Fuzzy Optim Decis Making (2017) 16:51–70
DOI 10.1007/s10700-016-9238-8

Exponential distance-based fuzzy clustering for interval-valued data

Pierpaolo D'Urso[1] · **Riccardo Massari**[1] · **Livia De Giovanni**[2] · **Carmela Cappelli**[3]

Published online: 9 March 2016

Abstract In several real life and research situations data are collected in the form of intervals, the so called interval-valued data. In this paper a fuzzy clustering method to analyse interval-valued data is presented. In particular, we address the problem of interval-valued data corrupted by outliers and noise. In order to cope with the presence of outliers we propose to employ a robust metric based on the exponential distance in the framework of the Fuzzy C-medoids clustering mode, the Fuzzy C-medoids clustering model for interval-valued data with exponential distance. The exponential distance assigns small weights to outliers and larger weights to those points that are more compact in the data set, thus neutralizing the effect of the presence of anomalous interval-valued data. Simulation results pertaining to the behaviour of the proposed approach as well as two empirical applications are provided in order to illustrate the practical usefulness of the proposed method.

Keywords Interval-valued data · Outlier interval data · Fuzzy C-medoids clustering · Exponential distance · Robust clustering

1 Introduction

In the literature on data analysis, a great deal of attention is paid to statistical methods for interval-valued data, in different research areas. See, e.g., Denoeux and Masson

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

1 Dipartimento di Scienze Sociali ed Economiche, Sapienza University of Rome, P.le Aldo Moro, 5, 00185 Rome, Italy

2 Dipartimento di Scienze Politiche, LUISS Guido Carli, Viale Romania, 32, 00197 Rome, Italy

3 Dipartimento di Scienze Politiche, Università Federico II di Napoli, Via L. Rodinò, 22, 80138 Naples, Italy

(2000), Coppi and D'Urso (2002), D'Urso and Giordani (2004, 2006), Guru et al. (2004), Carvalho and Lechevallier (2009), Leite et al. (2012), Duarte Silva and Brito (2015).

In particular, in a classical cluster analysis framework different interesting methods have been suggested. Gowda and Diday (1991) proposed a clustering method for symbolic data. Guru et al. (2004) proposed a similarity measure for compare interval-valued data and a modified agglomerative method for clustering symbolic data. Carvalho et al. (2006) proposed a partitional dynamic clustering method for interval data based on adaptive Hausdorff distances. Carvalho and Lechevallier (2009) suggested clustering methods for interval data based on single adaptive distances. Recently an interesting line of research has focused on clustering of interval-valued data based on fuzzy approaches (Carvalho and Tenório 2010) and, in particular, on robust fuzzy clustering methods capable to neutralize the disruptive effects of outlier interval-valued data (D'Urso and Giordani 2006; D'Urso et al. 2015b. For an overview on different robust approaches to fuzzy clustering, refer to D'Urso and De Giovanni (2014).

In this regards, following a noise approach, D'Urso and Giordani (2006) proposed a robust Fuzzy C-means clustering for interval-valued data in which the outliers are assigned to the so-called noise cluster. Recently, D'Urso et al. (2015b) suggested a robust Fuzzy C-medoids clustering based on the trimmed approach, i.e. the clustering procedure is applied to the data after discarding a fixed fraction of outlying data. The "optimum" percentage of data discarded in the clustering process and, thus, not considered in the optimization problem, is determined combining a validity criterion with the trimming algorithm. Starting from the whole data set and having fixed a minimum retention percentage of the objects ($\geq$50 %), the number of clusters and a trimming step, the stopping rule corresponds to the greatest improvement of the validity criterion (Kim et al. 1996; D'Urso et al. 2015b).

In this paper, following the so-called metric approach, we propose a robust fuzzy version of the Partitioning Around Medoids (PAM) for interval-valued data. In particular, our clustering method inherits the advantages of the PAM clustering approach and of the fuzzy theory (D'Urso et al. 2015b) and it is capable to neutralize the negative effects of possible outliers in the dataset by considering a suitable robust metric, i.e. an exponential transformation of the Euclidean distance between interval-valued data.

The paper is organized as follows. In Sect. 2, we describe the robust fuzzy clustering for interval-valued data belonging to the metric approach. In Sect. 3, we present the result of a simulation study while in Sect. 4 we apply our method to two real world cases. In Sect. 5 some final remarks conclude the paper.

2 A robust fuzzy partitioning around medoids method for interval-valued data

There are different real cases in which the empirical information is imprecise, i.e., it is represented by intervals. In particular, we can distinguish the following situations:

– Interval-valued data may occur due to a lack of knowledge, that is when the true value of a variable is unknown and only an interval of values including the true

value is available. Thus, the available information is imprecise and therefore cannot be correctly expressed by means of a single value.
- Interval-valued data may arise as the result of aggregating huge databases, which are impossible, or at least very difficult, to analyse in the original form.
- The data are intrinsically interval-valued, i.e, the phenomena are naturally explained by using intervals. Examples are the monthly temperature in meteorological stations or the daily rate of exchange between euro and dollar or euro and sterling. For example, in case of daily temperatures or daily air pollution levels registered in different places or the mineral concentrations of food items, it could be more interesting to consider the minimum and maximum values registered than the average ones, because they offer more detailed information about the examined phenomenon taking into account the variability of the features involved. Data in which each observation is an interval of values –indicated by a minimum and a maximum– are called interval-valued data. Notice that the intervals not necessarily pertain to the observed maxima and minima, but, for instance, they could pertain to interquartile intervals, or to the middle 90 % the scores (Giordani and Kiers 2004).

An interval-valued datum can be formalized as $x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}]$, $i = 1, \ldots, I$; $j = 1, \ldots, J$, where x_{ij} represents the j-th interval-valued variable observed on the i-th object; $\underline{x}_{ij}$ and $\overline{x}_{ij}$ denote, respectively, the lower and upper bounds of the interval. Each object is represented geometrically by a *hyperrectangle* in $\Re^j$ having 2^J vertices. The 2^J vertices correspond to all the possible (lower bound, upper bound) combinations. In particular, in $\Re$ ($J = 1$) the generic object is represented by a segment; in $\Re^2$ ($J = 2$), it is represented by a rectangle with $2^2 = 4$ vertices, and so on (Cazes et al. 1997).

Alternatively, an interval valued datum can be represented in terms of its midpoint (*center*), $m_{ij} = \frac{\overline{x}_{ij} + \underline{x}_{ij}}{2}$, $i = 1, \ldots, I$; $j = 1, \ldots, J$, and of its radius (*spread*), $r_{ij} = \frac{\overline{x}_{ij} - \underline{x}_{ij}}{2}$, $i = 1, \ldots, I$; $j = 1, \ldots, J$. In this way, the lower and upper bounds of the interval-valued datum can be obtained as $m_{ij} - r_{ij}$ and $m_{ij} + r_{ij}$, respectively.

Thus, by considering the previous reformulation of the interval-valued data, we have: $x_{ij} = (m_{ij}, r_{ij})$; $i = 1, \ldots, I$; $j = 1, \ldots, J$. Note that the center-radius representation is simple and convenient because the range is a common measure of variability of a random variable and it is often employed for estimation purposes in various empirical applications.

Interval-valued data can be corrupted by noise and outliers. In particular, we can distinguish three possible types of outlier interval-valued data, i.e. outlier interval-valued data with outlier midpoint (center) (see, e.g., Fig. 1, case 1), with outlier radius (spread) (see, e.g., Fig. 1, case 2), with outlier midpoint and outlier radius (see, e.g., Fig. 1, case 3).

In order to deal with the above types of outlier, we propose a fuzzy clustering method based on a robust metric.

Let $\{\mathbf{x}_1 = (\mathbf{m}_1, \mathbf{r}_1), \ldots, \mathbf{x}_i = (\mathbf{m}_i, \mathbf{r}_i), \ldots, \mathbf{x}_I = (\mathbf{m}_I, \mathbf{r}_I)\}$ be a set of I vector objects (data matrix) and $\{\widetilde{\mathbf{x}}_1 = (\widetilde{\mathbf{m}}_1, \widetilde{\mathbf{r}}_1), \ldots, \widetilde{\mathbf{x}}_c = (\widetilde{\mathbf{m}}_c, \widetilde{\mathbf{r}}_c), \ldots, \widetilde{\mathbf{x}}_C = (\widetilde{\mathbf{m}}_C, \widetilde{\mathbf{r}}_C)\}$ a subset of the previous set with cardinality C, where $\widetilde{\mathbf{m}}_c$ and $\widetilde{\mathbf{r}}_c$ denote, respectively, the midpoint and radius vectors of the c-th medoid. The Fuzzy C-medoids clustering model

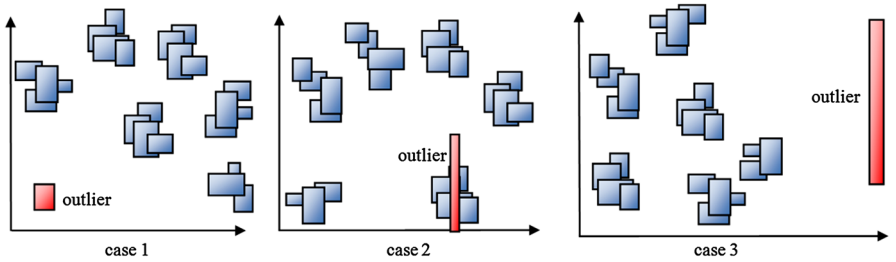

Fig. 1 Examples of different types of outlier interval-valued data in $\Re^2$

for interval-valued data with exponential distance (ExpFCMd-ID) can be formalized as follows:

$$
\begin{aligned}
\min : & \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\, {}_{exp}D^2(\mathbf{x}_i, \widetilde{\mathbf{x}}_c) \\
&= \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\left[1-\exp\left\{-\beta\left(\sum_{j=1}^{2^J}\|(\mathbf{m}_i+\mathbf{r}_i*\mathbf{h}_v)-(\widetilde{\mathbf{m}}_c+\widetilde{\mathbf{r}}_c*\mathbf{h}_v)\|^2\right)\right\}\right] \\
&= \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\left[1-\exp\left\{-\beta\left(2^J\|\mathbf{m}_i-\widetilde{\mathbf{m}}_c\|^2+2^J\|\mathbf{r}_i-\widetilde{\mathbf{r}}_c\|^2\right)\right\}\right] \\
&\approx \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\left[1-\exp\left\{-\beta\left(\|\mathbf{m}_i-\widetilde{\mathbf{m}}_c\|^2+\|\mathbf{r}_i-\widetilde{\mathbf{r}}_c\|^2\right)\right\}\right] \\
\text{s.t:} & \sum_{c=1}^{C} u_{ic}=1, \quad u_{ic}\geq 0
\end{aligned}
\tag{1}
$$

where ${}_{exp}D^2(\mathbf{x}_i, \widetilde{\mathbf{x}}_c)$ is the squared exponential distance between $\mathbf{x}_i = (\mathbf{m}_i, \mathbf{r}_i)$ and $\widetilde{\mathbf{x}}_c = (\widetilde{\mathbf{m}}_c, \widetilde{\mathbf{r}}_c)$ (Wu and Yang 2002; D'Urso and Giordani 2004); the symbol $*$ is the Hadamard product, that is the element-wise product of two matrices (vectors) of the same order; the vectors $\mathbf{h}_v$, $v = 1, \ldots, 2^J$ help us to define every vertex of the hyper-rectangle associated to each object separately, since their elements are equal to ± 1 in order to refer exactly to every vertex; u_{ic} represents the fuzzy membership of the i-th object to the c-th cluster; $m > 1$ is a weighting exponent that controls the fuzziness of the partition; β is a suitable parameter (positive constant) determined according to the variability of the data.

As for the squared exponential distance ${}_{exp}D^2(\mathbf{x}_i, \widetilde{\mathbf{x}}_c)$, we use the exponential version (Wu and Yang 2002) of the distance measure for interval-valued data proposed by D'Urso and Giordani (2004) and successively adopted by D'Urso and Giordani (2006) and D'Urso et al. (2015b). Notice that the exponential distance was adapted to the case of imprecise data by D'Urso and De Giovanni (2014). The exponential distance is a weighted distance that assigns different weights to each data point, according to whether a data point is noisy or not and thus it is more robust to the presence of outliers.

In fact, the exponential distance assigns small weights to outliers and larger weights to those points that are more compact in the data set.

The membership degrees in (1) can be obtained heuristically in many different ways. For instance, following Krishnapuram et al. (2001), we consider the Lagrangian function:

$$\begin{aligned} L_m(u_{ic},\lambda) &= \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m exp D^2(\mathbf{x}_i,\widetilde{\mathbf{x}}_c) - \lambda\left(\sum_{c=1}^{C} u_{ic} - 1\right) \\ &= \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^m\left[1-\exp\left\{-\beta\left(\|\mathbf{m}_i-\widetilde{\mathbf{m}}_c\|^2+\|\mathbf{r}_i-\widetilde{\mathbf{r}}_c\|^2\right)\right\}\right] - \lambda\left(\sum_{c=1}^{C} u_{ic} - 1\right) \end{aligned} \tag{2}$$

and, by taking the partial derivatives and setting them to 0 we obtain:

$$\frac{\partial L_m(\mathbf{u}_i,\lambda)}{\partial u_{ic}} = 0 \Leftrightarrow m u_{ic}^{m-1}\left[1-\exp\left\{-\beta\left(\|\mathbf{m}_i-\widetilde{\mathbf{m}}_c\|^2+\|\mathbf{r}_i-\widetilde{\mathbf{r}}_c\|^2\right)\right\}\right] - \lambda = 0 \tag{3}$$

$$\frac{\partial L_m(\mathbf{u}_i,\lambda)}{\partial \lambda} = 0 \Leftrightarrow \sum_{c=1}^{C} u_{ic} - u = 0. \tag{4}$$

From (3) we obtain:

$$u_{ic} = \left(\frac{\lambda}{m\left[1-\exp\left\{-\beta\left(\|\mathbf{m}_i-\widetilde{\mathbf{m}}_c\|^2+\|\mathbf{r}_i-\widetilde{\mathbf{r}}_c\|^2\right)\right\}\right]}\right)^{\frac{1}{m-1}} \tag{5}$$

and, by substituting (5) in (4), with some algebra:

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{c=1}^{C}\left(\frac{1}{\left[1-\exp\left\{-\beta\left(\|\mathbf{m}_i-\widetilde{\mathbf{m}}_c\|^2+\|\mathbf{r}_i-\widetilde{\mathbf{r}}_c\|^2\right)\right\}\right]}\right)^{\frac{1}{m-1}}}. \tag{6}$$

Finally, substituting (6) in (5) we obtain the iterative solutions:

$$u_{ic} = \frac{1}{\sum_{c'=1}^{C}\left[\frac{1-\exp\left\{-\beta\|\mathbf{m}_i-\widetilde{\mathbf{m}}_c\|^2+\|\mathbf{r}_i-\widetilde{\mathbf{r}}_c\|^2\right\}}{1-\exp\left\{-\beta\|\mathbf{m}_i-\widetilde{\mathbf{m}}_{c'}\|^2+\|\mathbf{r}_i-\widetilde{\mathbf{r}}_{c'}\|^2\right\}}\right]^{\frac{1}{m-1}}}. \tag{7}$$

Notice that when the objective function (1) is minimized, the subset $\widetilde{\mathbf{X}}$ corresponding to the solution provides a fuzzy partition via (7). However, the objective function (1) cannot be minimized by means of the alternating optimization algorithm, because the necessary conditions cannot be derived by differentiating it with respect to the medoids. Nonetheless, following Fu's heuristic algorithm for a crisp version of the objective function in (1), a fuzzy clustering algorithm to obtain a local optimal solution can be retrieved (Krishnapuram et al. 2001). The steps of our clustering procedure are shown in Algorithm 1.

Algorithm 1 FCMd-ID algorithm

1: Fix C, $max.iter$ and β;
2: Set $iter = 0$;
3: Pick initial medoids: $\widetilde{\mathbf{X}} \equiv \{\widetilde{\mathbf{x}}_1 = (\widetilde{\mathbf{m}}_1, \widetilde{\mathbf{r}}_1), \ldots, \widetilde{\mathbf{x}}_c = (\widetilde{\mathbf{m}}_c, \widetilde{\mathbf{r}}_c), \ldots, \widetilde{\mathbf{x}}_C = (\widetilde{\mathbf{m}}_C, \widetilde{\mathbf{r}}_C), \}$;
4: **repeat**
5: Store the current medoids $\widetilde{\mathbf{X}}_{OLD} = \widetilde{\mathbf{X}}$;
6: Compute u_{ic} by using (7);
7: Select the new medoids: $\widetilde{\mathbf{x}}_c = (\widetilde{\mathbf{m}}_c, \widetilde{\mathbf{r}}_c),\ c = 1, \ldots, C$:
8: **for** $c = 1$ to C **do**
9: $q = \text{argmin}_{1 \le i' \le I} \sum_{i''=1}^{I} u_{i''c}^{m} \left[1 - \exp\left\{-\beta \|\mathbf{m}_{i'} - \mathbf{m}_{i''}\|^2 + \|\mathbf{r}_{i'} - \mathbf{r}_{i''}\|^2\right\}\right]$
10: **return** $\Rightarrow \widetilde{\mathbf{x}}_c = \mathbf{x}_q$
11: **end for**
12: $iter \leftarrow iter + 1$;
13: **until** $\widetilde{\mathbf{X}}_{OLD} = \widetilde{\mathbf{X}}$ or $iter = max.iter$

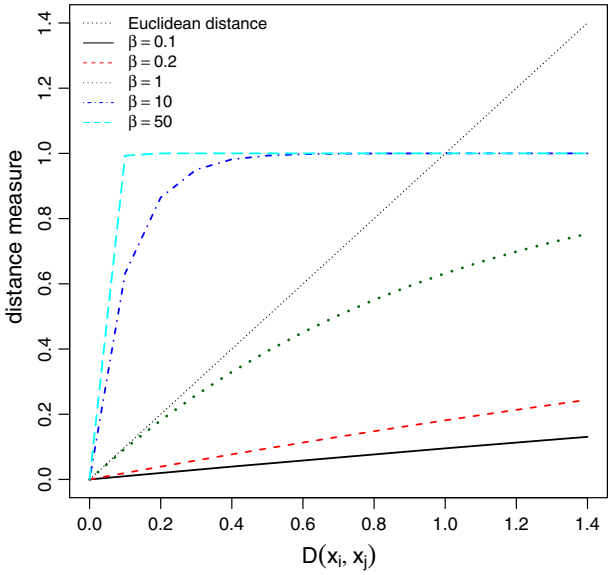

Fig. 2 Effect of β on the exponential distance ${}_{exp}D(\mathbf{x}_i, \mathbf{x}_j)$

The role of the parameter β is crucial both for the distance and for the detection of the membership degrees of each unit to each cluster. Figure 2 shows the effect of increasing values of β on the exponential distance (the 45° dotted line represents the Euclidean distance).

First, it should be noted that the exponential distance is bounded by 1. Second, as the value of β increases, the distance tends more rapidly to its maximum value. Hence, if β is too high, in the classification process each unit is a singleton, since it has no neighbours.

Figure 3 shows different membership curves for different values of β obtained with the ExpFCMd-ID model in the case of two clusters with midpoint of the medoids equal to 0.5 and 0.6, respectively. The curve with circle points represents the membership degrees obtained with the Fuzzy C-medoids clustering model for interval-valued

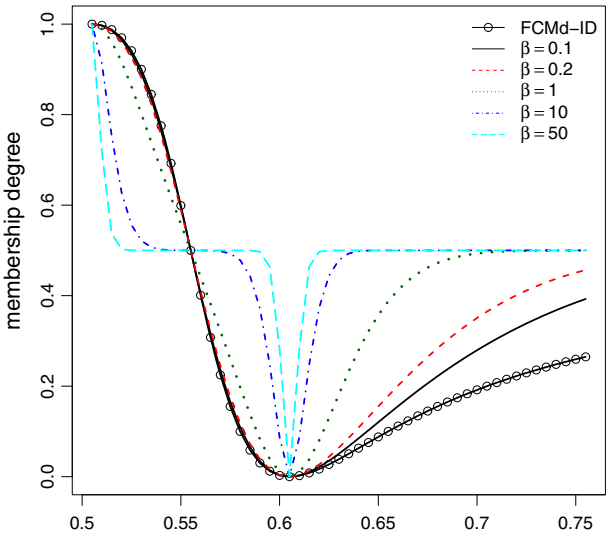

Fig. 3 Effect of the parameter β on the membership degrees (7)

data (FCMd-ID, D'Urso et al. 2015b). If β is very small the ExpFCMd-ID membership curve is very close to the FCMd-ID membership curve, but tends to 0.5 more rapidly as the distance with respect to the medoids increases. As β increases, even for rather small value of β, like $\beta = 1$, the ExpFCMd-ID membership curve is very different from to the FCMd-ID membership curve, and it assigns membership close or equal to 0.5 to data that are only slightly far from the midpoint medoids.

Based on the behaviour of the distance and of the membership degrees for varying values of β, and following Wu and Yang (2002), we set β as the inverse of a measure of the variability of the data:

$$\beta = \left[\frac{\sum_{i=1}^{I} D(\mathbf{x}_i, \widetilde{\mathbf{x}}_q)^2}{I}\right]^{-1} = \left[\frac{\sum_{i=1}^{I} \left(\|\mathbf{m}_i - \widetilde{\mathbf{m}}_q\|^2 + \|\mathbf{r}_i - \widetilde{\mathbf{r}}_q\|^2\right)}{I}\right]^{-1} \tag{8}$$

where

$$\widetilde{\mathbf{x}}_q = (\widetilde{\mathbf{m}}_q, \widetilde{\mathbf{r}}_q) : q = \underset{1 \le i \le I}{\operatorname{argmin}} \sum_{i'=1}^{I} \left(\|\mathbf{m}_i - \widetilde{\mathbf{m}}_{i'}\|^2 + \|\mathbf{r}_i - \widetilde{\mathbf{r}}_{i'}\|^2\right)$$

i.e., $\widetilde{\mathbf{x}}_q = (\widetilde{\mathbf{m}}_q, \widetilde{\mathbf{r}}_q)$ is the unit closest to all other units.

By looking at Fig. 3, if β is large (low variability) the model based on the exponential distance assigns membership degrees values approximately equal to 0.5 to all the units that are not close to the medoids. The case of low variability of data is consistent with the presence of well-separated clusters, and each unit that is not close to the medoids is likely to be an outlier. On the contrary, if β is small (large variability) the model

based on the exponential distance has a behaviour similar to the non-robust FCMdC-ID model, when data are not far away from the medoids, but tends to assign membership degrees approximately equal to 0.5 for units that are distant from the medoids. This case is compatible with the situation in which there are overlapping clusters, but also with the case in which there are well-separated clusters in a very noisy environment, i.e. with a large fraction of data that are outliers with respect to the well-separated clusters.

More in general, when there are C clusters ExpFCMd-ID assigns membership degrees approximately equal to $1/C (c = 1, \ldots, C)$.

See Wu and Yang (2002) for further insights on the robustness of the exponential distance.

3 A simulation study

To investigate the performance of ExpFCMd-ID, a simulation study has been carried out. The proposed model has also been compared with other fuzzy clustering models for interval-valued data, the Fuzzy C-medoids (FCMd-ID, D'Urso et al. 2015b), and two robust models, the Trimmed Fuzzy C-medoids (TrFCMd-ID, D'Urso et al. 2015b), and the Noise Cluster-based Fuzzy C-medoids (NcFCMd-ID), which is the PAM-based version of the Noise Cluster-based Fuzzy C-means model for interval-valued data (NcFCM-ID) proposed by D'Urso and Giordani (2006). Notice that we consider the NcFCMd-ID model to draw a comparison between PAM-based model, since this approach is inherently slightly more robust than the k-means based approach (García-Escudero and Gordaliza 2005).

Three data generation scenarios have been considered. In each scenario the simulated dataset is constructed in such a way that two well-separated clusters ($C = 2$) are generated.

In the *centers scenario*, the radii of the interval-valued data are all randomly generated from $U[0, 1]$ whereas the centers of the data belonging to the first cluster ($I/2$ observations) are drawn from a $U[0, 1]$ and those belonging to the second cluster ($I/2$ observations) from a $U[1.5, 2.5]$. Thus, in the centers scenario, the observation objects are distinguished with respect to the values of the centers.

In the *spreads scenario*, the centers of the interval-valued data are randomly generated from a $U[0, 1]$, whereas the radii of the interval-valued data belonging to the first cluster ($I/2$ observations) are drawn from a $U[0, 1]$ and those belonging to the second cluster ($I/2$ observations) from a $U[1.5, 2.5]$. Therefore, in the spreads scenario, the observation objects are distinguished with respect to the radii.

Finally, in the *centers and spreads scenario*, the centers and the radii of the interval-valued data belonging to the first cluster ($I/2$ observations) are all randomly generated from $U[0, 1]$, whereas the centers and the radii belonging to the second cluster ($I/2$ observations) are drawn from a $U[1.5, 2.5]$.

Each simulated dataset is composed by eighty objects ($I = 80$) and two interval-valued variables ($J = 2$).

For the purpose of evaluating the robustness of the proposed model in presence of outliers, three different percentages of outliers (10, 20 and 30 %) have been added to

the 80 objects, thus yielding to three datasets for each scenario. In the centers scenario, the centers of the outliers are generated from a Gaussian distribution $N(4.5, 2)$ and the values of the radii from a $U[0, 1]$; in the spreads scenario, the radii of the outliers are generated from a Gaussian distribution $N(4.5, 2)$ and the values of the centers from a $U[0, 1]$; in the centers and spreads scenario, both the centers and the radii of the outliers are generated from a Gaussian distribution $N(4.5, 2)$.

For each scenario the data generating process has been replicated 100 times.

The data generation processes are summarized in Table 1. The expected values of the random generated variables for the three scenarios are depicted in Fig. 4. For the sake of completeness, on the x- and y-axis are reported the expected values of the midpoints, while the minimum and maximum of the expected values of the two interval-valued variables are reported near the vertices of the rectangles representing the variables in $\Re^2$.

The main goal of this simulation study is to evaluate the proposed model with respect to two aspects:

Table 1 Data and outlier generation processes for the three scenarios

Data generation scenario	Centers	Spreads
Centers		
Cluster 1 ($i = 1, \ldots, I/2$)	$U[0, 1]$	$U[0, 1]$
Cluster 2 ($i = I/2 + 1, \ldots, I$)	$U[1.5, 2.5]$	$U[0, 1]$
Outliers	$N(4.5, 2)$	$U[0, 1]$
Spreads		
Cluster 1 ($i = 1, \ldots, I/2$)	$U[0, 1]$	$U[0, 1]$
Cluster 2 ($i = I/2 + 1, \ldots, I$)	$U[0, 1]$	$U[1.5, 2.5]$
Outliers	$U[0, 1]$	$N(4.5, 2)$
Centers and spreads		
Cluster 1 ($i = 1, \ldots, I/2$)	$U[0, 1]$	$U[0, 1]$
Cluster 2 ($i = I/2 + 1, \ldots, I$)	$U[1.5, 2.5]$	$U[1.5, 2.5]$
Outliers	$N(4.5, 2)$	$N(4.5, 2)$

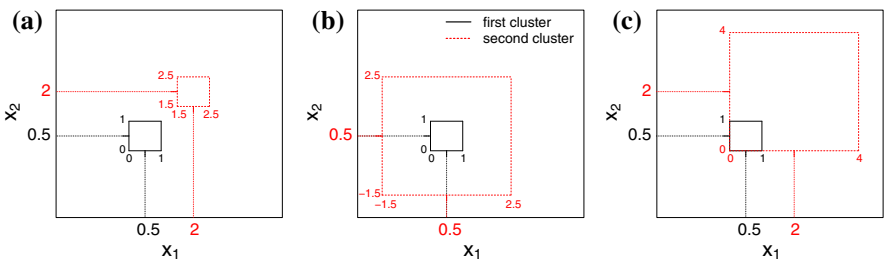

Fig. 4 Expected values of the interval-valued random variables

1. the capability to identify the two equal sized and well-separated clusters generated according to the schemes illustrated in Table 1 when a given percentage of outliers is added to the dataset;
2. the ability to identify cluster prototypes which are not too distant from the "ideal" centers, given by the expected values of the underlying generative random variables, irrespective of the number of outliers added to the dataset.

In particular, to evaluate the robustness of a clustering model with respect to *misclassification* in the presence of outliers, the *Fuzzy Rand Index* (*FRI*, henceforth) (Anderson et al. 2010) has been used. *FRI* allows to compare the "theoretical" hard partitions in two clusters generated with the generation processes illustrated in Table 1 with the fuzzy partitions obtained. *FRI* ranges between 0 and 1. The closer *FRI* is to 1, the better is the classification performance of the model.

To assess the robustness with respect to prototype detection in the presence of outliers, we considered the index of robustness detection, rd, illustrated in D'Urso et al. (2015b) that compares the medoids obtained in the presence of outliers with the ideal centers. The index in case of $\alpha \cdot I$ outliers ($0 < \alpha < 1$) is defined as follows:

$$rd((\tilde{x}_{1,ideal}, \tilde{x}_{2,ideal}), (\tilde{x}_{1,\alpha\cdot I}, \tilde{x}_{2,\alpha\cdot I})) = \frac{d(\tilde{x}_{1,ideal}, \tilde{x}_{1,\alpha\cdot I}) + d(\tilde{x}_{2,ideal}, \tilde{x}_{1,\alpha\cdot I}) + d(\tilde{x}_{1,ideal}, \tilde{x}_{2,\alpha\cdot I}) + d(\tilde{x}_{2,ideal}, \tilde{x}_{2,\alpha\cdot I})}{2 \cdot d(\tilde{x}_{1,ideal}, \tilde{x}_{2,ideal})} \tag{9}$$

where $\tilde{x}_{c,\alpha\cdot I}$ denotes the medoid of cluster c in the case of $\alpha \cdot I$ outliers, $\tilde{x}_{c,ideal}$ the ideal center of cluster c and d the distance for interval-valued data. Notice that $rd((\tilde{x}_{1,ideal}, \tilde{x}_{2,ideal}), (\tilde{x}_{1,\alpha\cdot I}, \tilde{x}_{2,\alpha\cdot I})) \geq 1$, where equality holds only if the two elements of $(\tilde{x}_{1,ideal}, \tilde{x}_{2,ideal})$ are equal to the two elements of $(\tilde{x}_{1,\alpha\cdot I}, \tilde{x}_{2,\alpha\cdot I})$. The more rd departs from 1, the worse is the capability of the model to detect the ideal centers of the clusters.

For each scenario, both indices are averaged over the 100 simulation runs.

We have also set two values of the fuzziness parameter m, 1.5 and 2 respectively, to detect how the clustering performance is affected by this parameter.

Results for both the mean values of *FRI* and *rd*, averaged over the 100 replications of the simulation, are presented in Table 2 with respect to different percentages of outlier and fuzziness parameters.

As it can be seen from Table 2, the average values of *FRI* recorded for ExpFCMd-ID are always very high and close to 1. Another remarkable finding is that the clustering performance of ExpFCMd-ID is slightly affected by the percentage of outliers. This is in line with the discussion at the end of Sect. 2 (see Fig. 3). Notice that as the fraction of outliers increases, the variability of the data increases and so the value of β decreases. In our case, the value of β ranges in (0.1, 0.2), decreasing as the percentage of outliers increases. In such situation, ExpFCMd-ID is capable of detect the two well-separated clusters, while data that are not compact in the two clusters are considered as outliers.

This result hold true even if we increase the percentage of outliers for values higher than 30 %.

In Fig. 5 are reported the average *FRI* values obtained with ExpFCMd-ID when the simulated datasets are contaminated with a percentage of outliers up to 90 %.

Table 2 Performances of the models

Scenario:	Centers				Spreads				Centers and spreads			
	$m = 1.5$		$m = 2.0$		$m = 1.5$		$m = 2.0$		$m = 1.5$		$m = 2.0$	
% Ouliers	*FRI*	*rd*	*FRI*	*rd*	*FRI*	*rd*	*FRI*	*rd*	*FRI*	*rd*	*FRI*	*rd*
FCMd-ID												
10	0.83	1.24	0.74	1.34	0.80	1.26	0.68	1.34	0.88	1.35	0.72	1.34
20	0.70	1.28	0.54	1.36	0.68	1.28	0.58	1.37	0.71	1.38	0.61	1.36
30	0.64	1.29	0.52	1.36	0.57	1.30	0.52	1.39	0.66	1.39	0.50	1.36
ExpFCMd-ID												
10	0.95	1.07	0.80	1.06	0.95	1.08	0.80	1.07	0.99	1.05	0.88	1.04
20	0.96	1.08	0.82	1.08	0.96	1.09	0.82	1.08	0.99	1.06	0.90	1.06
30	0.96	1.11	0.83	1.10	0.96	1.11	0.83	1.10	0.99	1.08	0.90	1.07
TrFCMd-ID												
10	0.92	1.20	0.80	1.24	0.88	1.22	0.76	1.25	0.92	1.31	0.84	1.35
20	0.88	1.22	0.76	1.28	0.86	1.22	0.75	1.28	0.88	1.32	0.84	1.38
30	0.86	1.24	0.74	1.29	0.86	1.25	0.74	1.30	0.86	1.34	0.78	1.39
NcFCMd-ID												
10	0.90	1.21	0.75	1.25	0.86	1.21	0.75	1.26	0.92	1.33	0.80	1.35
20	0.84	1.24	0.73	1.28	0.80	1.26	0.67	1.28	0.82	1.34	0.78	1.37
30	0.80	1.26	0.70	1.30	0.78	1.27	0.65	1.31	0.80	1.36	0.72	1.40

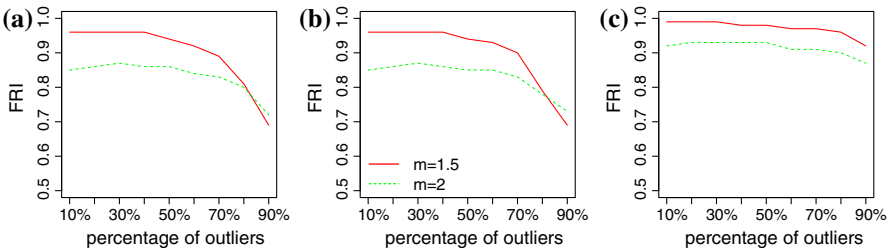

Fig. 5 Classification performance of the ExpFCMd-ID model

It should be noted that ExpFCMd-ID is capable to identify the presence of the two well-separated clusters, even in very noisy environment, and that the classification performance degrades very slowly as the percentage of outliers increases. Indeed, even with 90 % of outliers added to the simulated datasets, the average value of *FRI* is higher than 0.7, which is a remarkable results given the noisiness of the dataset.

A similar pattern can be seen for the capability of detecting prototypes. The values of *rd* obtained with ExpFCMd-ID are always close to 1, irrespective of the scenario, of the percentage of outliers and of the values of the fuzziness parameter. In addition, the value of *rd* increases very slowly as the percentage of outliers increases.

From a comparative viewpoint, ExpFCMd-ID outperforms the non-robust FCMd-ID model, but also the two robust models taken into consideration in this simulation

study, TrFCMd-ID and NcFCMd-ID. Indeed, as for the classification performance, the *FRI* values obtained with ExpFCMd-ID are always higher and closer to 1 than those obtained with the remaining models. By the same token, the rd values recorded with the proposed model are always lower and closer to 1 than those observed with FCMd-ID, TrFCMd-ID and NcFCMd-ID.

4 Applications

4.1 Air pollution data

In this section we illustrate an air quality study based on daily emissions of nitrogen monoxide (NO). NO concentration has been detected in fourteen monitoring stations located in Rome and its surroundings. The list of the stations is reported in the first column of Table 3. Data were collected in 2012, from January 1 December 31.

Data are drawn from the database BRACE[1], which is maintained by ISPRA (Istituto per la Protezione e la Ricerca Ambientale), the Italy's Institute for Environmental Protection and Research.

Daily maximum and minimum values were collected during the period considered. Then these values were averaged over each quarter yielding to four interval-valued variables.

Note that the same data source has been employed in D'Urso et al. (2015a), but the data were differently treated. In D'Urso et al. (2015a) the log-differences of the daily emissions of NO were considered, while in this paper we analysed the average minimum and maximum values of the air pollutant in each quarter. Indeed, the focus in D'Urso et al. (2015a) was on daily rates of change of the pollutant concentration, while in this paper we are interested in the (average) daily excursions of NO concentration.

For comparison's sake we make use both of FCMd-ID and of the proposed robust model ExpFCMd-ID. By adopting the Fuzzy Silhouette criterion, which is an extension to the fuzzy framework of the well-known Silhouette criterion (Campello and Hruschka 2006), two clusters are detected with both models.

In Table 3 are reported the membership degrees of each station to each cluster both for FCMd-ID (second and third columns respectively) and for ExpFCMd-ID (last two columns). As it can be seen, results are similar. Indeed, the values of β is very close to 0 (see Sect. 2). In Fig. 6 are reported the evolution over time observed for the maximum and minimum values of NO concentration (solid lines), which indicates the average daily excursion in each quarter, and the midpoint values (dashed line). With both methods the first cluster, whose medoid is Ciampino (Fig. 6a), is more characterized by non-urban stations or station situated in a park (Malagrotta, Tenuta del Cavaliere and Villa Ada). The daily excursion is rather low in each quarter. Again, the composition of the second cluster is similar with both models, since it comprises stations located in residential areas, like Arenula, Cinecittà, Cipro and L.go Magna Grecia, and the daily excursion is more pronounced especially at the beginning and at the end of the period observed. One difference is that with FCMd-ID the medoid of the

[1] http://www.brace.sinanet.apat.it/web/struttura.html. Data retrieved on 2015-05-03.

Table 3 Membership degrees

	FCMd-ID		ExpFCMd-ID	
	Cluster 1* (Ciampino)	Cluster 2* (L.go Magna Grecia)	Cluster 1* (Ciampino)	Cluster 2* (Cipro)
Arenula	0.090	0.910	0.030	0.970
Bufalotta	0.950	0.050	0.793	0.207
C.so Francia	0.076	0.924	0.462	0.538
Castel Di Guido	0.951	0.049	0.692	0.308
Ciampino	1.000	0.000	1.000	0.000
Cinecittà	0.094	0.906	0.001	0.999
Cipro	0.055	0.945	0.000	1.000
Fermi	0.022	0.978	0.319	0.681
L.go Magna Grecia	0.000	1.000	0.080	0.920
L.go Perestrello	0.076	0.924	0.001	0.999
Malagrotta	0.995	0.005	0.964	0.036
Tenuta Del Cavaliere	0.998	0.002	0.987	0.013
Tiburtina	0.049	0.951	0.374	0.626
Villa Ada	0.993	0.007	0.979	0.021

*Medoid is reported in brackets

second cluster is L.go Magna Grecia (Fig. 6b), while with ExpFCMd-ID it is Cipro (Fig. 6c). The more striking difference, is that C.so Francia is allocated in the second cluster with a high membership degree with FCMd-ID, while with ExpFCMd-ID its membership degrees are approximately equally split across the two clusters (Table 3). This indicates that this station is considered an outlier when one adopts ExpFCMd-ID (see the related discussion at the end of Sect. 2). Indeed, evolution over time of NO excursion recorded in C.So Francia is at odds with the medoids, as it can be seen in Fig. 6d, since both the midpoint values and the daily excursion are higher than for the other station.

Also notice that D'Urso et al. (2015a) found that C.so Francia is an outlier, thus corroborating our findings obtained with ExpFCMd-ID. This is likely due to its characteristics, since it is a very large road, situated in a residential area, but used mainly to access to/depart from Rome centre. In conclusion, our findings, also corroborated by previous analysis, shows that FCMd-ID does not individuate a possible outlier (C.so Francia), whereas ExpFCMd-ID does.

4.2 Bicycle riders data

We analysed data gathered on a sample of Toronto inhabitants. The survey focused particularly on the commuting behaviour, but several socio-demographic character-

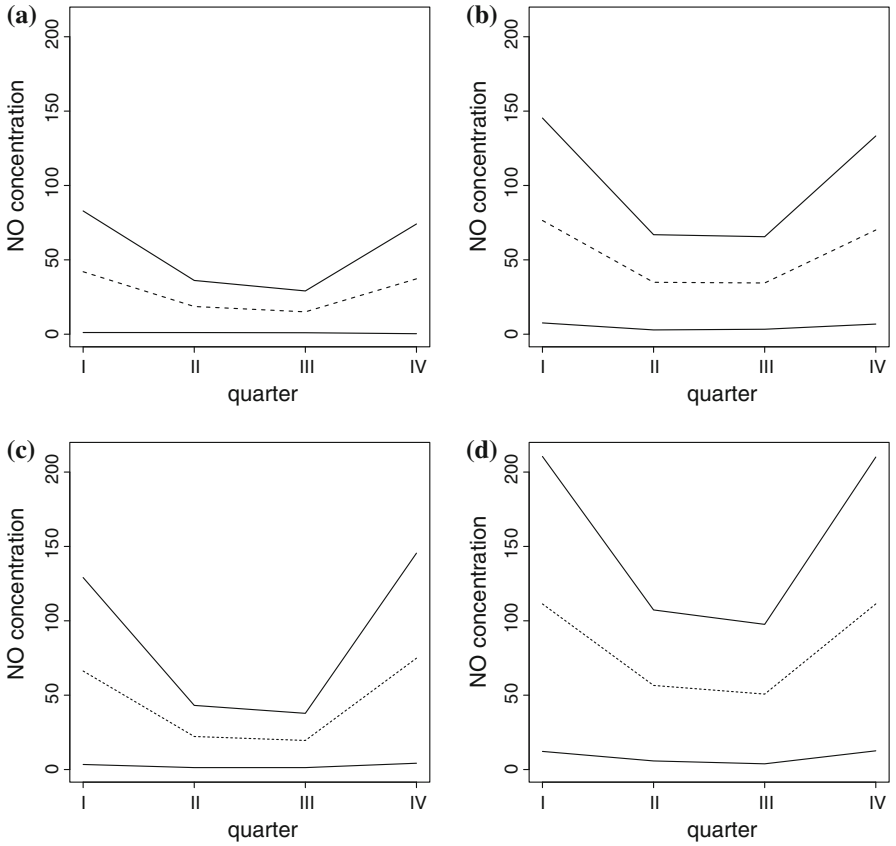

Fig. 6 Medoids and outliers. **a** Ciampino (1st medoid, FCMd-ID and ExpFCMd-ID), **b** L.go Magna Grecia (2nd medoid, FCMd-ID), **c** Cipro (2nd medoid, ExpFCMd-ID), **d** C.so Francia (outlier, ExpFCMd-ID)

istics were collected. Data are freely available on the Toronto Open Data website[2]. The survey provides four interval valued variables, i.e. age, commuting time, commuting distance and household income. Other socio-demographic categorical variables are gender, health status, education degree and working status. In the clustering process we considered only the interval-valued data, while the categorical variable were employed for an ex-post evaluation of the clusters obtained.

We have considered only individuals who live in Central Toronto and use for their commute principally bicycle or e-bike (electric-assisted bicycle).

After dropping individuals with missing or anomalous values for the categorical socio-demographic variables, we ended up with a sample of 458 individuals. The main characteristics of the sample are reported in the first column of Table 4.

As it can be seen, 13.76 % of the sample use e-bike for commuting reasons.

[2] Retrieved at http://www1.toronto.ca/City%20Of%20Toronto/Information%20&%20Technology/Open%20Data/Data%20Sets/Assets/Files/E-Bike_Survey_Responses.xls, on 2015-05-03.

Table 4 Interval-valued variables and socio-demographic characteristics of the respondents (percentages)

	Sample	Cluster 1	Cluster 2	Cluster 3
Transport				
Bicycle	86.24	89.39	88.84	75.89
E-bike	13.76	10.61	11.16	24.11
Interval-valued variables				
Age				
18–34	50.44	93.65	15.23	17.06
35–49	38.43	4.92	67.38	61.76
50–64	11.14	1.43	17.39	21.18
Distance (in km)				
0–2	8.30	7.45	7.12	11.81
5–10	48.03	51.76	49.66	38.00
10–20	34.28	33.60	34.05	35.99
20–35	9.39	7.19	9.16	14.19
Time (in min.)				
0–15	22.05	25.07	19.40	20.03
16–25	48.91	50.19	49.77	44.99
30–44	20.52	18.82	22.12	21.53
45–60	8.52	5.92	8.72	13.45
Household income (in 1000 Canadian $)				
0–20	7.21	10.13	1.71	9.72
20–39	20.52	31.38	3.94	24.00
40–59	23.14	26.50	7.01	41.10
60–79	26.86	21.86	41.37	14.70
80–99	22.27	10.12	45.98	10.48
Socio-demographic characteristics				
Gender				
Female	36.46	37.98	36.33	33.61
Male	63.54	62.02	63.67	66.39
Health status				
Excellent	67.25	71.89	66.91	58.38
Very good	25.33	22.06	26.81	29.65
Poor or fair	7.42	6.04	6.28	11.97
Education degree				
University degree	42.58	45.95	45.11	31.87
Post graduate	24.24	26.66	24.27	19.28
High school or other	33.19	27.39	30.62	48.85

Table 4 continued

	Sample	Cluster 1	Cluster 2	Cluster 3
Working status				
Employed	69.43	72.19	72.77	58.73
Self employed	19.65	12.21	22.44	30.41
Other	10.92	15.60	4.78	10.86

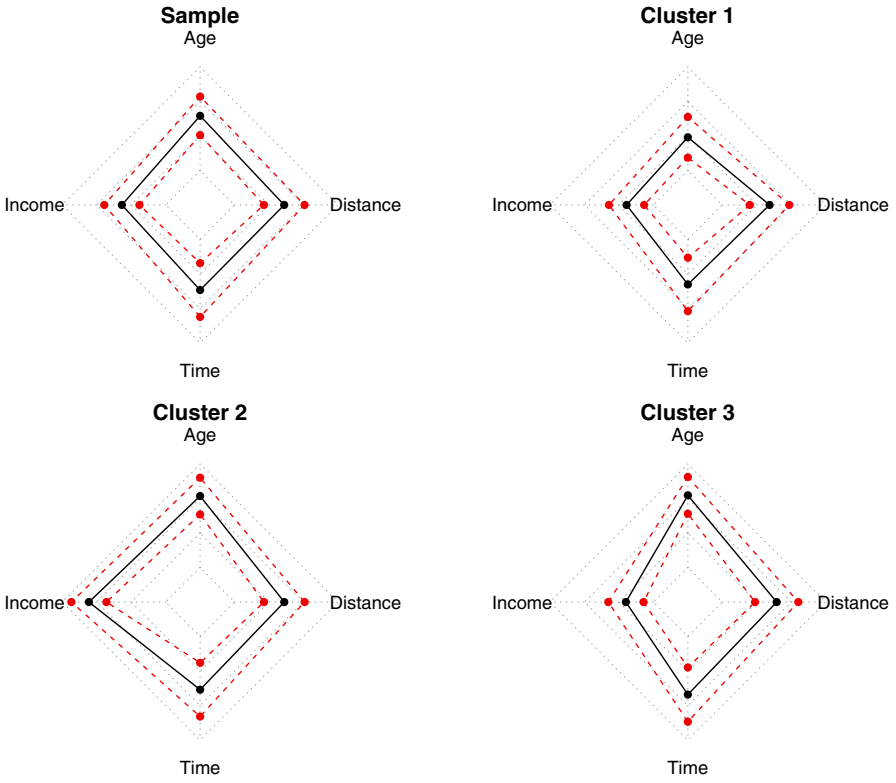

Fig. 7 Radar plots for interval-valued variables (mean values)

As for the interval-valued variables, most of the respondents are aged between 18 and 34 years, commute for relatively short distance and time, and they are almost equally distributed across income classes, apart the first class, in which falls only 7 % of the sample.

Concerning the categorical socio-demographic variables, the sample is composed mainly by individuals with an excellent health status (which is likely related to the transport used), with an university degree and employee.

The mean values of the four interval-valued variable are displayed in the first panel of Fig. 7 ("Sample", the black points, connected with solid lines, refer to the midpoint

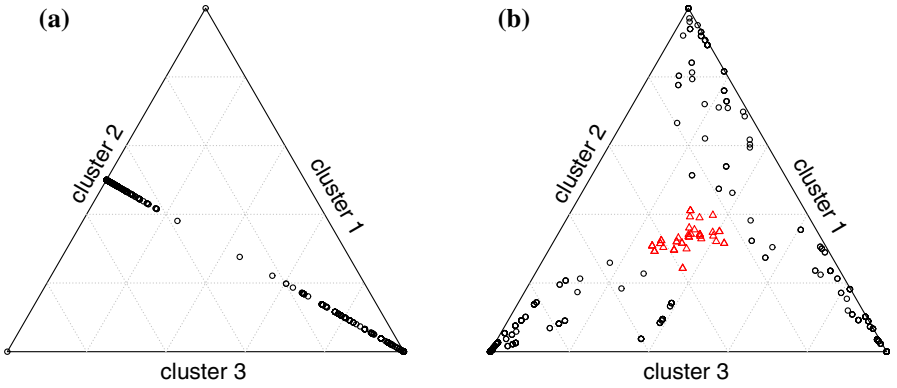

Fig. 8 Ternary plots for FCMD-ID and ExpFCMd-ID

of each variable, while the red points, connected with dashed lines, refer to the radii). This figure is reported as a benchmark for the corresponding figures for the clusters.

Both FCMd-ID and ExpFCMd-ID were applied to the dataset containing the four interval-valued variables. The value of β obtained by applying ExpFCMd-ID is equal to 0.2, thus indicating a large variability in the data and, possibly, a noisy environment. By adopting the Fuzzy Silhouette criterion, three clusters are identified when applying the ExpFCMd-ID to data. In Fig. 8 are reported the membership degrees obtained with both FCMd-ID (Fig. 8a) and ExpFCMd-ID (Fig. 8b). When FCMd-ID is used the three clusters partition is almost identical to the two clusters partition, as it can be seen from the fact that the points representing the membership degrees are aligned on a straight line (see Fig. 8a). The only difference between the partitions with two and three clusters is that in the latter one unit is "forced" to be the medoid of the added cluster. For this reason, we no longer discuss the results of FCMd-ID.

Figure 8b shows the membership degrees for the three clusters solution obtained with ExpFCMd-ID. The red triangle shaped dots represent the outliers, i.e. the individuals for which the membership degrees are approximately equally split across clusters. The presence of these outlier is likely to prevent the non-robust FCMd-ID clustering model to detect a sensible result with three clusters.

As previously said, Fig. 7 displays the radar plots computed on the interval-valued variables, for the whole sample and for the three clusters. The values reported for the three clusters are the weighted mean of the midpoints and of the radii of each variables, with weights given by the membership degrees.

With respect to the whole sample, the first cluster is composed mainly by younger individuals, with lower income and who travel for shorter distances. Clusters 2 and 3 are more similar, with the exception of the average household income.

These results are also confirmed by the values reported in the last three columns of Table 4. These values are the weighted percentage of individuals in each cluster for each category of the socio-demographic variables observed on the sample. The weighted percentage are computed by summing the membership degrees of the individuals with a given attribute, divided by the sum of membership degrees for the considered cluster.

Table 4 allows also for a further insights about the socio-demographic characteristics of individuals in each cluster. As it can be seen, individuals in the first cluster are characterised by an excellent health status, most of them have an university degree and are employees. Conversely, in the third cluster there is a higher (weighted) percentage of individuals with high school diploma (or similar undergraduate degree) and there are more self employed workers. The second cluster presents intermediate values between the first and the third for most of the categories of the socio-demographic variables.

5 Final remarks

In several real life and research situations data are collected in the form of intervals. To analyze interval-valued data, usually researchers summarize the original data into single values, such as the centers or the medians of the intervals, but by doing so some important information in the original data may be lost. Indeed, in the last years several efforts have been made to extend methods or develop new approaches to analyse these type of data taking into account their interval structure.

In this paper we have addressed the problem of clustering interval-valued data corrupted by noise and outliers considering the Fuzzy C-medoids model. To deal with the presence of outliers we have implemented a robust metric based on the exponential distance that assigns small weights to outliers and larger weights to those points that are more compact in the data set.

We have presented the results of simulation studies pertaining to three possible types of outlier interval-valued data: outlier midpoint (center)–inlier radius (spread); inlier midpoint–outlier radius; outlier midpoint–outlier radius.

The proposed model has been evaluated by comparison with other fuzzy clustering models for interval-valued data considering two aspects: the capability of identifying the natural clusters (even when the simulated datasets are contaminated with an increasing number of outliers) and the ability to identify cluster prototypes which are not too distant from the "ideal" centers, given by the expected values of the generative random variables, irrespective of the number of outliers. Results show that the proposed approach is more able to distinguish the natural clusters as well as to detect prototypes.

Eventually, we have analysed two real interval-valued data sets concerning the emissions of nitrogen monoxide detected in fourteen monitoring stations located in Rome and its surroundings and the commuting behaviour of bicycle riders that live in the centre of Toronto and use for their commute principally bicycle or e-bike (electric-assisted bicycle).

In both cases the proposed approach has shown to be useful in identifying the "natural" clusters even in presence of outliers.

The use of the robust metric in the framework of regression trees based method for change point detection in interval-valued time series affected by outliers is the subject of ongoing research.

Also the adoption of the exponential distance for self-organizing maps is worth exploring in future research.

Finally, the utilization of the exponential transformation of other types of distance measures (Xu 2012) in a fuzzy partitioning around medoids procedure or in other typologies of fuzzy clustering, e.g. entropy-based fuzzy clustering (Dey et al. 2011) and bi-objective fuzzy clustering (Hung 2007), will be considered for the interval-valued data case.

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

References

Anderson, D. T., Bezdek, J. C., Popescu, M., & Keller, J. M. (2010). Comparing fuzzy, probabilistic, and possibilistic partitions. *IEEE Transactions on Fuzzy Systems*, *18*(5), 906–918.

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.

Cazes, P., Chouakria, A., Diday, E., & Schektrman, Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique Appliquée*, *45*(3), 5–24.

Coppi, R., & D'Urso, P. (2002). Fuzzy k-means clustering models for triangular fuzzy time trajectories. *Statistical Methods and Applications*, *11*(1), 21–40.

De Carvalho, Fd A T, & Lechevallier, Y. (2009). Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition*, *42*(7), 1223–1236.

De Carvalho, Fd A T, & Tenório, C. P. (2010). Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems*, *161*(23), 2978–2999.

De Carvalho, Fd A T, De Souza, R. M., Chavent, M., & Lechevallier, Y. (2006). Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, *27*(3), 167–179.

Denoeux, T., & Masson, M. (2000). Multidimensional scaling of interval-valued dissimilarity data. *Pattern Recognition Letters*, *21*(1), 83–92.

Dey, V., Pratihar, D. K., & Datta, G. L. (2011). Genetic algorithm-tuned entropy-based fuzzy c-means algorithm for obtaining distinct and compact clusters. *Fuzzy Optimization and Decision Making*, *10*(2), 153–166.

Duarte Silva, A. P., & Brito, P. (2015). Discriminant analysis of interval data: An assessment of parametric and distance-based approaches. *Journal of Classification*, *32*(3), 516–541. doi:10.1007/s00357-015-9189-8.

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

D'Urso, P., & Giordani, P. (2004). A least squares approach to principal component analysis for interval valued data. *Chemometrics and Intelligent Laboratory Systems*, *70*(2), 179–192.

D'Urso, P., & Giordani, P. (2006). A robust fuzzy k-means clustering model for interval valued data. *Computational Statistics*, *21*(2), 251–269.

D'Urso, P., De Giovanni, L., & Massari, R. (2015a). Time series clustering by a robust autoregressive metric with application to air pollution. *Chemometrics and Intelligent Laboratory Systems*, *141*, 107–124.

D'Urso, P., De Giovanni, L., & Massari, R. (2015b). Trimmed fuzzy clustering for interval-valued data. *Advances in Data Analysis and Classification*, *9*(1), 21–40.

García-Escudero, L. A., & Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, *22*(2), 185–201.

Giordani, P., & Kiers, H. A. (2004). Three-way component analysis of interval-valued data. *Journal of Chemometrics*, *18*(5), 253–264.

Gowda, K. C., & Diday, E. (1991). Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, *24*(6), 567–578.

Guru, D. S., Kiranagi, B. B., & Nagabhushan, P. (2004). Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. *Pattern Recognition Letters*, *25*(10), 1203–1213.

Hung, T. W. (2007). The bi-objective fuzzy c-means cluster analysis for tsk fuzzy system identification. *Fuzzy Optimization and Decision Making*, *6*(1), 51–61.

Kim, J., Krishnapuram, R., & Davé, R. (1996). Application of the least trimmed squares technique to prototype-based clustering. *Pattern Recognition Letters*, *17*(6), 633–641.

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.

Leite, D., Ballini, R., Costa, P., & Gomide, F. (2012). Evolving fuzzy granular modeling from nonstationary fuzzy data streams. *Evolving Systems*, *3*(2), 65–79.

Wu, K. L., & Yang, M. S. (2002). Alternative c-means clustering algorithms. *Pattern Recognition*, *35*(10), 2267–2278.

Xu, Z. (2012). Fuzzy ordered distance measures. *Fuzzy Optimization and Decision Making*, *11*(1), 73–97.