023) has to be added as a first

attempt to use the CUB model to a-posteriori correct Likert-type data in cluster analysis. In particular, Biasetton et al. (2023) suggested recoding Likert-type data in fuzzy numbers using estimates of the CUB model with covariates to derive the parameters of the fuzzy number membership function.

As a follow-up to the method presented by Biasetton et al. (2023), this paper suggests a novel clustering method in which the original Likert-type data are not pre-transformed and their uncertainty is handled including the estimates of the CUB model as segmentation variables of the algorithm. In particular, the CUB model with covariates is used to estimate two latent components of the respondent's answer behaviour, i.e. feelings and uncertainty, which are then included as input information, along with the original respondents' answer, of a clustering algorithm for mixed data. The reader should note that the CUB model allows for modeling one ordinal variable, i.e., one question, at a time. Thus, the CUB model doesn't allow for the consideration of complex multivariate dependence structures among different ordinal variables. The use of a clustering algorithm allow to partially overcome this limitation by providing a multivariate description of responses, looking at similarities among respondents, without, however, aiming to study the dependence among ordinal variables.

In this paper, we suggest combining the CUB model with covariates with the Fuzzy C-Medoids Clustering of Mixed Data (FCMd-MD) model (D'Urso & Massari, 2019) to capture both the uncertainty/imprecision related to Likert-type scale data and the uncertainty associated with the assignment of a unit to each cluster. Furthermore, the FCMd-MD model allows one to automatically obtain the importance of each segmentation variable group (i.e., original Likert-type data, feelings, uncertainties) in the creation of the final partition. In the following, the suggested model will shortly be indicated as Fuzzy C-Medoids Clustering of CUB Data (FCMd-CUB). Compared to the a-posteriori fuzzification of the Likert-type data, the FCMd-CUB allows us to model the uncertainty adding individuals' characteristics, making the clustering analysis more precise and flexible. Furthermore, the FCMd-CUB model differs from the clustering model suggested by Biasetton et al. (2023) since it has the advantage of not requiring the pre-recoding of Likert-type data into fuzzy data.

The paper is organised as follows. In Sect. 2, the suggested method for clustering consumers' satisfaction using Likert-type data is described. In particular, in Sect. 2.1 the basic notation of the CUB model is discussed; in Sect. 2.2 an overview of dissimilarity measures for mixed data is presented; in Sect. 2.3 the Fuzzy C-Medoids Clustering of CUB Data (FCMd-CUB) is described. In Sect. 3, the results obtained by applying the FCMd-CUB model to empirical data are shown. In Sect. 4, some concluding remarks and guidelines for future research are provided.

2 Methods

2.1 CUB model

The Combination of discrete Uniform and shifted Binomial random variables (CUB) model has been first introduced by Piccolo (2003) to analyse and model Likert-type data responses. The underlying assumption of this model is that individual responses to a Likert-type question are the result of the combination of two components named feeling and uncertainty. Specifically, the feeling component determines the level of respondent's agreement/ pleasantness towards the object investigated while the uncertainty component collects different non-measured factors, e.g. respondent laziness, difficulties in understanding the question,

ignorance of the topics, wording and length of the scale, that affect the final response. Thus, the lower the uncertainty, the more reliable the answers. The final choice on an ordinal rating scale to express the evaluation of an object, the satisfaction with a service or the agreement with a sentence (according to the question asked) is the result of latent pairwise comparisons between the specific score and all the other possible scores (i.e. items of the Likert-type scale), and each comparison corresponds to a success (the object obtains a high score) or a failure (the object receives a low score).

Therefore, the probability distribution of the response variable can be modelled by a shifted binomial. Usually, each final answer is characterised by a certain degree of feeling and a certain degree of uncertainty. However, when the respondent is completely uncertain toward a question (i.e., no feeling component), the probability of choosing each evaluation is the same for each response, and the uniform distribution can properly represent the decision-making process. Consequently, the probability distribution of the response variable for the CUB model is a mixture of a shifted binomial and a uniform distribution.

Consider R as the response variable representing the ratings expressed by a sample of respondent when answering a question that required to be answered on a Likert-type scale. Assume that R can only take an integer value from 1 to m, with $m > 3$, where m is the highest value of the scale. Let us say, without loss of generality, that 1 represents the worst evaluation and m the best evaluation. The CUB model, which models the probability that R takes the generic value r, with $1 \leq r \leq m$, is formulated as follows:

$$Pr(R = r) = \pi \left[\binom{m-1}{r-1} (1-\xi)^{r-1} \xi^{m-r} \right] + (1-\pi) \left[\frac{1}{m} \right] \tag{1}$$

where (1-π), with $\pi \in (0; 1]$, measures the weight of the uncertainty of the responses since it increases with the importance of the discrete Uniform distribution and $(1 - \xi)$, with $\xi \in [0; 1]$, is a measure of perception towards the satisfaction aspect, as it increases the probability of giving high scores to the responses. The parameter $(1-\pi)$ is called uncertainty since it charges for the inherent uncertainty that arises when perception is translated into an evaluation and, therefore, measures the overall uncertainty of the respondent's assessment. The parameter $(1 - \xi)$ can be considered as a measure of the feeling towards the object. The CUB model reported in Formula 1 has been further improved by allowing to include the respondents' characteristics within the model. Covariates can affect feelings and uncertainty: the generalisation of the CUB model with covariates allows one to identify an individual value of feeling and uncertainty for each respondent (for a comprehensive treatment see Piccolo & Simone, 2019). Note that feelings and uncertainty are not constrained to be affected by the same set of covariates, so if p covariates affect the uncertainty and q covariates affect the feelings, the CUB $(p; q)$ model with covariates is formulated as follows:

$$Pr(R_i = r|\boldsymbol{x}_i, \boldsymbol{z_i}) = \pi_i \left[\binom{m-1}{r-1} (1-\xi_i)^{r-1} \xi_i^{m-r} \right] + (1-\pi_i) \left[\frac{1}{m} \right] \tag{2}$$

$$\begin{cases} logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x_i \beta} \\ logit(\xi_i) = log\left(\frac{\xi_i}{1-\xi_i}\right) = n\boldsymbol{z_i}\gamma \end{cases} \equiv \begin{cases} \pi_i = \frac{1}{1+e^{-\beta_0 - \beta_1 x_{1,i} - \ldots - \beta_p x_{p,i}}} \\ \xi_i = \frac{1}{1+e^{-\gamma_0 - \gamma_1 z_{1,i} - \ldots - \gamma_q z_{q,i}}} \end{cases} \tag{3}$$

where R_i represents the response of the i-th respondent, $i = 1, ..., n$; $\boldsymbol{x}_i$ is the vector of p covariates that affect the uncertainty of the i-th respondent; $\boldsymbol{z}_i$ is the vector of q covariates that affect the feeling of the i-th respondent; $\boldsymbol{\beta}$ is the vector of $(p + 1)$ unknown coefficients to be estimated for the uncertainty; $\boldsymbol{\gamma}$ is the vector of $(q + 1)$ unknown coefficients to be estimated for the feeling.

Model's parameters are estimated using a maximum likelihood function. An Expectation-Maximization (EM) algorithm provides a computational solution for calculating the maximum likelihood estimates. Operatively the CUB model algorithm use the frequency distribution of the observed Likert-type data and, starting from an initial random estimation of π and ξ, iteratively update them in order to fit the CUB model to the observed frequency distribution. On the version considering covariates (Eqs. 2 and 3), on each iteration the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ unknown parameters are estimated and updated.

In this work, to estimate the CUB models the `R` library proposed by Iannario (2016) and Simone (2020) has been adopted.

2.2 Mixed data

As described in Sect. 1, the segmentation variables will be of three different kinds: the original Likert-type variables ($\mathbf{Y}$), which are ordinal variables; the feelings (Ξ) and the uncertainties (Π) which are both probabilities, i.e. continuous variables in the range [0; 1]. Therefore, the distance or dissimilarity measure to use in the clustering algorithm has to be able to handle simultaneously different types of data.

In this paper, following D'Urso and Massari (2019), the Gower (1971) dissimilarity has been adopted and the final dissimilarity measure between units i and j is calculated as a weighted sum of different dissimilarity measures, one for each kind of information considered (Likert data, feeling and uncertainty component) as follows:

$$\begin{aligned} {}_{w}d_{ij}^2 &= {}_{y}d^2(\mathbf{y}_i, \mathbf{y}_j) + \left[w_{\xi}^2 \cdot {}_{\xi}d^2(\xi_i, \xi_j) + w_{\pi}^2 \cdot {}_{\pi}d^2(\pi_i, \pi_j)\right] \\ &= {}_{y}d^2(\mathbf{y}_i, \mathbf{y}_j) + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_{s}d^2(\mathbf{X}_i, \mathbf{X}_j) \\ &= {}_{y}d_{ij}^2 + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_{s}d_{ij}^2 \end{aligned} \tag{4}$$

where ${}_{y}d^2$, ${}_{\xi}d^2$ and ${}_{\pi}d^2$ are suitable dissimilarity measures for the original Likert-type data ($\mathbf{Y}$ - ordinal variables), feelings and uncertainties components (Ξ, Π - continuous variables), respectively; w_{ξ}^2 and w_{π}^2 are the weights of feeling and uncertainty, respectively, in the calculation of the final dissimilarity. The sum of the weights is constrained to be unitary, and the weight of the original Likert-type data is set equal to 1. Note that, for comparison's sake across attribute types, within the clustering algorithm the dissimilarity measures (for $\mathbf{Y}$,Ξ and Π) are normalised to vary in the range [0; 1].

According to Eq. (4), ${}_{y}d^2$ is a suitable dissimilarity measure for ordinal data while both ${}_{\xi}d^2$ and ${}_{\pi}d^2$ are suitable dissimilarity measures for probabilities.

Researchers can select the best dissimilarity measures for their applications based on data at hand and on dissimilarity measure's properties. A review of distance measures and dissimilarity indices for different types of attributes, including ordinal and interval-valued data, can be found in D'Urso and Massari (2019) and D'Urso et al. (2022). We suggest the adoption of the Kendall distance for ordinal data, to compute the dissimilarity between respondents with respect to their response to Likert-type questions, and the Euclidean distance to obtain the dissimilarity between respondents with respect to their feelings and uncertainties, as suggested by Everitt et al. (2011).

Remark 1 **Likert-type data.** Likert-type scales are a set of items formulated in linguistic expression coded into natural numbers and characterised by a rank order. While Likert (1932) suggested that the distance between two consecutive response categories on a 5-points

scale were equal, classifying Likert-type data as interval data, nowadays many researchers in different fields believe that the distance between two scale points cannot be defined or presumed equal, classifying Likert-type data as ordinal data (Dolnicar, 2019). This discussion is not new (for more details seeGardner, 1975) and researchers must be aware of it, since the decision on how to classify Likert-type data impacts the kind of statistical analysis that can be implemented.

In this study, Likert-type data are considered ordinal data. As a result, Likert-type data cannot be analysed by statistical method defined on a vector space and since the distance between two consecutive items of the scale cannot be either defined or presumed equal, the distance or dissimilarity measure to be adopted for such data cannot be a distance for continuous data.

Remark 2 **Kendall distance.** The Kendall distance measures the correspondence between the ranking of 2 vectors. The total number of possible pairings of such 2 vectors of size n, is $\frac{1}{2}n(n-1)$. Ordering the pairs by the values of the first vector, if the two vectors (namely x and y) are correlated, then they would have the same relative rank orders: for each y_i, count the number of $y_j > y_i$ (concordant pairs - c) and the number of $y_j < y_i$ (discordant pairs - d). The Kendall distance (Kendall, 1949; Abdi, 2007; D'Urso & Vitale, 2022) is then defined as:

$$d_{Kendall}(x, y) = 1 - \tau_{Kendall} = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{5}$$

where n_c is the total number of concordant pairs and n_d is the total number of discordant pairs. Note that suitable extension of the Kendall distance corrected for ties, such as the Kemeny distance (D'Urso & Vitale, 2022, see), can been considered. Moreover, the choice of the Kendall distance can represent a limit of this study: despite the characteristics of considering the order of ordinal variables instead of assuming equal distance between contiguous categories, it cannot perfectly capture the variability and the difference in ratings.

2.3 Clustering for mixed data with CUB parameters: FCMd-CUB

Since the segmentation data of the clustering algorithm are of different types (the original data are ordinal data, while both the feeling and the uncertainty components are continuous variables), a clustering algorithm of mixed data must be adopted (for a detailed review on this kind of clustering algorithm see D'Urso & Massari, 2019). In this study we suggest using the Fuzzy C-Medoids Clustering of Mixed Data (FCMd-MD) model developed by D'Urso and Massari (2019), where the input data are the observed data together with the estimated information derived from the CUB model, i.e. estimates of the individual feeling and uncertainty for each question. Thus, in the following we will briefly address the suggested model as the Fuzzy C-Medoids Clustering of CUB Data (FCMd-CUB). Following D'Urso and Massari (2019), the FCMd-CUB objective function to be minimised is as follows:

$$\begin{cases} \min: & \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m} w d_{ic}^{2} = \\ & \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m} \left\{ {}_{y}d^{2}(\mathbf{y}_i, \widetilde{\mathbf{y}}_c) + \left[w_{\xi}^{2} \cdot {}_{\xi}d^{2}(\xi_i, \widetilde{\xi}_c) + w_{\pi}^{2} \cdot {}_{\pi}d^{2}(\pi_i, \widetilde{\pi}_c) \right] \right\} = \\ & \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m} \left\{ {}_{y}d_{ij}^{2} + \sum_{s\in\{\xi,\pi\}} w_{s}^{2} \cdot {}_{s}d_{ij}^{2} \right\} \\ \text{(s.t.)} & \sum_{c=1}^{C} u_{ic} = 1,\ u_{ic} \geq 0 \\ & w_{\xi} + w_{\pi} = 1,\ w_{\xi} \geq 0,\ w_{\pi} \geq 0 \end{cases} \tag{6}$$

where:

- u_{ic} represents the membership degree of the i-th object to the c-th cluster;
- $m > 1$ represents a weighting exponent that control the fuzziness of the partition. The closer it is to 1 the closer the partition is to a crisp one. As recommended by Krishnapuram et al. (2001), in this study, m has been set at the value of 1.5;
- ${}_w d_{ic}^2$ represents the weighted dissimilarity measure between the i-th object and the c-th medoid:

Note that the CUB parameters' weights are automatically estimated within the clustering procedure by solving a Lagrangian problem. In the following the steps to solve the Lagrangian problem are described.

Let us consider the following Lagrangian function:

$$\mathcal{L}(\mathbf{u}_i, \mathbf{w}_s, \lambda, \gamma) = \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^m \left({}_y d_{ij}^2 + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_s d_{ij}^2 \right) - \lambda \left(\sum_{c=1}^{C} u_{ic} - 1 \right) - \gamma \left(\sum_{s\in\{\xi,\pi\}} w_s - 1 \right) \tag{7}$$

where $\mathbf{u}_i = (u_{i1}, ..., u_{iC})'$ is the vector of i-th observation membership degrees to all the C clusters and λ and γ are the Lagrange multipliers for the equality constraints.

The membership degree u_{ic} for each unit and for each cluster is computed. We set the partial derivatives of (7) with respect to u_{ic} ($\forall i \in \mathbb{Z} \mid 1 \leq i \leq n$ and $\forall c \in \mathbb{Z} \mid 1 \leq c \leq C$) and with respect to λ equal to zero, yielding:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{u}_i, \mathbf{w}_s, \lambda, \gamma)}{\partial u_{ic}} = 0 &\iff m u_{ic}^{m-1} \left({}_y d_{ic}^2 + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_s d_{ic}^2 \right) - \lambda = 0 \\ &\iff u_{ic} = \left(\frac{\lambda}{m \left({}_y d_{ic}^2 + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_s d_{ic}^2 \right)} \right)^{\frac{1}{m-1}} \end{aligned} \tag{8}$$

$$\frac{\partial \mathcal{L}(\mathbf{u}_i, \mathbf{w}_s, \lambda, \gamma)}{\partial \lambda} = 0 \iff \sum_{c=1}^{C} u_{ic} = 1 \tag{9}$$

Combining (8) and (9) we obtain:

$$\frac{\lambda}{m}^{\frac{1}{m-1}} = \frac{1}{\sum_{c=1}^{C} \left[\left(\frac{1}{\left({}_y d_{ic}^2 + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_s d_{ic}^2 \right)} \right)^{\frac{1}{m-1}} \right]} \tag{10}$$

and substituting (10) in (8) we finally obtain:

$$u_{ic} = \frac{1}{\sum_{c'=1}^{C} \left[\left(\frac{\left({}_y d_{ic}^2 + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_s d_{ic}^2 \right)}{\left({}_y d_{ic'}^2 + \sum_{s\in\{\xi,\pi\}} w_s^2 \cdot {}_s d_{ic'}^2 \right)} \right)^{\frac{1}{m-1}} \right]} \tag{11}$$

The same procedure is applied to obtain $\mathbf{w}_s = (w_\xi, w_\pi)$. We set the partial derivatives of (7) with respect to w_s $\forall s \in \{\xi, \pi\}$ and γ equal to zero, yielding:

$$\frac{\partial \mathcal{L}(\mathbf{u}_i, \mathbf{w}_s, \lambda, \gamma)}{\partial w_s} = 0 \iff w_s = \frac{\gamma}{2} \frac{1}{\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m {}_s d_{ic}^2} \tag{12}$$

$$\frac{\partial \mathcal{L}(\mathbf{u}_i, \mathbf{w}_s, \lambda, \gamma)}{\partial \gamma} = 0 \iff \sum_{s \in \{\xi, \pi\}} w_s = 1 \tag{13}$$

Combining (12) and (13) we can obtain:

$$\frac{\gamma}{2} = \frac{1}{\sum_{s \in \{\xi, 5\pi\}} \frac{1}{\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m {}_s d_{ic}^2}} \tag{14}$$

which substituted in (12) leads to

$$w_s = \frac{1}{\sum_{s' \in \{\xi, \pi\}} \frac{\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic}^m {}_s d_{ic}^2}{\sum_{i=1}^{n} \sum_{c=1}^{C} u_{ic s'}^m d_{ic}^2}} \tag{15}$$

Based on the obtained solutions of the minimisation problem (namely Eqs. 11 and 15) the FCMd-CUB algorithm have been developed and coded using the software R. The algorithm is schematically illustrated in Algorithm 1.

Algorithm 1 FCMd-CUB algorithm

Fix C number of clusters, $m = 1.5$, $max.iteration$ and generate randomly the degree matrix U;
Set $iter = 0$
Set $\mathbf{w} \leftarrow (0.5, 0.5)$ ▷ Initialize feeling and uncertainty weights
$(\mathbf{y}, \xi, \pi)_c | c = 1, ..., C$ ▷ Pick initial random medoids
$(\mathbf{y}, \xi, \pi)_{OLD,c} | c = 1, ..., C$ ▷ Define Old medoids as different from initial
while $((\mathbf{y}, \xi, \pi)_c \neq (\mathbf{y}, \xi, \pi)_{OLD,c} \wedge iter < max.iter)$ **do**
  $(\mathbf{y}, \xi, \pi)_{OLD,c} \leftarrow (\mathbf{y}, \xi, \pi)_c$ ▷ Store the current medoids
  Compute $\mathbf{u}_i$ $(i = 1, \ldots, n)$ using (11)
  Compute $\mathbf{w}$ by using (15)
  for $c = 1$ to C **do** ▷ Select the new medoids
    $q = argmin_{1 \leq i' \leq n} \sum_{i''=1}^{n} u_{i''c}^m \left\{ {}_y d_{i'i''}^2 + \sum_{s \in \{\xi, \pi\}} w_s^2 \cdot {}_s d_{i'i''}^2 \right\}$
    return $(\mathbf{y}, \xi, \pi)_c \leftarrow (\mathbf{y}, \xi, \pi)_q$
  end for
  $iter \leftarrow iter + 1$
end while

3 Case study

In this Section, an application of the proposed FCMd-CUB model to the clustering of tourists accordingly to their satisfaction against different aspects of the visited destination is presented. Data are drawn from the annual inbound survey of International Tourism in Italy conducted by the Bank of Italy (Banca d'Italia). The inbound-outbound frontier survey is the technique adopted for data collection. The stratified sampling method is applied (using different types of stratified variables for each type of frontier) and face-to-face interviews are carried out at national borders (including highways, railways, airports and ports). Sampling

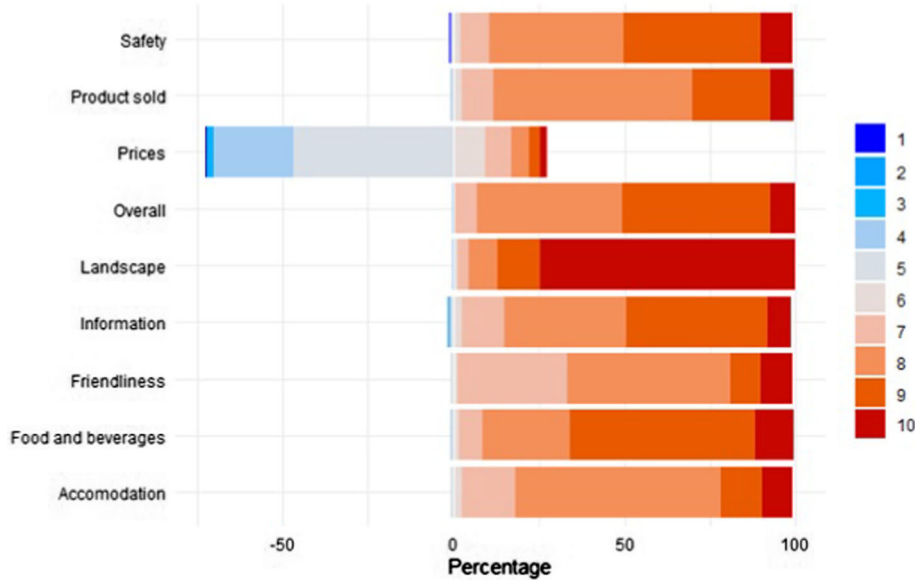

Fig. 1 Percentage distribution for each item

is done independently on each type of frontier. Tourists are interviewed at the end of the trip when they return to their place of habitual residence. The interviews are conducted at different times of the day, on both working days and holidays, and month by month, with a fixed number of interviews per period of the survey. In this study, a sample of 3,127 foreign tourists who spent at least one night in the Venice municipality in 2017 have been analysed. The respondents were asked to report their level of satisfaction with 9 different aspects of the destination: hospitality and friendliness of locals ("Friendliness"); hotels and other accommodation ("Accommodation"); food and beverages ("Food & Beverages"); prices and cost of living ("Prices"); landscape and natural environment ("Landscape"); quality and variety of products offered in stores ("Product sold"); information and tourist services ("Information"); safety ("Safety"); overall level of satisfaction with the destination ("Overall"). The level of satisfaction has been collected using a 10-point Likert-type scale from 1 = "Very unsatisfied" to 10 = "Very satisfied". Together with satisfaction, socio-demographic characteristics and trip information were asked through the survey. The sample is made up of people who travel for tourism, holiday, or fun and whose vacation is cultural.

Cluster analysis was conducted using the level of satisfaction with the different aspects of the destination as segmentation variables. Figure 1 represents the observed percentage distribution of the level of satisfaction in each aspect of the destination. In general, satisfaction is quite high regardless of the aspect considered (mainly equal to or greater than 6), except for the aspect related to prices and cost of living ("Prices").

Since the frequencies of unsatisfied tourists are minimal for most of the aspects considered, the 10-point Likert-type scale has been recoded into a 7-point Likert-type scale grouping the low items of the scale (from 1 to 4) into a singular one. Figure 2 shows the procedure adopted to recode the scale and Fig. 3 represents the percentage distribution observed of the level of satisfaction for each aspect of the destination on the new scale.

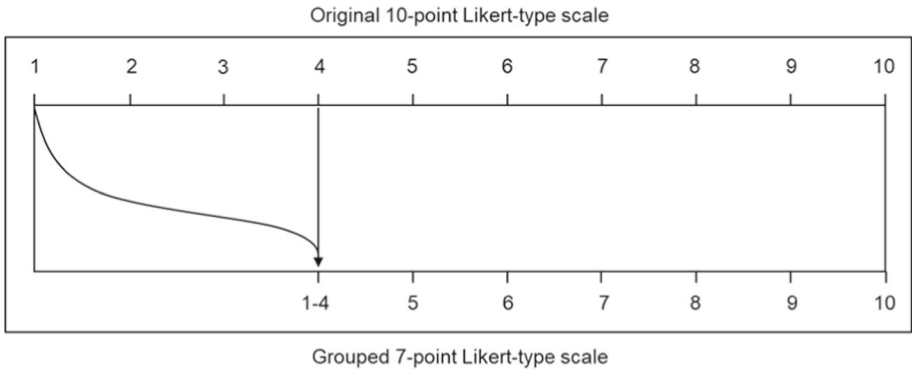

Fig. 2 Recoding of the 10-point Likert-type scale into the 7-point Likert-type scale

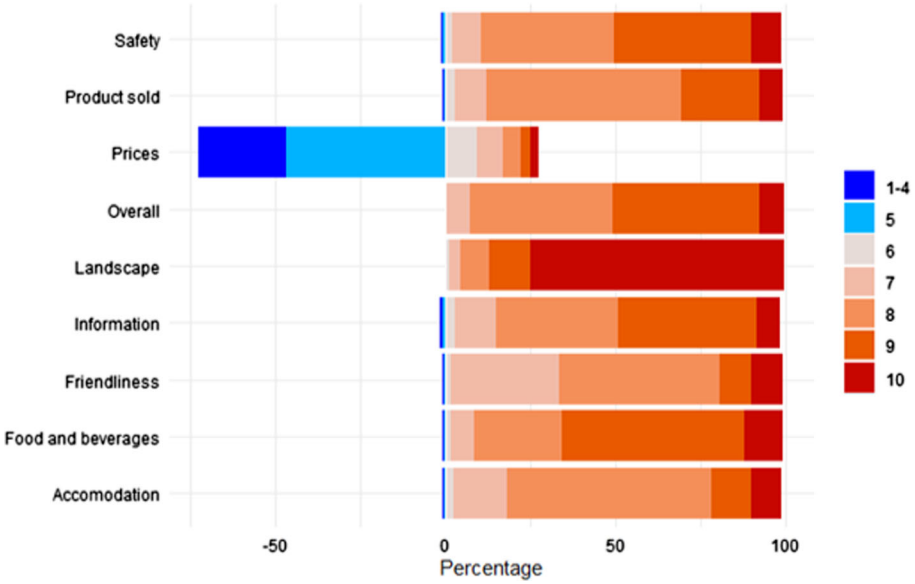

Fig. 3 Percentage distribution for each item on the modified scale

Feeling and uncertainty have been estimated for each respondent and for each question (i.e. segmentation variable) using Eqs. (2) and (3). The set of covariates of the CUB model is described in Table 1 and includes both socio-demographic and trip characteristics.

The modified Likert-type data, feelings and uncertainty for each Likert-type variable have been used as input variables of the FCMd-CUB model. The Gower dissimilarity discussed in Sect. 2.2 has been adopted using the Kendall distance for ordinal data (i.e. the modified Likert-type data) and the Euclidean distance for feelings and uncertainties. The FCMd-CUB model has been run for $C = 2, \ldots, 10$ clusters.

Remark 3 **Cluster validity.** Since the FCMd clustering algorithm requires one to a-priori define the number of clusters to extract, a validation procedure is adopted to define the final optimal partition. Due to its particularly satisfactory results in recognising the true number of clusters (see Arbelaitz, Gurrutxaga, Muguerza, Pérez, & Perona, 2013, for an extensive

Table 1 Description of the CUB model covariates

Variable *Socio-demographic characteristics*	**Description**
Age	<=34 / >34
Employment status	Self-employed / Employed / Other
Country of origin	EU countries / Non EU countries
Trip information	
N◦ of people that shared expenses	Count
N◦ of night spent in holiday	Count
Accommodation type	Hotel / Other

simulation study) the Fuzzy Silhouette index (SILF) ((see Campello & Hruschka 2006), i.e. the fuzzy version of the Average Silhouette Width (ASW) criterion (Kaufman & Rousseeuw, 2009), has been adopted. The SILF is a composite index which allows one to account for both the separability between the clusters and the homogeneity of units inside the same cluster. The SILF index represents the weighted average of individual silhouettes width, λ_i, with weights derived from the fuzzy membership matrix $U = u_{ic} : i = 1, ..., I; c = 1, ..., C$ and is defined as follows:

$$SILF = \frac{\sum_{i=1}^{N}(u_{ir} - u_{iq})^{\alpha}\lambda_i}{\sum_{i=1}^{N}(u_{ir} - u_{iq})^{\alpha}}, \lambda_i = \frac{(b_i - a_i)}{\max(b_i, a_i)} \tag{16}$$

where a_i is the average distance between the i-th unit and the units belonging to the cluster r, 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 $q \neq r$; $(r, q) \in [1, ..., c]$ are respectively the first- and second-best clusters (accordingly to the membership degrees) to which the i-th unit is associated; $(u_{ir} - u_{iq})^{\alpha}$ is the weight of each λ_i, calculated upon the fuzzy membership matrix U where (u_{ir}, u_{iq}) are the first and second largest element of the i-th row; $\alpha \geq 0$ is an optional user defined weighting coefficient (the ASW index is obtained by setting $\alpha = 0$).

The index has to be maximised, at least locally, in fact the higher the value of SILF, 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.

The SILF index has been computed for $c \in [2, \ldots, 10]$. The best partition suggested by this index is $C = 2$ followed by $C = 4$. In the upcoming discussion, we will focus on the four-clusters partition. Indeed, from both a managerial and practical standpoint, opting for the two-clusters partition is unappealing, as it lacks general informativeness and utility for devising new policies and strategies.

As an indication of the dimension of the clusters the weighted sample size of the 4 clusters can be considered. The sum of the membership degrees by cluster is 900.6, 325.2, 785.8 and 1115.3 for cluster from 1 to 4 respectively.

In order to label the final clusters, the weighted percentage distribution of the modified level of satisfaction (7-point Likert-type scale) for each aspect of the destination and for each cluster is represented in Fig. 4. Note that the black dots in the plot represent the category assumed by the medoid of the cluster on each of the response. The weighted percentages are computed using membership degrees as weights.

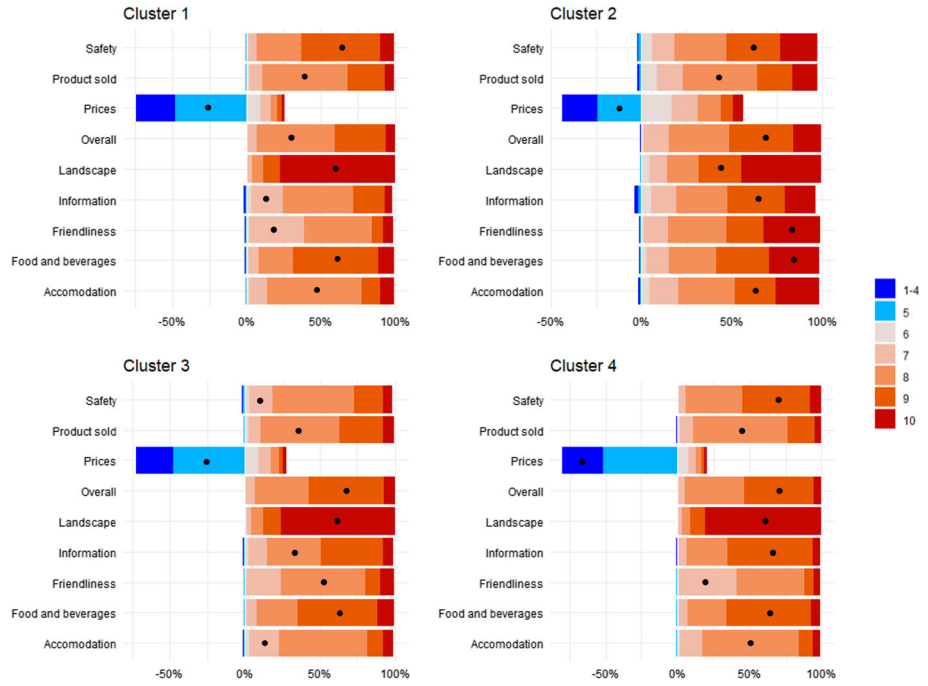

Fig. 4 Weighted distribution of the segmentation variables with medoids' indication

Cluster 2 is the smaller and groups tourists who are in general more satisfied with all aspects of the destination, except "Landscape", compared to the tourists in the other clusters. We can therefore labelled this cluster as the "Enthusiasts". Cluster 4 is the largest and brings together tourists who are more satisfied with the landscape and natural environment but less satisfied with all other aspects of the destination than other tourists. This cluster can therefore be labelled as the "Unfulfilled". Clusters 1 and 3 are very similar. However, cluster 1 groups tourists who rated the "Safety" and "Accommodation" aspects of the destination higher, while Group 3 tourists better evaluated the services offered by the Venice municipality, such as information and tourist services, they appreciated more the hospitality and friendliness of locals, and the quality of food and beverages offered by the city. Therefore, cluster 1 can be labeled as the "Moderates" and cluster 3 as the "Experiential tourists".

Figure 5 shows, the estimated values of feeling and Uncertainty for the final medoids. As we can observe, the "Enthusiasts" (Cluster 2) show the highest level of uncertainty, while the "Experienced tourists" (Cluster 3) exhibit less uncertain, especially regarding "Landscape" and "Prices". The weights attributed by the automatic optimization within the FCMd-CUB algorithm to the feeling and uncertainty components are respectively 0.515 and 0.485, respectively, not highlighting any particular dominance between the two components.

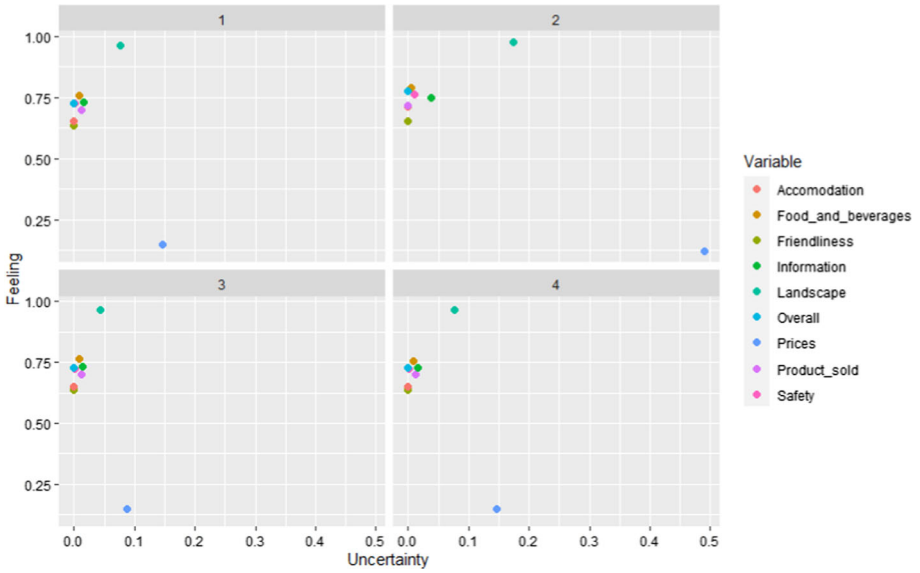

Fig. 5 Feeling and uncertainty parameter for each medoid by cluster

Table 2 Weighted distribution of the CUB model covariates

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	*p*-value
Socio-demographic characteristics					
Age <=34 (%)	32.72	34.35	36.11	34.77	
Self-employed (%)	6.25	12.78	6.75	4.94	***
Employed (%)	87.60	67.06	86.66	89.93	***
EU (%)	74.77	64.66	74.07	75.42	***
Trip information					
No of people	1.739	0.659	1.590	2.184	***
No of night	0.58	0.220	0.515	0.727	***
Hotel (%)	69.64	72.11	66.60	68.22	

Weighted percentage distributions and weighted means are reported. Significance of the χ^2 test of independence and ANOVA test, respectively for binary and quantitative variables, is reported. All test results are not significant unless indicated otherwise: ***Significant at $p \leq 0.01$, **Significant at $p \leq 0.05$, *Significant at $p \leq 0.1$

As an indication of the influence of covariates on the cluster, their weighted distribution or weighted average is reported in Table 2 along with the significance of the test.

4 Conclusions

In customer satisfaction analysis, Likert-type scales are frequently used to gather personal feelings and evaluations about a post-experience or post-use. However, the information obtained using such scales is uncertain and imprecise. Therefore, Likert-type data need to be

preprocessed before being used in further analysis, such as cluster analysis. This study aims to present an alternative way to a-posteriori handle the uncertainty and vagueness naturally embedded in Likert-type data when a cluster analysis is conducted. Although in the literature the most common preprocessing operation is fuzzification, in this study we suggest modelling the Likert-type data uncertainty by means of the CUB model with covariates (Piccolo, 2003). Furthermore, the estimates of the CUB model parameters, i.e. feelings and uncertainty, are used, together with the original respondents' answers to the Likert-type questions, as input variables of the clustering algorithm. Although any clustering algorithm for mixed data can be adopted, we suggest applying the FCMd-MD model (D'Urso & Massari, 2019) to obtain a more realistic multivariate description of the units and to identify the importance of each segmentation information in the creation of the final partition. Therefore, in this study we suggest to combine the CUB model with the FCMd-MD algorithm, obtaining the so-called FCMd-CUB model.

The main advantages in using the suggested a posteriori correction are: 1) the possibility of modeling individual uncertainty arising from Likert-type data using additional respondents' features collected through the questionnaire; 2) the removal of the elicitation problem, i.e. the necessity to define the membership function of the fuzzy numbers, since no fuzzy recoding is involved. From a business point of view, the FCMd-CUB model is of particular interest since it allows one to obtain more reliable clusters with the possibility to effectively direct marketing and managerial resources to specific customers.

Limitations of the present work include: 1) the choice of the dissimilarity measure to use for ordinal variables; 2) the estimates of the CUB model parameters; 3) the inability to study the multivariate dependence structure among ordinal variables. With respect to the first point, in this work the Kendall distance has been used because it is suitable for ordinal variables, including Likert-type data. However, it's important to note that the Kendall distance only allows to measure similarities or dissimilarities between items (i.e. ratings) and doesn't take into account the difference in terms of the number of items between two responses. Thus, a different dissimilarity measure for ordinal variables should be considered. Concerning the uncertainty and feeling parameters estimated using the CUB model with covariates, it's important to note that these parameters represent an approximation of individual characteristics. In fact, the individual parameters are computed with minor adjustment over the global parameters based on individual covariates: this results in the impossibility to obtain parameters perfectly suited for the individual (in fact, despite their answer could be different, two people with equal covariates will get the same parameters). Further studies should be devoted to the improvement of this point. Finally, as mentioned in the introduction, the FCMd-CUB model is unable to study the complex multivariate dependence structure among questions and further studies should be devoted to overcome this limitation.

Acknowledgements We sincerely thank prof. Domenico Piccolo and prof. Rosaria Simone for the fruitful discussion that led to the improvement of the present work. We also thank the reviewers for providing comments and suggestions that let us better convey the value of the present work.
This study was carried out within: the BIRD 2022 project titled "Fuzzy theory in Unsupervised Machine Learning algorithm and Sentiment Analysis; the MICS (Made in Italy - Circular and Sustainable) Extended Partnership and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) - MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 - D.D. 1551.11-10-2022, PE00000004); the MOST-Sustainable Mobility National Research Centre and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR)-MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4-D.D. 1033 17 June 2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

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

Declarations

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abdi, H. (2007). The Kendall rank correlation coefficient. *Encyclopedia of measurement and statistics* (pp. 508–510). Sage.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243–256.

Biasetton, N., Disegna, M., Barzizza, E., & Salmaso, L. (2023). A new adaptive membership function with CUB uncertainty with application to cluster analysis of Likert-type data. *Expert Systems with Applications, 213*, 118893.

Campello, R. J., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems, 157*(21), 2858–2875.

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

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55–75.

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

Dolnicar, S. (2019). Market segmentation analysis in tourism: a perspective paper. *Tourism Review Ahead-of-print Ahead-of-Print*.

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). Wiley.

D'Urso, P., & De Giovanni, L. (2014). Robust clustering of imprecise data. *Chemometrics and Intelligent Laboratory Systems, 136*, 58–80.

D'Urso, P., De Giovanni, L., & Massari, R. (2014). Self-organizing maps for imprecise data. *Fuzzy Sets and Systems, 237*, 63–89.

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

D'Urso, P., Giovanni, L. D., Disegna, M., Massari, R., & Vitale, V. (2021). A tourist segmentation based on motivation, satisfaction and prior knowledge with a socio-economic profiling: A clustering approach with mixed information. *Social Indicators Research, 154*(154), 335–360.

D'Urso, P., & Leski, J. M. (2020). Fuzzy clustering of fuzzy data based on robust loss functions and Ordered Weighted Averaging. *Fuzzy Sets and Systems, 389*, 1–28.

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

D'Urso, P. , De Giovanni, L. & Vitale, V. (2022). A robust method for clustering football players with mixed attributes. *Annals of Operations Research* 1–28.

D'Urso, P., Disegna, M., Massari, R., & Prayag, G. (2015). Bagged fuzzy clustering for fuzzy data: An application to a tourism market. *Knowledge-Based Systems, 73*, 335–346.

D'Urso, P., & Vitale, V. (2022). A Kemeny distance-based robust fuzzy clustering for preference data. *Journal of Classification, 39*(3), 600–647.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.
Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research, 45*(1), 43–57.
Gil, M.Á. & González-Rodríguez, G. (2012). Fuzzy vs. Likert scale in statistics. In *Combining experimentation and theory*: *A hommage to abe mamdani* (pp. 407–420). Springer.
Gil, M. A., Lubiano, M., la Rosa, De., de Sàa, S., & Sinova, B. (2015). Analyzing data from a fuzzy rating scale-based questionnaire: A case study. *Psicothema, 27*(2), 182–191.
Gower, J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics 857–871.
Harpe, S. E. (2015). How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning, 7*(6), 836–850.
Hu, H.-Y., Lee, Y.-C., & Yen, T.-M. (2010). Service quality gaps analysis based on fuzzy linguistic servqual with a case study in hospital out-patient services. *The TQM Journal, 22*(5), 499–515.
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*(3), 561–577.
Iannario, M. , Piccolo, D. & Simone, R. (2016). CUB: A class of mixture models for ordinal data. R package version 0.1, available at: https://cran.r-project.org/web/packages/CUB/CUB.pdf, accessed October 29 2016.
Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1212–1218.
Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Wiley.
Kendall, M. G. (1949). Rank correlation methods. *Journal of the Institute of Actuaries, 75*(1), 140–141.
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*(5), 774–792.
Krishnapuram, R., Joshi, A., Nasraoui, O., & Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Transactions on Fuzzy Systems, 9*(4), 595–607.
la Rosa, De., de Sàa, S., Gil, M., Gonzalez-Rodriguez, G., López, M. T., & Lubiano, M. (2015). Fuzzy rating scale-based questionnaires and their statistical analysis. *IEEE Transactions on Fuzzy Systems, 23*(1), 111–126.
Li, Q. (2013). A novel likert scale based on fuzzy sets theory. *Expert Systems with Applications, 40*(5), 1609–1618.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology, 22*(140), 5–55.
Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica, 5*, 85–104.
Piccolo, D., & Simone, R. (2019). The class of cub models: Statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications, 28*(3), 389–435.
Shirahama, N., Watanabe, S., Moriya, K., Koshi, K., & Matsumoto, K. (2021). A new method of subjective evaluation using visual analog scale for small sample data analysis. *Journal of Information Processing, 29*, 424–433.
Simone, R. (2020). FastCUB: Fast EM and best-subset selection for CUB models for rating data. R package, available on CRAN at https://cran.r-project.org/package=FastCUB.
Stavroulakis, P. J., Papadimitriou, S., Tsioumas, V., Koliousis, I. G., Riza, E., & Kontolatou, E. O. (2020). Strategic competitiveness in maritime clusters. *Case Studies on Transport Policy, 8*(2), 341–348.