Annals of Operations Research
https://doi.org/10.1007/s10479-023-05180-1

ORIGINAL RESEARCH

Fuzzy clustering with entropy regularization for interval-valued data with an application to scientific journal citations

Pierpaolo D'Urso[1] · Livia De Giovanni[2] · Leonardo Salvatore Alaimo[1] · Raffaele Mattera[1] · Vincenzina Vitale[1]

Accepted: 5 January 2023

Abstract
In recent years, the research of statistical methods to analyze complex structures of data has increased. In particular, a lot of attention has been focused on the interval-valued data. In a classical cluster analysis framework, an interesting line of research has focused on the clustering of interval-valued data based on fuzzy approaches. Following the partitioning around medoids fuzzy approach research line, a new fuzzy clustering model for interval-valued data is suggested. In particular, we propose a new model based on the use of the entropy as a regularization function in the fuzzy clustering criterion. The model uses a robust weighted dissimilarity measure to smooth noisy data and weigh the center and radius components of the interval-valued data, respectively. To show the good performances of the proposed clustering model, we provide a simulation study and an application to the clustering of scientific journals in research evaluation.

Keywords Fuzzy clustering · Robust clustering · Weighting system · Interval-valued data · Entropy · Journal citation indices · Evaluation

✉ Livia De Giovanni
ldegiovanni@luiss.it

Pierpaolo D'Urso
pierpaolo.durso@uniroma1.it

Leonardo Salvatore Alaimo
leonardo.alaimo@uniroma1.it

Raffaele Mattera
raffaele.mattera@uniroma1.it

Vincenzina Vitale
vincenzina.vitale@uniroma1.it

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

[2] Department of Political Sciences and Data Lab, Luiss University, Viale Romania, 32-00197 Rome, Italy

Published online: 02 March 2023

1 Introduction

In recent years, the research of statistical methods to analyze complex structures of data has increased. In particular, a lot of attention has been focused on the interval-valued data (Denoeux & Masson, 2000; D'Urso & De Giovanni, 2014; D'Urso & Leski, 2016).

In the literature on data analysis, a great deal of attention is paid to statistical methods to treat interval-valued data, in different research areas (Coppi et al., 2006; Denoeux & Masson, 2000; D'Urso & Giordani, 2005; Giordani & Kiers, 2004; D'Urso & Leski, 2016; D'Urso & De Giovanni, 2014).

In a classical cluster analysis framework, a variety of interesting methods have been suggested. In particular, Gowda and Diday (1991) hinted a clustering method for symbolic data; Guru et al. (2004) proposed a similarity measure to compare interval-valued data and a modified agglomerative method for clustering symbolic data. De Carvalho and Lechevallier (2009) proposed a partitional dynamic clustering method for interval data based on adaptive Hausdorff distances; De Carvalho et al. (2006) suggested clustering methods for interval data based on single adaptive distances.

An interesting line of research has focused on the clustering of interval-valued data based on fuzzy approaches, where the weighting exponent m controls the extent of membership sharing between fuzzy clusters (De Carvalho & Tenório, 2010; Denoeux & Masson, 2000; D'Urso et al., 2015b; D'Urso & Giordani, 2006a; D'Urso et al., 2017). Li and Mukaidono (1995) remarked that this unusual parameter is unnatural and doesn't have a physical meaning. The parameter m may be removed in the objective function of the clustering model; when this is the case, the procedure cannot generate the membership update equations (Coppi & D'Urso, 2006). For this purpose, Li and Mukaidono (1995, 1999) suggested a new approach to fuzzy clustering by proposing the so-called Maximum Entropy Inference Method. The underlying idea is presented in the paper by Miyamoto and Mukaidono (1997), where the trade-off between fuzziness and compactness is dealt with by introducing a unique objective function reformulating the maximum entropy method in terms of regularization of the Fuzzy c-Means (FCM) function.

In the literature, many authors proposed the entropy-based approach as a regularization in fuzzy clustering modeling. In particular, Yao et al. (2000) proposed an entropy-based fuzzy clustering method which automatically identifies the number and initial locations of cluster centers. Successively, it removes all data points having dissimilarity larger than a threshold with the chosen cluster center; the procedure is repeated until all data points are removed. Ichihashi (2000) and Miyagishi et al. (2000) suggested a generalized objective function with additional variables. These authors consider a covariance matrix and show an equivalence between their Kullback–Leibler (KL) fuzzy clustering and the Gaussian mixture model. The method of fuzzy clustering using the KL information is called entropy-based method of FCM. Ménard and Eboueya (2002) suggested an axiomatic derivation of the Maximum Entropy Inference (and also of the possibilistic) clustering approach, based on a unifying principle of physics, that of Extreme Physical Information (EPI) defined by Frieden and Binder (2000). Coppi and D'Urso (2006) suggested fuzzy unsupervised clustering models based on Shannon entropy regularization to classify time-varying data. Zarinbal et al. (2014) proposed a new fuzzy clustering method based on FCM and the relative entropy is added to the objective function as a regularization function to maximize the dissimilarity between clusters. Kahali et al. (2019) presented an entropy-based FCM segmentation method that incorporates the uncertainty of classification of individual pixels within the classical framework of FCM. Gao et al. (2019) showed a novel method considering noise intelligently based on the existing

FCM approach, called adaptive-FCM and its extended version (adaptive-REFCM) in combination with relative entropy. More recently, Ashtari et al. (2020) proposed an entropy-based regularization approach to fuzzify the partition and to weight features, enabling the method to capture more complex patterns, identify significant features, and yield better performance facing high-dimensional data.

Note that the models cited above utilizing entropy-based regularization regard ordinary point data.

Following this line of research, in this paper a new robust fuzzy clustering model for interval-valued data with entropy as a regularization function is proposed. The model is named Robust Entropy-based Fuzzy c-Medoids clustering for interval-valued data (EFCMd-ID).

The paper is organized as follows. In Sect. 2.1, the basic notation and the family of robust dissimilarity measures for interval-valued data are described; in Sect. 2.2, the motivation of the use of the Shannon entropy regularization in fuzzy clustering is discussed. Then in Sect. 2.3, the modeling details and the algorithm of the proposed EFCMd-ID model for interval-valued data along with the Robust Entropy-based Fuzzy c-Means clustering variant (EFCM-ID) are presented. In Sect. 3, a detailed simulation study and comparison with other fuzzy and not fuzzy clustering models for interval-valued data is proposed. In Sect. 4, the results obtained by the application of the EFCMd-ID model on empirical data are shown. In Sect. 5, some concluding remarks and the lines for future research are provided.

2 Robust entropy-based fuzzy *c*-medoids clustering for interval-valued data (robust EFCMd-ID model)

2.1 Robust dissimilarity measure for interval-valued data

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} indicates 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, i.e., they represent the minimum and maximum values of the j-th interval-valued variable with respect to the i-th object. Each object is represented geometrically by a hyper-rectangle in $\Re^J$ having 2^J vertices. All 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).

Then, assuming J interval-valued variables are observed on I objects, the entire dataset can be stored in the so-called *interval-valued matrix* as follows:

$$\mathbf{X} \equiv \{x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}] : i = 1, \ldots, I;\ j = 1, \ldots, J\}. \tag{1}$$

By denoting with

$$\mathbf{M} \equiv \left\{ m_{ij} = \frac{\overline{x}_{ij} + \underline{x}_{ij}}{2} : i = 1, \ldots I;\ j = 1, \ldots, J \right\}, \tag{2}$$

the *midpoint matrix* (*center matrix*), where m_{ij} is the midpoint (center) of the associated interval value for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, and with

$$\mathbf{R} \equiv \left\{ r_{ij} = \frac{\overline{x}_{ij} - \underline{x}_{ij}}{2} : i = 1, \ldots, I;\ j = 1, \ldots, J \right\}, \tag{3}$$

the *radius matrix*, where r_{ij} is the radius (spread) of the associated interval for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, we can reformulate the interval-valued matrix (1) as follows:

$$\tilde{\mathbf{X}} \equiv \{\tilde{x}_{ij} = [m_{ij}, r_{ij}] : i = 1, \ldots, I;\ j = 1, \ldots, J\} = \{\tilde{\mathbf{x}}_i = [\mathbf{m}_i, \mathbf{r}_i] : i = 1, \ldots, I\}. \tag{4}$$

where $\mathbf{m}_i$ and $\mathbf{r}_i$ denote, respectively, the i-th row of $\mathbf{M}$ and $\mathbf{R}$.

Then, $\tilde{x}_{ij} = [m_{ij}, r_{ij}]$ represents an alternative formalization of the interval-valued datum $x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}]$. In this way, the lower and upper bounds of the interval-valued datum can be obtained as $\underline{x}_{ij} = m_{ij} - r_{ij}$ and $\overline{x}_{ij} = m_{ij} + r_{ij}$, respectively.

The generic interval-valued datum pertaining to the i-th object with respect to the j-th interval-valued feature can be shown as the pair (m_{ij}, r_{ij}), $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where m_{ij} denotes the midpoint and r_{ij} denotes the radius of the interval.

In the literature, several metrics have been suggested for interval-valued data. In this paper, we adopt a robust weighted dissimilarity measure.

The robustness of the dissimilarity measure for interval-valued data is obtained by considering the exponential version (Wu & Yang, 2002; Zhang & Chen, 2004) of the distance measure for interval-valued data proposed by D'Urso and Giordani (2004) and successively adopted by D'Urso et al. (2017).

The dissimilarity measure is weighted as the dissimilarity between each pair of objects is measured by separately considering the midpoints and the radii of the interval-valued data and using a suitable weighting system for such components (D'Urso & Giordani, 2006b).

In formula, the robust weighted dissimilarity measure between objects i and i' is:

$$d^2_{exp}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'}) = \left\{1 - exp\left[-\beta[w_m^2 d^2(\mathbf{m}_i, \mathbf{m}_{i'}) + w_r^2 d^2(\mathbf{r}_i, \mathbf{r}_{i'})\right]\right\} \tag{5}$$

where $d^2(\mathbf{m}_i, \mathbf{m}_{i'}) = \|\mathbf{m}_i - \mathbf{m}_{i'}\|^2$ is the squared Euclidean distance between the midpoints and $d^2(\mathbf{r}_i, \mathbf{r}_{i'}) = \|\mathbf{r}_i - \mathbf{r}_{i'}\|^2$ is the squared Euclidean distance between the radii, while w_m and w_r are the weights for the midpoint component and the radius component, respectively, and $\beta > 0$.

The exponential dissimilarity measure (5) assigns small weights to noisy objects and large weights to those objects that are more compact in the data set (Wu & Yang, 2002), and it is superiorly bounded by 1.

Following Wu and Yang (2002), β is set as the inverse of the variability of the data:

$$\beta = \left(\frac{\sum_{i=1}^{I} d^2(\mathbf{m}_i, \mathbf{m}_q) + d^2(\mathbf{r}_i, \mathbf{r}_q)}{I}\right)^{-1} \tag{6}$$

where $\mathbf{m}_q, \mathbf{r}_q$ is the unit closest to all other units.

See Wu and Yang (2002), D'Urso et al. (2015a) and D'Urso et al. (2017) for further insights on the robustness of the exponential distance and on the role of β.

Moreover, we assume the following conditions: (*i*) $w_m + w_r = 1$ (*normalization condition*) and (*ii*) $w_m \geq w_r \geq 0$ (*coherence condition*).

The *coherence condition* excludes that the radius component, which represents the uncertainty around the midpoint of the interval-valued data, has more importance than the midpoint component.

The *normalization condition* assesses, in a comparative fashion, the contributions of the midpoint and radius components to the dissimilarity measure computation.

2.2 Shannon entropy regularization in a fuzzy clustering framework

We focus on the entropy regularization approach in a fuzzy clustering framework. It is known that the maximum entropy principle, as applied to fuzzy clustering, provides a new perspective on facing the problem of fuzzifying the clustering of the objects, whilst ensuring the maximum compactness of the obtained clusters (Coppi & D'Urso, 2006; Gao et al., 2019). The first objective is achieved by maximizing the entropy (and, therefore, the uncertainty) of the assignment of the objects into the clusters. The Shannon entropy measure is employed in the objective function of the Fuzzy c-Medoids or Fuzzy c-Means model to deal with the uncertainty of the clustering. The second objective is obtained by minimizing the overall distance of the objects from the cluster prototypes (i.e. to maximize cluster compactness).

The trade-off between fuzziness and compactness is dealt with by introducing a unique objective function reformulating the maximum entropy method in terms of "regularization" of the Fuzzy c-Means objective function (Miyamoto & Mukaidono, 1997; Kahali et al., 2019) and of the Fuzzy c-Medoids objective function.

The novelty of the proposal is the use of entropy regularization for fuzzy clustering of interval-valued data.

Additionally, given the nature of the data (i.e., interval-valued), a weighted dissimilarity measure proposed by D'Urso and Giordani (2006b) is adopted. Here, the dissimilarity between each pair of objects is measured by separately considering the midpoints and the radii of the interval-valued data and using a suitable weighting system for such components.

2.3 Modeling

2.3.1 Robust entropy-based fuzzy c-medoids clustering (EFCMd-ID) model

Let $\mathbf{X}$ be an $I \times J$ interval-valued data matrix. Given the dissimilarity measure shown in Eq. (5), in which we assume that the weights (i.e., w_m and w_r) are objectively computed during the clustering process. We have set $w_m = (1 - w)$ and $w_r = w$. In this way, the *normalization condition* is satisfied and the *coherence condition* turns to $0 \leq w \leq 0.5$. Following a Partitioning Around Medoid (PAM) approach (Kaufmann & Rousseeuw, 1987), the Robust Entropy-based Fuzzy c-Medoids clustering (EFCMd-ID) model is characterized as follows:

$$\begin{aligned}
&\text{min: } J_{EFCMd-ID}(\mathbf{U}, \tilde{\mathbf{X}}, w) = \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic} d^2_{exp}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_c) + p \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic} log(u_{ic}) = \\
&\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic} \left\{1 - exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)]\right]\right\} + p \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic} log(u_{ic}) \\
&\sum_{c=1}^{C} u_{ic} = 1, u_{ic} \geq 0 \\
&0 \leq w \leq 0.5
\end{aligned} \tag{7}$$

where u_{ic} indicates the membership degree of the i-th unit in the c-th cluster and $\mathbf{U}$ is the related $I \times C$ matrix; $d^2_{exp}(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_c)$ is the squared version of Eq. (5) between the i-th unit and the medoid in the c-th cluster; $\mathbf{m}_i$ and $\mathbf{r}_i$ are the midpoints and radii of the i-th unit, respectively; $\tilde{\mathbf{m}}_c$ and $\tilde{\mathbf{r}}_c$ are the medoids of the midpoints and radii in the c-th cluster,

respectively; $p \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic} log(u_{ic})$ is the *fuzzy entropy function*; p is a factor called *degree of fuzzy entropy* that represents the extent of fuzziness uncertainty of the partition (Coppi & D'Urso, 2006; Li & Mukaidono, 1995, 1999).

By solving the constrained quadratic minimization problem shown in Eq. (7) *via* the Lagrangian multiplier method, we obtain the optimal solutions u_{ic} and w. In particular, by considering the following Lagrangian function:

$$L_m(u_{ic}, \lambda, w) = \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic} \left\{1 - exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)\right]\right\} + \\ + p \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic} log(u_{ic}) - \lambda \left(\sum_{c=1}^{C} u_{ic} - 1 \right) \tag{8}$$

and setting the first partial derivatives with respect u_{ic} and λ equal to zero, we obtain:

$$\frac{\partial L_m(u_{ic}, \lambda, w)}{\partial u_{ic}} = 0 \Leftrightarrow 1 - exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)\right]] \\ + p(log(u_{ic}) + 1) - \lambda = 0 \tag{9}$$

$$\frac{\partial L_m(u_{ic}, \lambda, w)}{\partial \lambda} = 0 \Leftrightarrow \sum_{c=1}^{C} u_{ic} - 1 = 0. \tag{10}$$

From Eq. (9), we obtain:

$$log(u_{ic}) = \frac{1}{p}\left[\lambda - \left\{1 - exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)\right]\right\} - 1\right] \tag{11}$$

and then

$$u_{ic} = \exp\left\{\frac{\lambda}{p} - \frac{1}{p}\left\{1 - exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)\right]\right\} - 1\right\}. \tag{12}$$

By considering Eq. (10):

$$exp\left(\frac{\lambda}{p} - 1\right) = \frac{1}{\sum_{c=1}^{C} exp\left[-\frac{1}{p}[1 - exp[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)]]\right]} \tag{13}$$

and by replacing Equation (13) in Equation (12), we obtain:

$$u_{ic} = \frac{exp\left[-\frac{1}{p}[1 - exp[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)]]\right]}{\sum_{c'=1}^{C} \exp\left[-\frac{1}{p}[1 - exp[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_{c'}) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_{c'})]]\right]}. \tag{14}$$

The normalization condition for w is implicitly satisfied. To take into account the *coherence condition*, we derive with respect to w and select the minimum between the obtained value and 0.5:

$$\frac{\partial L_m(u_{ic}, \lambda, w)}{\partial w} = 0$$

$$w = \frac{\sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic} d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)]\right]}{\sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic} (d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)) exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \tilde{\mathbf{m}}_c) + w^2 d^2(\mathbf{r}_i, \tilde{\mathbf{r}}_c)]\right]}. \tag{15}$$

Note that (15) can be solved only using an iterative method.

The fuzzy clustering algorithm that minimizes (7) is built by adopting an estimation strategy based on the Fu and Albus heuristic algorithm (Fu & Albus, 1977; Krishnapuram et al., 1999, 2001). Indeed, the alternating optimization estimation procedure cannot be adopted because the necessary conditions cannot be derived by differentiating the objective function in (7) with respect to the medoids. The fuzzy clustering procedure is illustrated in Algorithm 1.

Algorithm 1 robust Entropy-based Fuzzy c-Medoids Clustering for Interval-Value Data (EFMd-ID) algorithm

Fix C, $max.iter$ and generate randomly the degree matrix $\mathbf{U}$;
Set $iter = 0$;
Compute β according to (6);
Pick initial medoids: $\{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_C\}$;
repeat
  Store the current medoids;
  Compute $\mathbf{u}_i$ $(i = 1, \ldots, I)$ by using (12);
  Compute w by using (15);
  Select the new medoids:
  for $c = 1$ to C **do**
    $q = \arg\min_{1 \le i' \le I} \sum_{i''=1}^{I} u_{i''c} \left\{1 - exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_{i'}, \mathbf{m}_{i''}) + w^2 d^2(\mathbf{r}_{i'}, \mathbf{r}_{i''})\right]\right\} + p \sum_{i''=1}^{I} \sum_{c=1}^{C} u_{i''c} log(u_{i''c})$
    return $\Rightarrow \tilde{\mathbf{x}}_c = \tilde{\mathbf{x}}_q$
  end for
  $iter \leftarrow iter_{OLD} + 1$;
until *current medoids=old medoids* or $iter = max.iter$

2.3.2 Robust entropy-based fuzzy c-means clustering (EFCM-ID) model

The Robust Entropy-based Fuzzy c-Means clustering (EFCM-ID) model is characterized as follows:

$$\begin{aligned} \min: & J_{EFCM-ID}(\mathbf{U}, \tilde{\mathbf{X}}, w) = \\ & \sum_{i=1}^{I} \sum_{c=1}^{k} u_{ic} \left\{1 - exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \mathbf{m}_c) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_c)\right]\right\} \\ & + p \sum_{i=1}^{I} \sum_{c=1}^{C} u_{ic} log(u_{ic}) \\ & \sum_{c=1}^{C} u_{ic} = 1, u_{ic} \ge 0 \\ & 0 \le w \le 0.5 \end{aligned} \tag{16}$$

where $\mathbf{m}_c$ and $\mathbf{r}_c$ are the centroids of the midpoints and radii in the c-th cluster.

The optimal solutions for u_{ic} and w are obtained as in the EFMd-ID model.

The centroids for the midpoints and radii are obtained by minimizing the objective function with respect to $\mathbf{m}_c$ and $\mathbf{r}_c$ component-wise, respectively:

$$\mathbf{m}_c = \frac{\sum_{i=1}^{I} u_{ic} exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \mathbf{m}_c) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_c)]\right] \mathbf{m}_i}{\sum_{i=1}^{I} u_{ic} exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \mathbf{m}_c) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_c)]\right]} \tag{17}$$

$$\mathbf{r}_c = \frac{\sum_{i=1}^{I} u_{ic} exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \mathbf{m}_c) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_c)]\right] \mathbf{r}_i}{\sum_{i=1}^{I} u_{ic} exp\left[-\beta[(1-w)^2 d^2(\mathbf{m}_i, \mathbf{m}_c) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_c)]\right]} \tag{18}$$

Note that Eqs. (17) and (18) can be solved only using an iterative method.
The fuzzy clustering procedure is illustrated in Algorithm 2.

Algorithm 2 robust Entropy-based Fuzzy c-Means Clustering for Interval-Value Data (EFM-ID) algorithm

1: Fix C, $max.iter$ and generate randomly the degree matrix $\mathbf{U}$;
2: Set $iter = 0$;
3: Compute β according to (6);
4: Set initial centroids: $\{\mathbf{x}_1, \ldots, \mathbf{x}_C\}$;
5: **repeat**
6: Store the current centroids;
7: Compute $\mathbf{u}_i$ $(i = 1, \ldots, I)$ by using (12);
8: Compute w by using (15);
9: Compute the new centroids according to Eqs. (17) and (18);
10: $iter \leftarrow iter_{OLD} + 1$;
11: **until** *current centroids=old centrids* or $iter = max.iter$

2.3.3 Other models

As variants of the proposed fuzzy clustering models (7) and (14) other related models can be suggested, either fuzzy entropy-based not robust or fuzzy not entropy-based.

In particular:

- Entropy-based Fuzzy c-Medoids clustering model for interval-valued data with (not robust) weighted dissimilarity measure (not robust version of EFCMd-ID).
- Entropy-based Fuzzy c-Means clustering model for interval-valued data with (not robust) weighted dissimilarity measure (not robust version of EFCM-ID).
- Robust Fuzzy c-Medoids clustering model for interval-valued data (FCMd-ID with exponential weighted dissimilarity measure 5) (D'Urso et al., 2016): fixing p=0 (removing the entropy term) and considering the fuzziness exponent m for the membership degrees in (7).
- Robust Fuzzy c-Means clustering model for interval-valued data (FCM-ID with exponential weighted dissimilarity measure 5): fixing p=0 (removing the entropy term) and considering the fuzziness exponent m for the membership degrees in (16).

The models are summarized in Table 1.

Table 1 Variants of the proposed fuzzy clustering models (7) and (16)

	Fuzzy entropy-based	Fuzzy
Not robust dissimilarity	Not robust EFCMd-ID, EFCM-ID	
Robust dissimilarity	Robust EFCMd-ID, EFCM-ID	Robust FCMd-ID, FCM-ID

3 Simulation study

The performances of the proposed Robust Entropy-based Fuzzy c-Medoids clustering model for interval-valued data with weighted dissimilarity measure, i.e. the EFCMd-ID model, have been evaluated by carrying out a simulation study. The proposed model has been compared with the Robust Entropy-based Fuzzy c-Means clustering model for interval-valued data with weighted dissimilarity measure i.e. the EFCM-ID model, with the Robust Fuzzy c-Medoids clustering model for interval-valued data (FCMd-ID with exponential weighted dissimilarity measure) and with its EFCMd-ID not robust version.

Eighty objects ($I = 80$), two interval-valued variables ($J = 2$) and three percentages of noisy data in the dataset (0% to 15% step 5%) have been considered. Two clusters ($C = 2$) are generated in each simulation. Five values of the degree of fuzzy entropy p (0.05 to 0.30 step 0.05) for the entropy-based models and four values of the fuzziness parameter m ($m = 1.0, 1.3, 1.5, 2.0$) have been considered.

In the data generation scheme the midpoints 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 midpoints and the radii belonging to the second cluster ($I/2$ observations) from $U[1.5, 2.5]$.

To evaluate the robustness of the proposed model in presence of noisy data, $0.05 \cdot I$ to $0.15 \cdot I$ noisy objects have been added to the 80 objects. The midpoints and the radii of the noisy objects are generated from a Gaussian distribution $N(4.5, 2)$. Each data generation scheme has been replicated 100 times.

The data generation is summarized in Table 2.

The simulated scenario is presented in Fig. 1.

To assess the robustness with respect to misclassification in the presence of noisy data, an extension of the Adjusted Rand Index (ARI) for fuzzy partitions based on the Normalized Degree of Concordance (D'Ambrosio et al., 2021) has been used. The index allows the comparison of the hard partition in two clusters with the fuzzy partition obtained as a result of the robust model. The normalized degree of concordance varies between 0 and 1, and it always equals 1 when comparing a fuzzy partition with itself. The index has been then averaged over the 100 simulation runs.

The boxplots of the values of the extended ARI over 100 simulations are presented in Figs. 2, 3, 4 and 5, along with the boxplots of the values of the weight of the radii.

Some comments follow, with respect to the boxplots of the extended ARI.

The model FCMd-ID is less robust to the presence of noisy data than the other models. Considering the three robust models, EFCMd-ID presents better performances than EFCM-ID and FCMd-ID, in particular as the percentage of noisy data increases, especially for small values of the degree of fuzzy entropy. The weights of the radii are in the region of 0.5, always below, as expected.

Table 2 Data and noisy data generation scheme

Data generation scheme		Midpoints	Radii
Midpoints and radii	Cluster 1 ($i = 1, \ldots, I/2$)	$U[0, 1]$	$U[0, 1]$
	Cluster 2 ($i = I/2 + 1, \ldots, I$)	$U[1.5, 2.5]$	$U[1.5, 2.5]$
	Noisy data	$N(4.5, 2)$	$N(4.5, 2)$

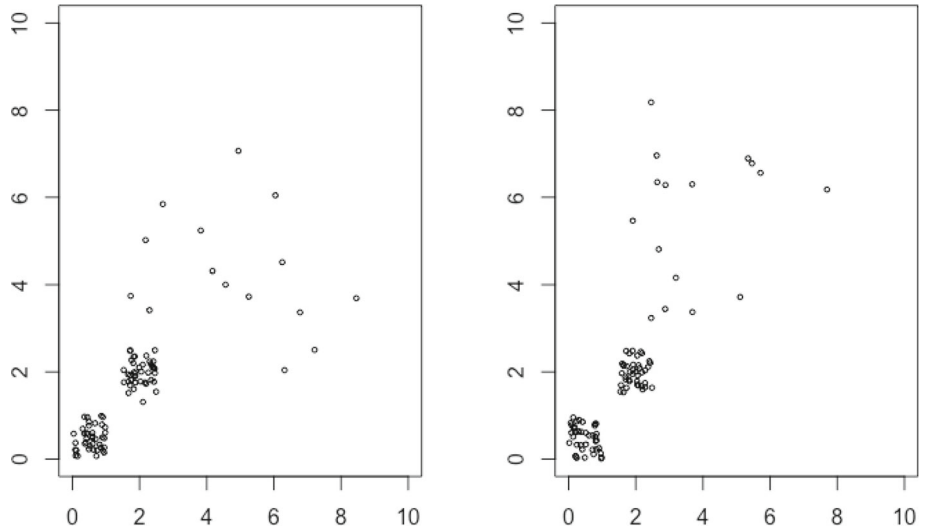

Fig. 1 Simulated midpoint-radius scenario. The midpoints are presented in the left figure, the radii in the right figure

4 Application: robust clustering of scientific journals

In this Section, an application of the proposed EFCMd-ID model to the clustering of scientific journals in the field of research evaluation is presented.

Institutional bodies in many countries evaluate the quality of the outcomes of the research of the universities and research institutes providing an up-to-date assessment of the state of research in the various scientific fields, in order to promote the improvement of research quality in the assessed institutions and to allocate the Ordinary Financing Fund for the University system on a performance basis.

To define the quality profiles of the research outputs, the peer review method is adopted. When considered appropriate to the characteristics of the field, peer review can be informed by the use of international citation indicators.

The Journal Citation Report^{TM} (JCR) from Clarivate provides transparent, publisher-neutral data and statistics needed to make confident decisions in the evolving scholarly publishing landscape. Publishers and editors can make confident business decisions - understand how journals are performing and benchmark them against others. Librarians can make confident collection management decisions - understand which journals are the most important to the institution's and researchers' success. Researchers can make confident decisions about where to submit manuscripts - using Journal Citation Reports as a definitive list and guide to discover and select the most appropriate journals to read and publish research findings.

Among the indicators proposed by JCR, the 5-Year Journal Impact Factor (5-Year JIF) and the $Immediacy$ Index^{TM} have been considered in the application.

The Journal Impact Factor^{TM} is the average number of times articles from the journal published in the past two years have been cited in the JCR year. The Impact Factor is calculated by dividing the number of citations in the JCR year by the total number of

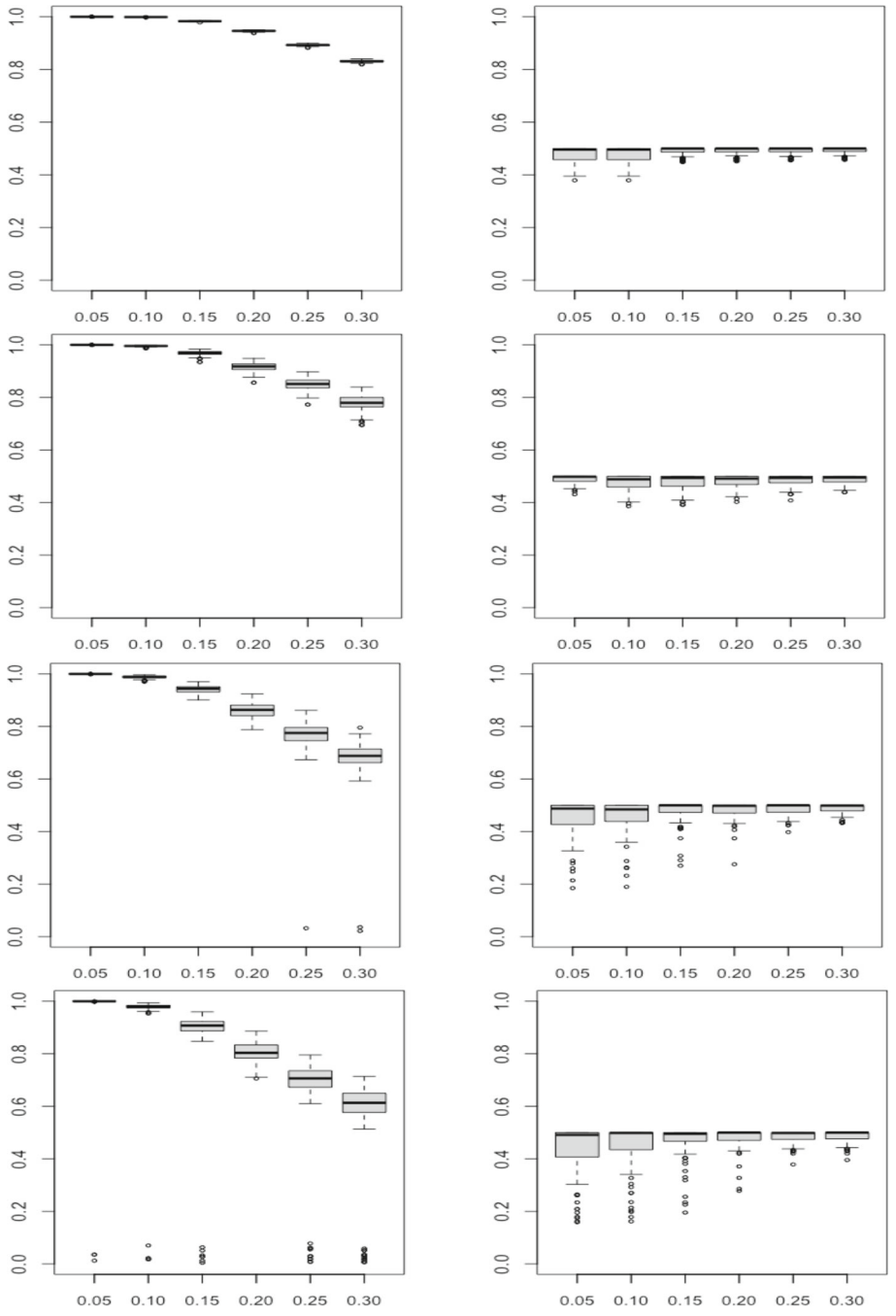

Fig. 2 Robust EFCMd-ID. The extended ARI is shown in the left panel, while the weight of the radii is on the right panel. From top to bottom, there are scenarios with 0%, 5%, 10% and 15% of noisy data, respectively. Five values of the degree of fuzzy entropy p (0.05 to 0.30 step 0.05) are considered

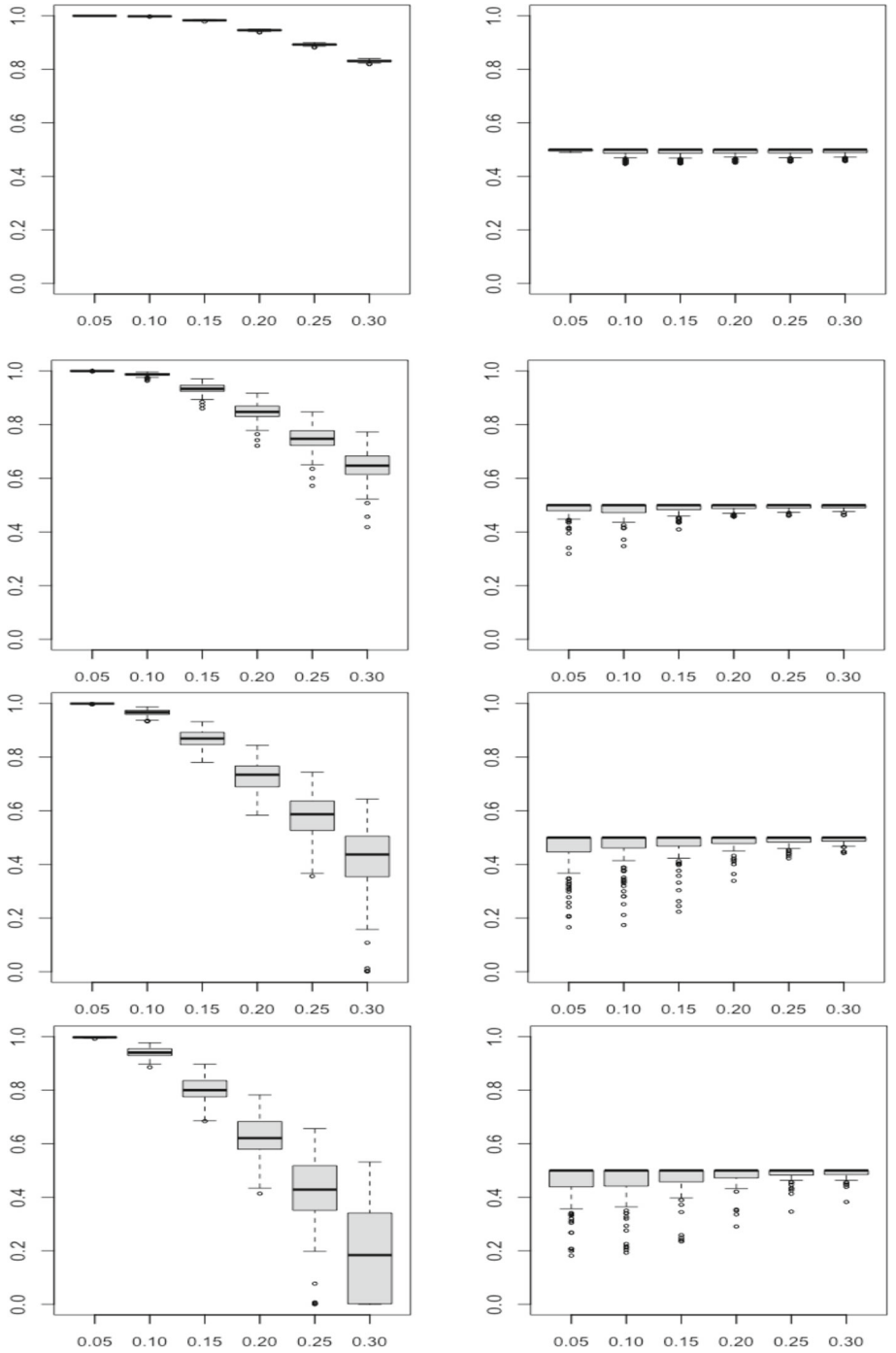

Fig. 3 Robust EFCM-ID. The extended ARI is shown in the left panel, while the weight of the radii is on the right panel. From top to bottom, there are scenarios with 0%, 5%, 10% and 15% of noisy data, respectively. Five values of the degree of fuzzy entropy p (0.05 to 0.30 step 0.05) are considered

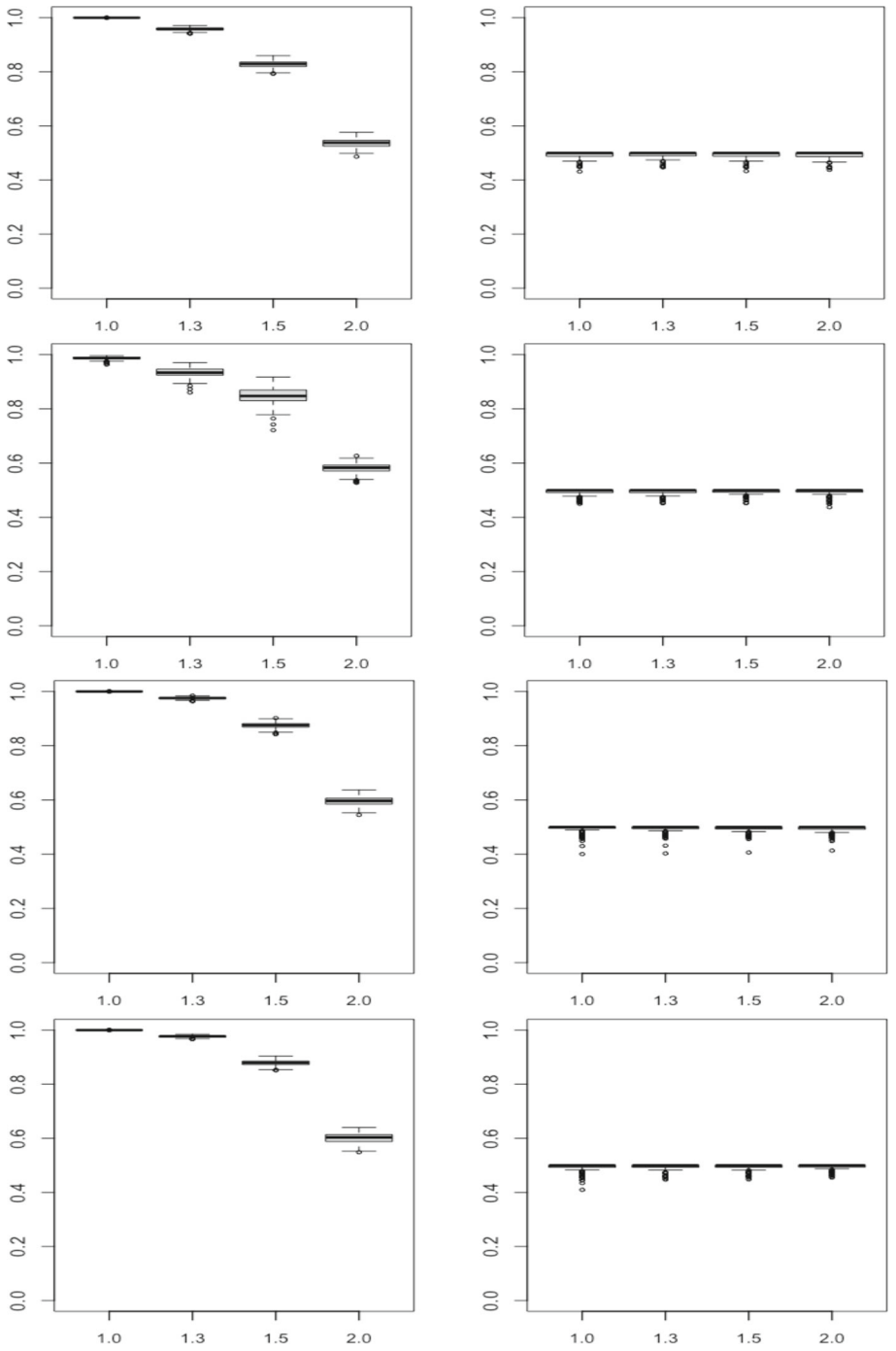

Fig. 4 FCMd-ID with exponential weighted dissimilarity measure. The extended ARI is shown in the left panel, while the weight of the radii is on the right panel. From top to bottom, there are scenarios with 0%, 5%, 10% and 15% of noisy data, respectively. Four values of the fuzziness parameter m (1.0, 1.3, 1.5, 2.0) are considered

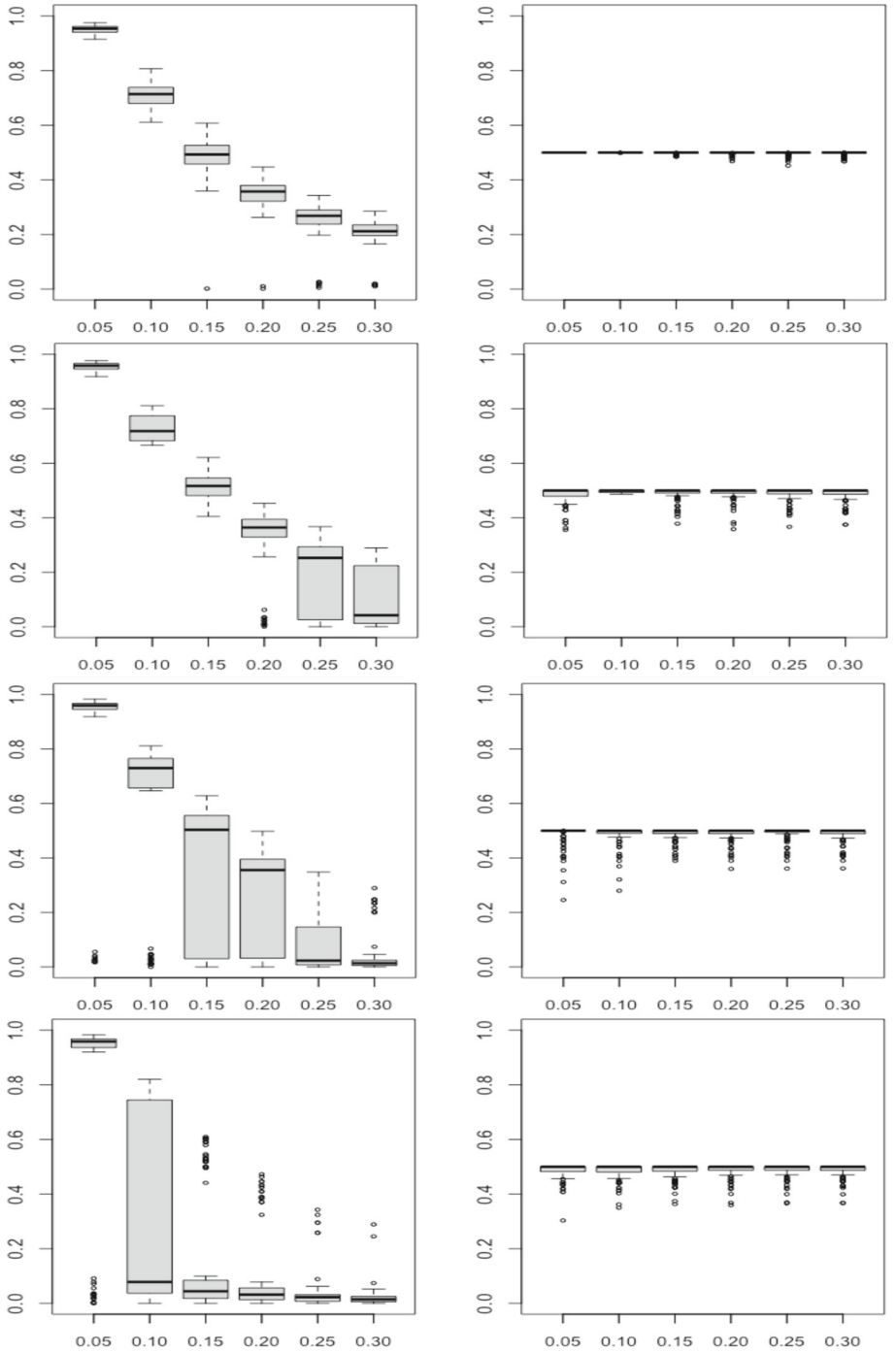

Fig. 5 Not robust EFCMd-ID. The extended ARI is shown in the left panel, while the weight of the radii is on the right panel. From top to bottom, there are scenarios with 0%, 5%, 10% and 15% of noisy data, respectively. Five values of the degree of fuzzy entropy p (0.05 to 0.30 step 0.05) are considered

articles published in the two previous years. Citing articles may be from the same journal; most citing articles are from different journals.

The 5-year Journal Impact Factor is the average number of times articles from the journal published in the past five years have been cited in the JCR year. It is calculated by dividing the number of citations in the JCR year by the total number of articles published in the five previous years. The 5-Year Impact Factor is available only in JCR 2007 and subsequent years.

The Immediacy Index is the average number of times an article is cited in the year it is published. The journal Immediacy Index indicates how quickly articles in a journal are cited. The Immediacy Index is calculated by dividing the number of citations to articles published in a given year by the number of articles published in that year. Because it is a per-article average, the Immediacy Index tends to discount the advantage of large journals over small ones. However, frequently issued journals may have an advantage because an article published early in the year has a better chance of being cited than one published later in the year. Many publications that publish infrequently or late in the year have low Immediacy Indexes. For comparing journals specializing in cutting-edge research, the Immediacy Index can provide a useful perspective.

Journals are organized into categories and groups. Groups are used to organize the 254 categories of JCR into broad discipline areas. Groups in JCR have no associated metrics and aren't used for rankings. Categories may be in more than one group.

The category "Health Care Sciences & Services" in the group "Clinical Medicine" has been considered. Health Care Sciences & Services covers resources on health services, hospital administration, health care management, health care financing, health policy and planning, health economics, health education, history of medicine, and palliative care.

The units (objects) are 74 journals, the variables 5-Year JIF and $Immediacy$ index ($J = 2$). The variables have been collected in the period 2017–2021 and the minimum and maximum value in the period has been computed for each variable. The reformulation with midpoint and radius has been used. The data are presented in Fig. 6 and Table 4. We observe that the two indexes give different information as expected. The EFCMd-ID model has been run over five values of the degree of fuzzy entropy p ($p = 0.05 - 0.30$ step 0.05) and $C = 2$, 3, 4, 5, 6 clusters.

Remark 1 Because of its particularly satisfactory results in recognizing the true number of clusters [for a reference, see the extensive simulations carried out in Arbelaitz et al. (2013)], we select the optimal C according to the Fuzzy Silhouette criterion (Campello & Hruschka, 2006), that is a fuzzy version of the Average Silhouette Width (ASW) criterion (Kaufman & Rousseeuw, 1990). The Fuzzy Silhouette index (FS) measures cohesion and separation of a partition. This index represents the weighted average of individual silhouettes width, λ_i, with weights derived from the fuzzy membership matrix $\mathbf{U} = \{u_{ic} : i = 1, \ldots, I;\ c = 1, \ldots, C\}$:

$$\mathrm{FS} = \frac{\sum_{i=1}^{I}(u_{ip} - u_{iq})^{\alpha} \cdot \lambda_i}{\sum_{i=1}^{I}(u_{ip} - u_{iq})^{\alpha}}, \qquad \lambda_i = \frac{(b_i - a_i)}{\max\{b_i, a_i\}} \tag{19}$$

where a_i is the average distance between the i-th unit and the units belonging to the cluster p ($p = 1,...,C$) with which i is associated with the highest membership degree; b_i is the minimum (over clusters) average distance of the i-th unit to all units belonging to the cluster q with $q \neq p$; $(u_{ip} - u_{iq})^{\alpha}$ is the weight of each λ_i calculated upon $\mathbf{U}$, where p and q are, respectively, the first and second best clusters (according to the membership degree) to

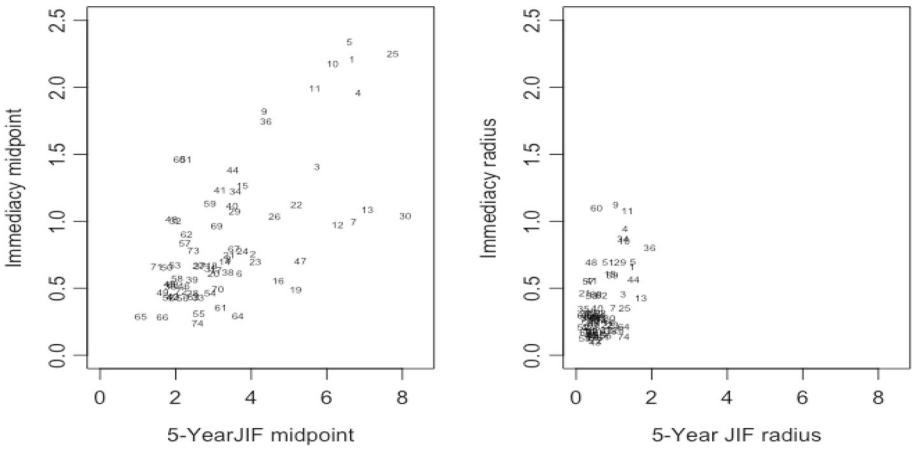

Fig. 6 Midpoints and radii of the variables 5-year JIF and $Immediacy$ index. The midpoints are presented in the left figure, the radii in the right figure

which the i-th unit is associated; $\alpha \geq 0$ is an optional user defined weighting coefficient. The traditional Silhouette coefficients is obtained by setting $\alpha = 0$.

The higher the value Fuzzy Silhouette index, the better the assignment of the units to the clusters simultaneously obtaining the minimization of the intra-cluster distance and the maximization of the inter-cluster distance.

Remark 2 An empirical rule for selecting a suitable cut-off point of the highest membership values has been suggested by Dembélé and Kastner (2003) and also used by Belacel et al. (2004). Dembélé and Kastner (2003) and Belacel et al. (2004) studied the cut-off point of the highest membership value with the fuzziness parameter in a fuzzy clustering framework. In particular, Dembélé and Kastner (2003) proposed a new method which enabled the computation of the upper bound value for m and showed that Fuzzy c-Means clustering of microarray data, combined with threshold-based gene selection, offers a convenient way of defining subsets of gene which are more tightly associated with a given cluster. In our paper, the aim is not to investigate the relationship between m and the cut-off for the membership degrees. Hence, the chosen cut-off point of 0.7 for a partition in two clusters for the membership degrees is compatible with the indications suggested in literature; i.e., for the simulation studies, see D'Urso and Maharaj (2009) and Maharaj et al. (2010), and for the applications see Dembélé and Kastner (2003) and D'Urso and Giordani (2006b).

The results are presented in Table 3 - Fuzzy Silhouette.

The optimal number of clusters is $C = 2$, the degree of fuzzy entropy $p = 0.20$. The cluster numerosity is 30, 22 and 22 journals have a fuzzy membership. The medoids are journal 39: "Journal of interprofessional care" and journal 8: "Supportive care in cancer" (highlighted in bold in Table 4).

Considering the midpoints (Fig. 6, left and Fig. 7), two clusters represent, respectively, journals with small values of the midpoint of the 5-Year JIF and medium-high values of the midpoint of the $Immediacy$ index (medoid: "Journal of interprofessional care"); and journals with high values of the midpoint of the 5-Year JIF and medium-high values of the midpoint of the $Immediacy$ index (medoid: "Supportive care in cancer"), for some journals smaller than in the other cluster.

Table 3 Fuzzy Silhouette for different values of the number of clusters C and of the degree of fuzzy entropy p

C	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.20$	$p = 0.25$	$p = 0.30$
2	0.278	0.279	0.284	**0.355**	0.325	0.326
3	0.253	0.242	0.128	0.252	0.272	0.286
4	0.279	0.263	0.189	0.124	0.198	0.133
5	0.234	0.200	0.032	0.066	0.098	0.061
6	0.207	0.200	0.138	0.055	0.108	0.106

Bold indicates the highest value and related pair (C, p)

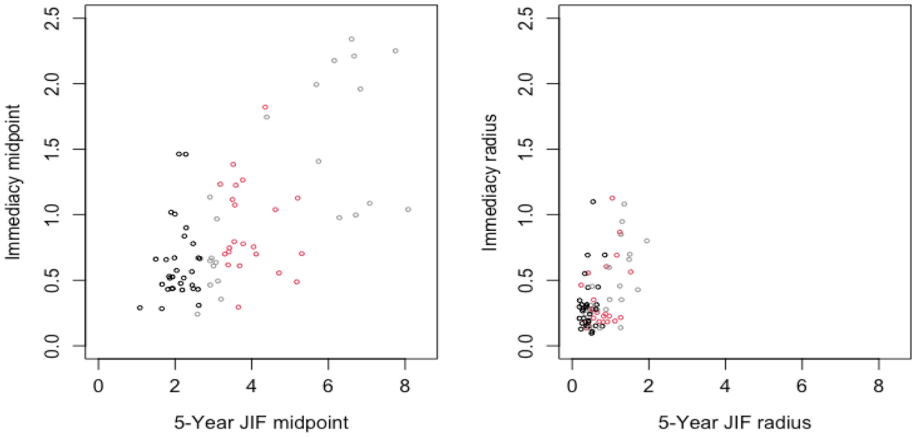

Fig. 7 Midpoints and radii of the variables 5-year JIF and $Immediacy$ index. The midpoints are presented in the left figure, the radii in the right figure. The journals in the two clusters are coloured red and black, respectively, the fuzzy journals grey

The 22 fuzzy journals (in italic in Table 4) show either the values of the midpoints of the two variables greater than the cluster with medoid "Supportive care in cancer", in particular journals 1, 4, 5, 10, 11, 25, 36; or the values of the midpoint of the 5-Year JIF greater than the cluster with medoid "Supportive care in cancer", in particular journals 3, 7, 12, 13, 30; or the values of the midpoints of the two variables in the middle with respect to the medoids of the two clusters. The memberships demonstrate the ability of the model to smooth the presence of noisy journals, without altering the medoids.

Considering the radii (Fig. 6, right and Fig. 7, greater dispersion is observed with respect to the *Immediacy* index. Noisy also with respect to the radii are journals 4, 11, 5, 36 (high 5-Year JIF and $Immediacy$ radius), journals 3, 7, 13, 25 (high 5-Year JIF radius). Journal 74 is a singleton as it shows a high radius of 5-Year JIF and a small radius of the $Immediacy$ index.

The value of the weight of the radius component is always greater than 0.5, demonstrating the smaller variability of the radii, resulting in a weight equal to 0.5.

Table 4 Fuzzy memberships

Journal		5-Year 2021 JIF		$Immediacy$ Index		$c = 1$	$c = 2$
		Midpoint	Radius	Midpoint	Radius		
1	*Journal of clinical epidemiology*	6.670	1.485	2.212	0.660	0.495	0.505
2	Medical care	4.046	0.374	0.756	0.135	0.082	0.918
3	*Journal of general internal medicine*xx	5.741	1.244	1.407	0.457	0.414	0.587
4	*Health affairs*	6.834	1.302	1.960	0.947	0.496	0.504
5	*Academic medicine*	6.604	1.504	2.341	0.699	0.495	0.505
6	Quality of life research	3.684	0.705	0.612	0.185	0.105	0.895
7	*Journal of medical internet research*	6.713	0.978	0.997	0.353	0.486	0.514
8	**Supportive care in cancer**	3.402	0.556	0.718	0.212	**0.000**	**1.000**
9	Journal of pain and symptom management	4.350	1.047	1.821	1.126	0.266	0.734
10	*Medical education*	6.153	1.270	2.177	0.852	0.482	0.518
11	*Journal of the american medical informatics association*	5.684	1.357	1.994	1.081	0.456	0.544
12	*Value in health*	6.285	0.650	0.976	0.273	0.458	0.542
13	*Implementation science*	7.076	1.713	1.087	0.429	0.497	0.503
14	Health services research	3.300	0.351	0.701	0.180	0.229	0.771
15	Medical teacher	3.766	0.895	1.265	0.606	0.121	0.879
16	Journal of health economics	4.712	0.870	0.557	0.244	0.141	0.859
17	*Health policy*	3.065	0.484	0.638	0.265	0.382	0.619
18	*Health economics*	2.950	0.479	0.671	0.311	0.475	0.525
19	Health technology assessment	5.174	0.439	0.489	0.151	0.240	0.760
20	*Journal of palliative medicine*	3.002	0.469	0.611	0.260	0.439	0.562

Table 4 continued

Journal		5-Year 2021 *JIF*		*Immediacy* Index		$c = 1$	$c = 2$
		Midpoint	Radius	Midpoint	Radius		
21	Medical decision making	3.414	0.235	0.748	0.464	0.194	0.806
22	Palliative medicine	5.198	0.823	1.126	0.226	0.256	0.744
23	International journal of medical informatics	4.111	0.965	0.701	0.227	0.087	0.913
24	Health policy and planning	3.775	0.649	0.780	0.255	0.090	0.910
25	*Bmj quality & safety*	7.748	1.285	2.252	0.353	0.500	0.500
26	Pharmacoeconomics	4.618	1.117	1.039	0.190	0.145	0.855
27	International journal for quality in health care	2.606	0.250	0.672	0.314	0.749	0.251
28	American journal of managed care	2.447	0.636	0.463	0.314	0.822	0.178
29	Journal of medical systems	3.561	1.163	1.074	0.693	0.164	0.836
30	*Milbank quarterly*	8.079	0.885	1.040	0.278	0.500	0.500
31	*Statistical methods in medical research*	2.915	0.351	0.650	0.242	0.517	0.483
32	Journal of evaluation in clinical practice	1.997	0.378	1.004	0.315	0.890	0.110
33	Journal of school health	2.596	0.424	0.432	0.182	0.766	0.234
34	Telemedicine and e-health	3.582	1.243	1.225	0.869	0.193	0.807
35	Journal of manipulative and physiological therapeutics	1.864	0.196	0.517	0.347	0.919	0.082
36	*Journal of telemedicine and telecare*	4.391	1.948	1.745	0.802	0.338	0.663
37	European journal of cancer care	2.647	0.284	0.667	0.281	0.724	0.276
38	Health expectations	3.383	0.918	0.618	0.183	0.183	0.817
39	**Journal of interprofessional care**	2.438	0.461	0.566	0.242	**1.000**	**0.000**
40	Advances in health sciences education	3.498	0.560	1.116	0.352	0.140	0.861

Table 4 continued

Journal		5-Year 2021 *JIF*		*Immediacy* Index		$c = 1$	$c = 2$
		Midpoint	Radius	Midpoint	Radius		
41	Medical care research and review	3.178	0.415	1.233	0.557	0.282	0.718
42	International journal of technology assessment in health care	1.918	0.298	0.438	0.211	0.920	0.080
43	American journal of hospice & palliative medicine	1.939	0.508	0.438	0.096	0.916	0.084
44	Journal of rural health	3.515	1.528	1.384	0.565	0.222	0.778
45	Methods of information in medicine	1.844	0.530	0.530	0.280	0.921	0.080
46	Expert review of pharmacoeconomics & outcomes research	2.226	0.370	0.519	0.154	0.888	0.112
47	Jmir mhealth and uhealth	5.304	0.807	0.705	0.181	0.272	0.728
48	Gaceta sanitaria	1.896	0.403	1.018	0.693	0.881	0.119
49	Australian health review	1.661	0.410	0.470	0.149	0.920	0.080
50	International journal of health services	1.768	0.413	0.659	0.446	0.918	0.082
51	Hastings center report	2.278	0.852	1.461	0.694	0.727	0.273
52	American journal of medical quality	1.813	0.188	0.432	0.210	0.920	0.081
53	Journal of continuing education in the health professions	1.983	0.364	0.672	0.304	0.913	0.087
54	*Simulation in healthcare-journal of the society for simulation in healthcare*	2.916	0.600	0.465	0.296	0.521	0.479
55	Scandinavian journal of primary health care	2.616	0.611	0.309	0.153	0.749	0.251
56	Teaching and learning in medicine	2.186	0.786	0.427	0.150	0.879	0.121
57	Evaluation & the health professions	2.243	0.331	0.838	0.552	0.862	0.138
58	Journal of behavioral health services & research	2.041	0.230	0.576	0.128	0.908	0.092
59	*Disability and health journal*	2.911	0.961	1.134	0.599	0.431	0.569
60	Journal of health politics policy and law	2.101	0.544	1.463	1.099	0.748	0.252

Table 4 continued

Journal		5-Year 2021 JIF		$Immediacy$ Index		$c = 1$	$c = 2$
		Midpoint	Radius	Midpoint	Radius		
61	*Bmc palliative care*	3.198	0.720	0.357	0.136	0.306	0.694
62	Journal of public health policy	2.288	0.674	0.902	0.450	0.837	0.163
63	Current opinion in supportive and palliative care	2.473	0.188	0.438	0.298	0.826	0.174
64	International journal of integrated care	3.651	1.264	0.295	0.216	0.165	0.835
65	Technology and health care	1.082	0.259	0.291	0.173	0.855	0.145
66	Journal of palliative care	1.653	0.425	0.284	0.193	0.918	0.082
67	Patient-patient centered outcomes research	3.545	0.524	0.796	0.276	0.128	0.873
68	Families systems & health	1.930	0.364	0.527	0.292	0.919	0.081
69	*Journal of patient safety*	3.090	0.526	0.968	0.455	0.333	0.667
70	*Journal of managed care & specialty pharmacy*	3.116	0.452	0.495	0.137	0.361	0.639
71	Australian journal of primary health	1.497	0.268	0.662	0.267	0.910	0.090
72	population health management	2.147	0.522	0.477	0.109	0.897	0.103
73	Health informatics journal	2.470	0.622	0.781	0.281	0.790	0.210
74	*Journal of healthcare engineering*	2.583	1.262	0.242	0.139	0.690	0.310

Medoids in bold; fuzzy journals in italic

5 Final remarks

In this paper, a robust entropy-based fuzzy c-Medoids clustering model for interval-valued data is suggested. In particular, by considering a suitable weighted measure, we propose a robust fuzzy clustering model with an entropy regularization and Partition Around Medoid approach, the EFCMd-ID model. An important advantage of the use of the entropy regularization approach in a fuzzy clustering framework is the maximum entropy principle that provides the fuzzy clusterization of the observations while ensuring the maximum compactness of the obtained clusters (Coppi & D'Urso, 2006; Gao et al., 2019; Kahali et al., 2019). Robustness to noisy observations is obtained by the use of the exponential trasformation. The simulations have shown the ability of the model to tune properly the weight of the center and radius components of the interval-valued data and the degree of fuzzy entropy, besides robustness to noisy data, in a comparative assessment. An application to the clustering of scientific journals in the field of research evaluation is provided, useful for Institutional bodies to evaluate the quality of the outcomes of the research of the universities and research institutes, in order to promote the improvement of research quality in the assessed institutions and to allocate the Ordinary Financing Fund for the University system on a performance basis.

Funding Open access funding provided by Luiss University within the CRUI-CARE Agreement.

Declarations

Conflicts of interest The authors declare the absence of any conflict of interest.

References

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.

Ashtari, P., Haredasht, F. N., & Beigy, H. (2020). Supervised fuzzy partitioning. *Pattern Recognition, 97*, 107013.

Belacel, N., Cuperlovic-Culf, M., Laflamme, M., & Ouellette, R. J. (2004). Fuzzy j-means and VNS methods for clustering genes from microarray data. *Bioinformatics, 20*(11), 1690–701.

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., & Schektman, 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. (2006). Fuzzy unsupervised classification of multivariate time trajectories with the Shannon entropy regularization. *Computational Statistics & Data Analysis, 50*(6), 1452–1477.

Coppi, R., Giordani, P., & D'Urso, P. (2006). Component models for fuzzy data. *Psychometrika, 71*(4), 733.

D'Ambrosio, A., Amodio, S., Iorio, C., Pandolfo, G., & Siciliano, R. (2021). Adjusted concordance index: An extension of the adjusted rand index to fuzzy partitions. *Journal of Classification, 38*, 112–128.

De Carvalho, F., 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.

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

De Carvalho, F. D. A., & 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.

Dembélé, D., & Kastner, P. (2003). Fuzzy c-means method for clustering microarray data. *Bioinformatics, 19*(8), 973–80.

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

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. (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.

D'Urso, P., De Giovanni, L., & Massari, R. (2016). Garch-based robust clustering of time series. *Fuzzy Sets and Systems, 305*, 1–28.

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. (2005). A Possibilistic approach to latent component analysis for symmetric fuzzy data. *Fuzzy Sets and Systems, 150*(2), 285–305.

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

D'Urso, P., & Giordani, P. (2006b). A weighted fuzzy c-means clustering model for fuzzy data. *Computational Statistics & Data Analysis, 50*(6), 1496–1523.

D'Urso, P., & Leski, J. (2016). Fuzzy c-ordered medoids clustering for interval-valued data. *Pattern Recognition, 58*, 49–67.

D'Urso, P., & Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems, 160*(24), 3565–3589.

D'Urso, P., Massari, R., De Giovanni, L., & Cappelli, C. (2017). Exponential distance-based fuzzy clustering for interval-valued data. *Fuzzy Optimization and Decision Making, 16*(1), 51–70.

Frieden, B. R., & Binder, P. M. (2000). Physics from fisher information: A unification. *American Journal of Physics, 68*(11), 1064–1065.

Fu, K., & Albus, J. (1977). *Syntactic pattern recognition*. Berlin: Springer.

Gao, Y., Wang, D., Pan, J., Wang, Z., & Chen, B. (2019). A novel fuzzy c-means clustering algorithm using adaptive norm. *International Journal of Fuzzy Systems, 21*(8), 2632–2649.

Giordani, P., & Kiers, H. A. (2004). Principal component analysis of symmetric fuzzy data. *Computational Statistics & Data Analysis, 45*(3), 519–548.

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

Guru, D., 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.

Ichihashi, H. (2000). Gaussian mixture pdf approximation and fuzzy c-means clustering with entropy regularization. In *Proceedings of 4th Asian fuzzy systems symposium* (pp. 217–221).

Kahali, S., Sing, J. K., & Saha, P. K. (2019). A new entropy-based approach for fuzzy c-means clustering and its application to brain MR image segmentation. *Soft Computing, 23*(20), 10407–10414.

Kaufmann, L. & Rousseeuw, P. (1987). Clustering by means of medoids. In *Data analysis based on the L1-norm and related methods* (pp. 405–416).

Kaufman, L. & Rousseeuw, P. J. (1990). Finding groups in data. In *An introduction to cluster analysis. Wiley series in probability and mathematical statistics. Applied probability and statistics*.

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.

Krishnapuram, R., Joshi, A., & Yi, L. (1999). A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In *1999 IEEE international fuzzy systems conference proceedings, FUZZ-IEEE'99* (Volu. 3, pp. 1281–1286), IEEE.

Li, R.-P. & Mukaidono, M. (1995). A maximum-entropy approach to fuzzy clustering. In *Proceedings of 1995 IEEE international conference on fuzzy systems* (Vol. 4, pp. 2227–2232), IEEE.

Li, R.-P., & Mukaidono, M. (1999). Gaussian clustering method based on maximum-fuzzy-entropy interpretation. *Fuzzy Sets and Systems, 102*(2), 253–258.

Maharaj, E. A., D'Urso, P., & Galagedera, D. (2010). Wavelet-based fuzzy clustering of time series. *Journal of Classification, 27*(2), 231–275.

Ménard, M., & Eboueya, M. (2002). Extreme physical information and objective function in fuzzy clustering. *Fuzzy Sets and Systems, 128*(3), 285–303.
Miyagishi, K., Yasutomi, Y., Ichihashi, H., & Honda, K. (2000). Fuzzy clustering with regularization by KL information. In *16th Fuzzy System Symposium*, pages 549–550.
Miyamoto, S., & Mukaidono, M. (1997). Fuzzy c-means as a regularization and maximum entropy approach. In *Proceedings of IFSA* (pp. 1–7).
Wu, K.-L., & Yang, M.-S. (2002). Alternative c-means clustering algorithms. *Pattern Recognition, 35*(10), 2267–2278.
Yao, J., Dash, M., Tan, S., & Liu, H. (2000). Entropy-based fuzzy clustering and fuzzy modeling. *Fuzzy Sets and Systems, 113*(3), 381–388.
Zarinbal, M., Zarandi, M. F., & Turksen, I. (2014). Relative entropy fuzzy c-means clustering. *Information Sciences, 260*, 74–97.
Zhang, D.-Q., & Chen, S.-C. (2004). A comment on "Alternative c-means clustering algorithms". *Pattern Recognition, 37*(2), 173–174.