



# Entropy-based fuzzy clustering of interval-valued time series

Vincenzina Vitale<sup>1</sup> · Pierpaolo D'Urso<sup>1</sup> · Livia De Giovanni<sup>2</sup> · Raffaele Mattera<sup>1</sup>

Received: 1 February 2023 / Accepted: 11 February 2024  
© The Author(s) 2024

## Abstract

This paper proposes a fuzzy  $C$ -medoids-based clustering method with entropy regularization to solve the issue of grouping complex data as interval-valued time series. The dual nature of the data, that are both time-varying and interval-valued, needs to be considered and embedded into clustering techniques. In this work, a new dissimilarity measure, based on Dynamic Time Warping, is proposed. The performance of the new clustering procedure is evaluated through a simulation study and an application to financial time series.

**Keywords** Interval-valued time series · Fuzzy clustering · Dynamic time warping · FTSE-MIB index

**Mathematics Subject Classification** 62H30

## 1 Introduction

In this paper, we propose a fuzzy  $C$ -medoids-based clustering method with entropy regularization to group data objects characterized by two sources of complexity: they

---

✉ Vincenzina Vitale  
vincenzina.vitale@uniroma1.it  
Pierpaolo D'Urso  
pierpaolo.durso@uniroma1.it  
Livia De Giovanni  
ldegiovanni@luiss.it  
Raffaele Mattera  
raffaele.mattera@uniroma1.it

<sup>1</sup> Department of Social Sciences and Economics, Sapienza University of Rome, Piazzale Aldo Moro, 5, 00185 Rome, Italy

<sup>2</sup> Department of Political Sciences and Data Lab, Luiss University, Viale Romania, 32, 00197 Rome, Italy

are both time-varying and interval-valued data, i.e. interval-valued time series (ITS). Specifically, ITS are time series for which an interval covering the observed values rather than a single value is associated with each time point: they are thus characterized by lower and upper boundaries, as well as by centres and radii.

The proposed clustering procedure is based on a new dissimilarity measure suitable for ITS which builds on the Dynamic Time Warping (DTW) and accounts for within-series variability rather than intrinsic imprecision.

It is worth noting that there has been a growing interest in these complex data in recent years, as they have become much more widely available and are being used in a variety of contexts. However, as the next section will detail, research about ITS clustering is still relatively limited. The main reason for deeply investigating this area of research is motivated by the importance of clustering these complex data objects taking into account intra-series variability, without losing this source of information by considering a summary of it as the corresponding mean or median values. Developing effective methods for handling and clustering ITS is proved to be highly valuable, especially when studying the range of variation of a particular variable over time. For instance, ITS can be used for analyzing wind speed levels, daily asset prices, electricity demand in regions, or a person's blood pressure fluctuations over time, among others. In general, there is a growing need for methodologies that deal with this type of symbolic data. According to the above considerations, we propose a new clustering technique for ITS based on a fuzzy approach, whose main advantage is that it produces a blurred partition of the objects to the clusters; in this way, there is no crisp assignment of units to clusters, allowing the identification of the second-best cluster too (Everitt and Leese 2001). Moreover, fuzzy clustering is distribution-free (Hwang et al. 2007). The fuzziness is introduced in the objective function through a regularization term based on the Shannon entropy (Li and Mukaidono 1995, 1999; Miyamoto and Mukaidono 1997). We consider a medoids-based clustering technique (Fcmd, Krishnapuram et al. 1999, 2001), given that the identification of the "virtual" prototype, i.e. the time series centroid, in this context may be meaningless, unlike the time series medoid, which is instead an observed time series.

The paper is structured as follows. Section 2 contains a review of the literature on time series clustering while Sect. 3 provides a formal definition of the ITS and then focuses on the new proposed dissimilarity measure. Section 4 introduces the fuzzy clustering method with entropy regularization based on the proposed dissimilarity while Sect. 5 contains the simulation plan. Section 6 shows the results of applying the clustering method to the components of the FTSE-MIB by considering the monthly minimum and maximum prices, ranging from October 2018 to October 2022 while Sect. 7 contains conclusions and some general remarks.

## 2 Literature review

Regarding the general approach to time series clustering, we argue that it has been widely used and applied in various fields including economics, finance, environmental and social sciences. As stated in Caiado et al. (2015), we can distinguish among the following three main approaches: *observation-based*, *features-based*, and *model-based*.

Clustering techniques belonging to the first group compute the dissimilarity based on the observed values of the time series (D'Urso 2005; D'Urso et al. 2018, 2021a) while those belonging to the second group are based on several features like quantile cross-spectral densities (López-Oriona et al. 2022a, b, c), quantile autocovariance (Vilar et al. 2018; Lafuente-Rego et al. 2020), autocorrelation function (Alonso and Maharaj 2006; D'Urso and Maharaj 2009) and generalized cross-correlation (Alonso et al. 2021), cepstral coefficients (Maharaj and D'Urso 2011), periodogram (Caiado et al. 2006, 2009), wavelets decomposition (Maharaj et al. 2010; D'Urso and Maharaj 2012; D'Urso et al. 2023). Finally, techniques belonging to the third group compute the dissimilarity among the parameter estimates arising from suitable fitted time series models like ARIMA models (Piccolo 1990; Xiong and Yeung 2004; D'Urso et al. 2015a), GARCH and INGARCH models (Otranto 2008, 2010; Caiado and Crato 2010; D'Urso et al. 2013, 2016; Otranto and Mucciardi 2019; Cerqueti et al. 2022), extreme value analysis (D'Urso et al. 2017a), splines coefficients (Garcia-Escudero and Gordaliza 2005; D'Urso et al. 2021b) and copulas (De Luca and Zuccolotto 2011; Durante et al. 2014, 2015; De Luca and Zuccolotto 2017; Disegna et al. 2017).

Regarding clustering techniques for interval-valued data, in the literature (see Noirhomme-Fraiture and Brito 2011 for an extensive review), many works have been devoted to this scope (De Carvalho et al. 2006a, b; De Carvalho and Tenório 2010; D'Urso et al. 2015b, 2017b; Montanari and Calò 2013; de Carvalho and Simões 2017; Kejžar et al. 2021; Umbleja et al. 2021), while clustering techniques for interval-valued time series are almost unexplored; indeed, few papers deal with imprecise time series such as Coppi and D'Urso (2002) and Coppi and D'Urso (2003) that introduce three types of dissimilarity: the instantaneous, the velocity and the simultaneous dissimilarity measures, respectively; Maharaj et al. (2019) and D'Urso et al. (2023) that follow a features-based approach.

In this work, we follow an observation-based approach within the fuzzy framework, exploiting all the advantages of dynamic time-warping, as discussed in the next section.

### 3 A dissimilarity measure for interval-valued time series

#### 3.1 Interval-valued time series

An interval time-series (ITS) can be seen as an interval variable observed over  $T$  times; formally, we can define a interval data time matrix as follows:

$$\mathbf{X} \equiv \{x_{it} = (l_{it}, h_{it}), i = 1, \dots, N, t = 1, \dots, T\}$$

where  $i$  and  $t$  denote the units and the times, respectively; then  $x_{it} = (l_{it}, h_{it})$  is the interval variable observed on the  $i$ -th unit at time  $t$ : in particular,  $l_{it}$  is the minimum value observed on the  $i$ -th unit, at time  $t$ , while  $h_{it}$  is the maximum value observed on the same  $i$ -th unit at the same time  $t$ .

We can also consider the corresponding following matrices:

$$\begin{aligned} \mathbf{L} &\equiv \{l_{it}; i = 1, \dots, N, t = 1, \dots, T\} \\ \mathbf{H} &\equiv \{h_{it}; i = 1, \dots, N, t = 1, \dots, T\} \end{aligned}$$

where  $\mathbf{L}$  is the matrix of time series of minima and  $\mathbf{H}$  that of the maxima.

In the following section, the Dynamic Time Warping is presented, being used, then, to define the new dissimilarity for ITS.

### 3.2 Dynamic time warping

Let  $\mathbf{x}_i$  and  $\mathbf{x}_j$  be two time series of length  $T$  and  $T'$  respectively, where we do not restrict  $T = T'$ . The Dynamic Time Warping (DTW, Velichko and Zagoruyko 1970; Berndt 1994) between these two time series works finding the optimal match between them under certain restrictions. Therefore, the sequences are warped nonlinearly to match each other.

The so-called “warping path” that “realigns” the time indices of the multivariate time series so that each data point in  $\mathbf{x}_i$  is compared to the “closest” data point in  $\mathbf{x}_j$  is defined as

$$\Phi_l = (\varphi_l, \psi_l), \quad l = 1, \dots, L$$

where  $\varphi_l$  and  $\psi_l$ , for  $l = 1, \dots, L$ , are the set of realigned indices of  $1, \dots, T$  and  $1, \dots, T'$  respectively, subject to the following constraints:

1. boundary condition:  $\Phi_1 = (1, 1)$ ,  $\Phi_L = (T, T')$ ;
2. monotonicity condition:  $\varphi_1 \leq \dots \leq \varphi_l \leq \dots \leq \varphi_L$  and  $\psi_1 \leq \dots \leq \psi_l \leq \dots \leq \psi_L$ .

Among the possible wrapping curves, the DTW is the one that minimizes the total dissimilarity between  $\mathbf{x}_i$  and  $\mathbf{x}_j$  computed as:

$$d_{DTW}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^L d(\mathbf{x}_{i, \varphi_l}, \mathbf{x}_{j, \psi_l}) m_{l, \Phi} \quad (1)$$

where  $m_{l, \Phi}$  is a local weighting coefficient and  $d(\cdot, \cdot)$  is, usually, the Euclidean distance for multivariate time series.

Even if the DTW algorithm could be problematic with long time series, when used in the Partitioning around Medoids (PAM) method, its computational burden is reduced since the distance matrix is computed only once in the iterative process.

Based on DTW, we define a new dissimilarity measure suitable when dealing with interval-valued time series, as follows.

### 3.3 A dissimilarity measure for ITS

Given two interval-valued time-series  $\mathbf{x}_i$  and  $\mathbf{x}_j$  of length  $T$  and  $T'$  (with  $T \geq T'$  or  $T \leq T'$ ), respectively, we define a new dissimilarity measure based on the Dynamic Time Warping. Therefore, we propose:

$$ITSd(\mathbf{x}_i, \mathbf{x}_j) = d_{DTW}^*(\mathbf{l}_i, \mathbf{l}_j) + d_{DTW}^*(\mathbf{h}_i, \mathbf{h}_j) + |d_{DTW}^*(\mathbf{l}_i, \mathbf{h}_i) - d_{DTW}^*(\mathbf{l}_j, \mathbf{h}_j)| \quad (2)$$

where  $\mathbf{l}_i = (l_{i1}, \dots, l_{it}, \dots, l_{iT})'$ ,  $\mathbf{l}_j = (l_{j1}, \dots, l_{jt}, \dots, l_{jT})'$ ,  $\mathbf{h}_i = (h_{i1}, \dots, h_{it}, \dots, h_{iT})'$  and  $\mathbf{h}_j = (h_{j1}, \dots, h_{jt}, \dots, h_{jT})'$  are the vectors of the minima and maxima pertaining to the time series of the  $i$ -th and  $j$ -th unit, respectively.  $d_{DTW}^*(\cdot; \cdot)$  denotes the normalized  $d_{DTW}(\cdot; \cdot)$ , i.e. the latter is divided by its maximum so that it varies in  $[0, 1]$ . Specifically:

- $d_{DTW}^*(\mathbf{l}_i, \mathbf{l}_j)$  is the normalized DTW among the time series of the minima of units  $i$  and  $j$ ;
- $d_{DTW}^*(\mathbf{h}_i, \mathbf{h}_j)$  is the normalized DTW among the time series of the maxima of units  $i$  and  $j$ ;
- $d_{DTW}^*(\mathbf{l}_i, \mathbf{h}_i)$  is the normalized DTW among the time series of the minima and the maxima of the unit  $i$ ;
- $d_{DTW}^*(\mathbf{l}_j, \mathbf{h}_j)$  is the normalized DTW among the time series of the minima and the maxima of the unit  $j$ ;

We finally remark that  $ITSd_{ij}$  thus can take values in the interval  $[0, 3]$ .

In the next Section, we introduce the Entropy-based fuzzy clustering method ( $Fcmd_{ITS}$ ) for interval-valued time series based on the above dissimilarity measure.

### 4 Entropy-based fuzzy clustering of interval-valued time series

A fuzzy  $C$ -medoids clustering method with entropy regularization is defined to group  $N$  interval-valued time series into  $C$  homogeneous clusters based on the new dissimilarity measure (2), henceforth named  $Fcmd_{ITS}$ . The aim is to identify the matrix of fuzzy membership degrees  $\mathbf{U}_{N \times C}$  whose generic element  $u_{ic}$  is the membership degree of the  $i$ -th time series to the  $c$ -th cluster, with  $i = 1, \dots, N$ ,  $c = 1, \dots, C$  and the corresponding  $C$  prototypes, i.e. the  $C$  time series medoids; both are the solutions of the following constraint optimization problem:

$$\left\{ \begin{aligned}
 \min : J_{p,C}(\mathbf{U}, \tilde{\mathbf{X}}) &= \sum_{i=1}^N \sum_{c=1}^C u_{ic} ITSd(\mathbf{x}_i, \tilde{\mathbf{x}}_c) + p \sum_{i=1}^N \sum_{c=1}^C u_{ic} \ln(u_{ic}) \\
 &= \sum_{i=1}^N \sum_{c=1}^C u_{ic} \left[ d_{DTW}^*(\mathbf{l}_i, \tilde{\mathbf{l}}_c) + d_{DTW}^*(\mathbf{h}_i, \tilde{\mathbf{h}}_c) \right. \\
 &\quad \left. + |d_{DTW}^*(\mathbf{l}_i, \mathbf{h}_i) - d_{DTW}^*(\tilde{\mathbf{l}}_c, \tilde{\mathbf{h}}_c)| \right] + p \sum_{i=1}^N \sum_{c=1}^C u_{ic} \ln(u_{ic}) \\
 s.t. \sum_{c=1}^C u_{ic} &= 1, u_{ic} \geq 0.
 \end{aligned} \right. \tag{3}$$

where  $ITSd(\mathbf{x}_i, \tilde{\mathbf{x}}_c)$  is the dissimilarity (2) between the  $i$ -th ITS and the ITS medoid of the  $c$ -th cluster.

In this objective function, the fuzziness is due to the entropy regularization term; its extent is controlled through the weight factor  $p > 0$ , which is called *degree of*

fuzzy entropy since the higher  $p$  the higher the degree of fuzziness. In particular, the second term is  $-p \times$  the Shannon entropy, so that the optimization problem aims at maximizing both the entropy measure and the internal cohesion.

The iterative solutions of (3) are:

$$u_{ic} = \frac{1}{\sum_{c'=1}^C \left[ \frac{\exp\left(\frac{1}{p} [d_{DTW}^*(l_i, \tilde{l}_c) + d_{DTW}^*(h_i, \tilde{h}_c) + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_c, \tilde{h}_c)|]\right)}{\exp\left(\frac{1}{p} [d_{DTW}^*(l_i, \tilde{l}_{c'}) + d_{DTW}^*(h_i, \tilde{h}_{c'}) + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_{c'}, \tilde{h}_{c'})|]\right)} \right]} \quad (4)$$

The proof of the iterative solutions (4) is given below, while the computational steps of the proposed clustering method are given in Algorithm 1.

Starting from the vector of the degrees of membership  $\mathbf{u}_i$  and using the Lagrangian multiplier, we get:

$$\begin{aligned} L_p(\mathbf{u}_i, \lambda) = & \sum_{i=1}^N \sum_{c=1}^C u_{ic} [d_{DTW}^*(l_i, \tilde{l}_c) + d_{DTW}^*(h_i, \tilde{h}_c) \\ & + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_c, \tilde{h}_c)|] \\ & + p \sum_{i=1}^n \sum_{c=1}^C u_{ic} \ln(u_{ic}) - \lambda \left( \sum_{c=1}^C u_{ic} - 1 \right), \end{aligned} \quad (5)$$

where  $\mathbf{u}_i = (u_{i1}, \dots, u_{ic}, \dots, u_{iC})'$  and  $\lambda$  is the Lagrange multiplier. Therefore, we set the first derivatives of (5) with respect to  $u_{ic}$  and  $\lambda$  equal to zero, yielding:

$$\begin{aligned} \frac{\partial L_p(\mathbf{u}_i, \lambda)}{\partial u_{ic}} = 0 \Leftrightarrow & [d_{DTW}^*(l_i, \tilde{l}_c) + d_{DTW}^*(h_i, \tilde{h}_c) \\ & + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_c, \tilde{h}_c)|] \\ & + p[\ln(u_{ic}) + 1] - \lambda = 0, \end{aligned} \quad (6)$$

$$\frac{\partial L_p(\mathbf{u}_i, \lambda)}{\partial \lambda} = 0 \Leftrightarrow \sum_{c=1}^C u_{ic} - 1 = 0. \quad (7)$$

From (6) we obtain:

$$\ln(u_{ic}) = \frac{1}{p} \left( \lambda - [d_{DTW}^*(l_i, \tilde{l}_c) + d_{DTW}^*(h_i, \tilde{h}_c) + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_c, \tilde{h}_c)|] \right) - 1 \quad (8)$$

and, then:

$$u_{ic} = \exp \left\{ \frac{\lambda}{p} - \frac{1}{p} \left( [d_{DTW}^*(l_i, \tilde{l}_c) + d_{DTW}^*(h_i, \tilde{h}_c) + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_c, \tilde{h}_c)|] \right) - 1 \right\} \quad (9)$$

By considering (7):

$$\exp\left(\frac{\lambda}{p} - 1\right) = \frac{1}{\sum_{c'=1}^C \frac{1}{\exp\left(\frac{1}{p}\left[d_{DTW}^*(l_i, \tilde{l}_{c'}) + d_{DTW}^*(h_i, \tilde{h}_{c'}) + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_{c'}, \tilde{h}_{c'})|\right]\right)}}. \quad (10)$$

and by replacing Eq. (10) in (9), we have the solution as in (4).

**Algorithm 1** *Fcmd<sub>ITS</sub>* algorithm

- 1: Fix the number of clusters  $C$ , the weight factor  $p$  and maximum number of iterations (`max.iter`);
- 2: Randomly initialize  $\mathbf{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_N\}$  with  $\mathbf{u}_i = (u_{i,1}, \dots, u_{i,C})$ ;
- 3: Set  $iter = 0$ ;
- 4: Store the  $C$  initial medoids:  $(\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_i, \dots, \tilde{\mathbf{x}}_C)$ ;
- 5: **repeat**
- 6:   Store the current medoids  $(\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_i, \dots, \tilde{\mathbf{x}}_C)_{OLD} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_i, \dots, \tilde{\mathbf{x}}_C)$ ;
- 7:   Compute  $\mathbf{u}_i$  ( $i = 1, \dots, N$ ) by using (4);
- 8:   **for**  $c = 1, \dots, C$ , determine the index  $q_c \in \{1, \dots, N\}$  satisfying:

$$q_c = \arg \min_{1 \leq q \leq N} \sum_{i=1}^N u_{ic} \left[ d_{DTW}^*(l_i, \tilde{l}_q) + d_{DTW}^*(h_i, \tilde{h}_q) + |d_{DTW}^*(l_i, h_i) - d_{DTW}^*(\tilde{l}_q, \tilde{h}_q)| \right]$$

- 9:   Return  $\tilde{\mathbf{x}}_c = \tilde{\mathbf{x}}_{q_c}$  for  $c = 1, \dots, C$
- 10:    $iter \leftarrow iter + 1$ ;
- 11: **until**  $(\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_i, \dots, \tilde{\mathbf{x}}_C)_{OLD} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_i, \dots, \tilde{\mathbf{x}}_C)$  or  $iter = \text{max.iter}$

## 5 A simulation study

A simulation plan has been considered to assess the proposed fuzzy clustering method’s performance. Three different scenarios have been simulated as described in the next section.

### 5.1 Simulated scenarios

*First scenario* The first scenario considered 45 simulated interval-valued time series clustered in 3 equally sized groups with length  $T = 100$ ; they are shown in Fig. 1a while in (b) a sample of three ITS, one per group, is represented.

In detail, the time series of minima belonging to the first group ( $C1$ ) have been simulated by a random walk with drift so that:

$$l_t = \delta_0 + l_{t-1} + w_t,$$

with  $\delta_0 = -1$  and  $w_t \sim N(0, 4)$ ; then, we shifted it by 50 so that  $l_t = l_t + 50$ .

The time series of minima belonging to the second group ( $C2$ ) have been simulated by a random walk with drift so that:

for  $t \in [1, T/2]$ ,

$$l_t = \delta_0 + l_{t-1} + w_t,$$

with  $\delta_0 = 1.5$  and  $w_t \sim N(0, 4)$ ;

for  $t \in [(T/2) + 1, T]$ ,

$$l_t = \delta_0 + l_{t-1} + w_t,$$

with  $\delta_0 = 2$  and  $w_t \sim N(0, 4)$ .

The time series of minima belonging to the third group (C3) have been simulated by a Moving Average (MA) model so that:

for  $t \in [1, (T/2) + 10]$ ,

$$l_t = \mu + w_t + \phi w_{t-1},$$

with  $\mu = 0$   $\phi = -0.5$  and  $w_t \sim N(0, 0.25)$ ;

for  $t \in [(T/2) + 11, T]$ ,

$$l_t = \mu + w_t + \phi w_{t-1},$$

with  $\mu = -5$   $\phi = -0.5$  and  $w_t \sim N(0, 0.25)$ .

The time series of maxima have been simulated as follows:

$$h_t = l_t + U(1, 10) + 10$$

*Second scenario* The second scenario, plotted in Fig. 2, considered 45 simulated interval-valued time series clustered in 3 equally sized groups with length  $T = 100$  too. The time series of minima have been simulated as in the first scenario while the time series of maxima as follows:

$$h_t = l_t + U(1, 10) + 70, \text{ for } h_t \in C1,$$

$$h_t = l_t + U(1, 10) + 20, \text{ for } h_t \in C2,$$

$$h_t = l_t + U(1, 10) + 80, \text{ for } h_t \in C3.$$

*Third scenario* In the third scenario, plotted in Fig. 3, we considered 30 time series clustered in 3 equally sized groups, with length  $T = 200$ . The time series of minima belonging to the first group (C1) have been simulated by an Autoregressive (AR) model so that:

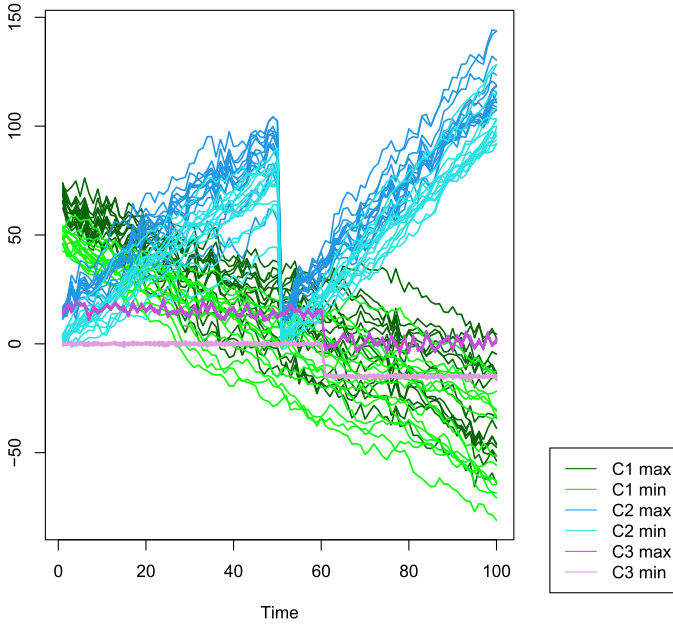
$$l_t = \alpha + \psi l_{t-1} + w_t,$$

with  $\psi = 0.7$ ,  $\alpha = 10(1 - \psi)$  and  $w_t \sim N(0, 1)$ .

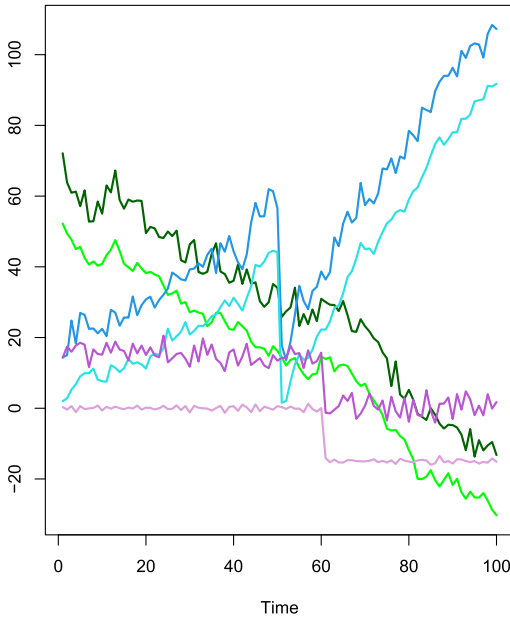
The time series of minima belonging to the second group (C2) have been simulated by an ARMA model so that:

$$l_t = \mu + \psi l_{t-1} + w_t + \phi w_{t-1},$$



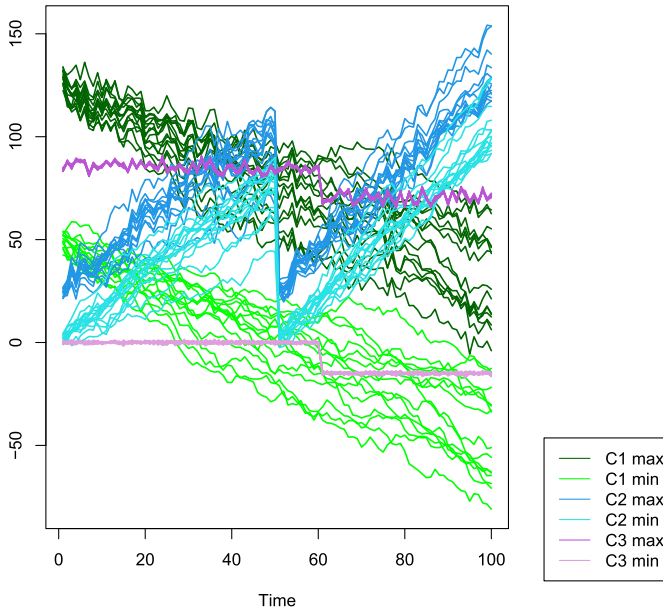


(a)

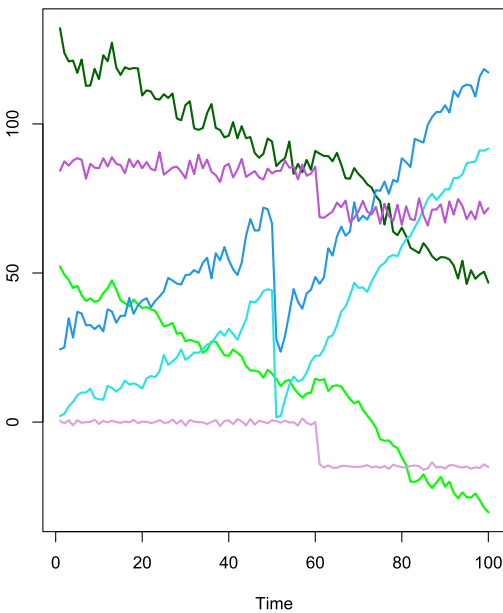


(b)

**Fig. 1** First scenario: **a** all simulated time series **b** a sample of three simulated ITS, one per group



(a)



(b)

**Fig. 2** Second scenario: **a** all simulated time series **b** a sample of three simulated ITS, one per group

with  $\mu = 20, \psi = -0.3, \phi = 0.5$  and  $w_t \sim N(0, 1)$ .

The time series of minima belonging to the third group (C3) have been simulated by an MA model so that:

for  $t \in [1, (T/2)]$ ,

$$l_t = \mu + w_t + \phi w_{t-1},$$

with  $\mu = 40, \phi = -0.5$  and  $w_t \sim N(0, 1)$ ;

for  $t \in [(T/2) + 1, T]$ ,

$$l_t = \mu + w_t + \phi w_{t-1},$$

with  $\mu = 50, \phi = -0.5$  and  $w_t \sim N(0, 1)$ .

The time series of maxima have been simulated as follows:

$$h_t = l_t + U(1, 6) + 10, \text{ for } h_t \in C1,$$

$$h_t = l_t + U(1, 6) + 20, \text{ for } h_t \in C2,$$

$$h_t = l_t + U(1, 6) + 30, \text{ for } h_t \in C3.$$

**Remark 1** A suitable pre-processing of the data may be required, such as normalization/standardization. Here, we propose the following normalization procedure:

$$l'_{it} = \frac{l_{it} - \min(l_{i1}, \dots, l_{iT})}{\max(h_{i1}, \dots, h_{iT}) - \min(l_{i1}, \dots, l_{iT})},$$

and

$$h'_{it} = \frac{h_{it} - \min(l_{i1}, \dots, l_{iT})}{\max(h_{i1}, \dots, h_{iT}) - \min(l_{i1}, \dots, l_{iT})},$$

$\forall t = 1, \dots, T$  and  $i = 1, \dots, N$ .

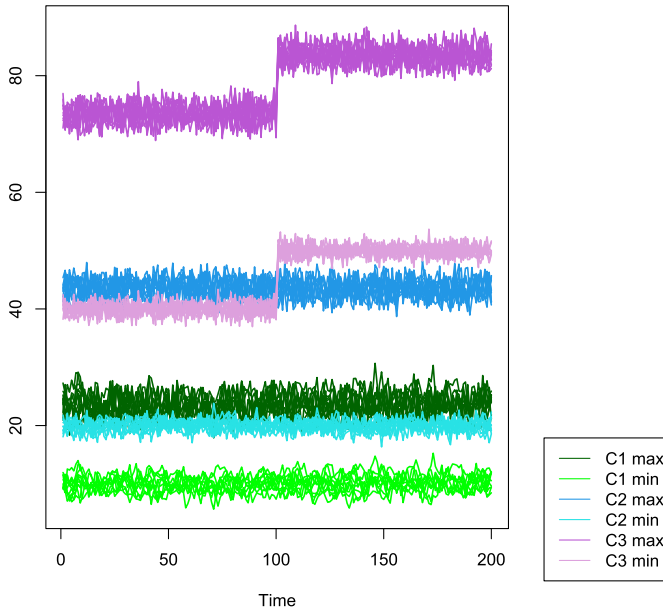
This normalization implies that the values are in the interval  $[0, 1]$  also ensuring that the order among minima and maxima of each time series is always preserved. It has been used for all clustering methods in the simulation.

### 5.2 Benchmark methods

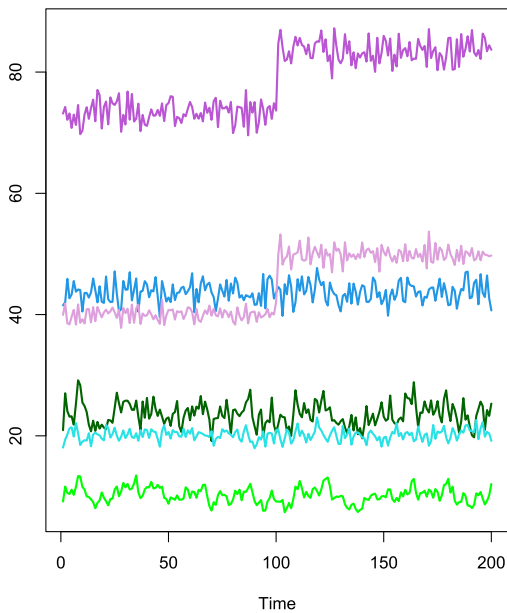
To compare our clustering method with some possible competitors, we consider the following dissimilarity measures:

$$1. d^2(x_i, \tilde{x}_c) = \sum_{t=1}^T \left( \|l_{it} - l_{ct}\|^2 + \|h_{it} - h_{ct}\|^2 \right). \tag{11}$$

The above dissimilarity (11) between the  $i$ -th time series and the  $c$ -th medoid can be obtained from the dissimilarity measure (4.1) proposed in Coppi and D'Urso



(a)



(b)

**Fig. 3** Third scenario: **a** all simulated time series **b** a sample of three simulated ITS, one per group

(2003) for LR fuzzy time trajectories by simply setting the centres  $c_{it} = 0$  and by considering the left and right spreads,  $l_{it}$  and  $r_{it}$ , as the minimum ( $l_{it}$ ) and maximum ( $h_{it}$ ) of the interval time series at time  $t$ ,  $\forall t = 1, \dots, T$  and  $i = 1, \dots, N$ .

The iterative solutions  $u_{ic}$ , for  $i = 1, \dots, N$  and  $c = 1, \dots, C$ , of the entropic *Fcmd* based on (11) are:

$$u_{ic} = \frac{1}{\sum_{c'=1}^C \left[ \frac{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T (\|l_{it} - \tilde{l}_{ct}\|^2 + \|h_{it} - \tilde{h}_{ct}\|^2) \right])\right)}{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T (\|l_{it} - \tilde{l}_{c't}\|^2 + \|h_{it} - \tilde{h}_{c't}\|^2) \right])\right)} \right]} \tag{12}$$

We name, henceforth, in the simulation study, this clustering method as *Fcmd<sub>EU</sub>*.

2. We extend the unweighted generalized Minkowski distance of order  $q$  (Billard and Diday 2006) to the case of time series as:

$$d(x_i, \tilde{x}_c) = \left( \sum_{t=1}^T \phi(x_{it}, \tilde{x}_{ct})^q \right)^{1/q} \tag{13}$$

where  $\phi(x_{it}, \tilde{x}_{ct})$  is the Ichino and Yaguchi dissimilarity<sup>1</sup> (Ichino and Yaguchi 1994) between the  $i$ -th time series and the  $c$ -th medoid at time  $t$ .

The iterative solutions  $u_{ic}$ , for  $i = 1, \dots, N$  and  $c = 1, \dots, C$ , of the entropic *Fcmd* based on the square of (13) and  $q = 2$  are:

$$u_{ic} = \frac{1}{\sum_{c'=1}^C \left[ \frac{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T \phi(x_{it}, \tilde{x}_{ct})^2 \right])\right)}{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T \phi(x_{it}, \tilde{x}_{c't})^2 \right])\right)} \right]} \tag{14}$$

We name, henceforth, in the simulation study, this clustering method as *Fcmd<sub>IY2</sub>*.

The iterative solutions  $u_{ic}$ , for  $i = 1, \dots, N$  and  $c = 1, \dots, C$ , of the entropic *Fcmd* based on (13) and  $q = 1$  are:

$$u_{ic} = \frac{1}{\sum_{c'=1}^C \left[ \frac{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T \phi(x_{it}, \tilde{x}_{ct}) \right])\right)}{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T \phi(x_{it}, \tilde{x}_{c't}) \right])\right)} \right]} \tag{15}$$

We name, henceforth, in the simulation study, this clustering method as *Fcmd<sub>IY1</sub>*.

<sup>1</sup> With  $\gamma = 0.5$ .

3. We extend the Euclidean Hausdorff distance (Billard and Diday 2006) to the case of time series as:

$$d(\mathbf{x}_i, \tilde{\mathbf{x}}_c) = \left( \sum_{t=1}^T v(x_{it}, \tilde{x}_{ct})^2 \right)^{1/2}, \quad (16)$$

where  $v(x_{it}, \tilde{x}_{ct})$  is the Hausdorff distance between the  $i$ -th time series and the  $c$ -th medoid at time  $t$ .

The iterative solutions  $u_{ic}$ , for  $i = 1, \dots, N$  and  $c = 1, \dots, C$ , of the entropic *Fcmd* based on the square of (16) are:

$$u_{ic} = \frac{1}{\sum_{c'=1}^C \left[ \frac{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T v(x_{it}, \tilde{x}_{ct})^2 \right]\right)}{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T v(x_{it}, \tilde{x}_{c't})^2 \right]\right)} \right]}. \quad (17)$$

We name, henceforth, in the simulation study, this clustering method as *Fcmd*<sub>H2</sub>. Then, we also consider:

$$d(\mathbf{x}_i, \tilde{\mathbf{x}}_c) = \sum_{t=1}^T v(x_{it}, \tilde{x}_{ct}). \quad (18)$$

The iterative solutions  $u_{ic}$ , for  $i = 1, \dots, N$  and  $c = 1, \dots, C$ , of the entropic *Fcmd* based on (18) are:

$$u_{ic} = \frac{1}{\sum_{c'=1}^C \left[ \frac{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T v(x_{it}, \tilde{x}_{ct}) \right]\right)}{\exp\left(\frac{1}{p} \left[ \sum_{t=1}^T v(x_{it}, \tilde{x}_{c't}) \right]\right)} \right]}. \quad (19)$$

We name, henceforth, in the simulation study, this clustering method as *Fcmd*<sub>H1</sub>.

We argue that, to the best of our knowledge, these benchmark methods are, however, a novelty in the state of the art of fuzzy clustering.

### 5.3 Simulation results

All the clustering methods, i.e. *Fcmd*<sub>ITS</sub>, *Fcmd*<sub>IY2</sub>, *Fcmd*<sub>IY1</sub>, *Fcmd*<sub>IH2</sub>, *Fcmd*<sub>IH1</sub> and *Fcmd*<sub>EU</sub> respectively, for each scenario, have been applied to 100 simulated datasets by setting  $C \in \{2, 3\}$  choosing the best  $C^*$  according to the Fuzzy silhouette Index (FS, Campello and Hruschka 2006), a well-known internal validity criterion that lies in  $[-1, 1]$ , so that the higher is the value of the Fuzzy Silhouette index, the

**Table 1** Ratio between the mean distance of benchmark method and that of  $Fcmd_{ITS}$  over 100 simulations

Method	Scenario 1	Scenario 2	Scenario 3
$Fcmd_{IY2}$	7.5	4.89	1.7
$Fcmd_{IY1}$	18.83	15.59	13.82
$Fcmd_{H2}$	10.96	10.37	3.24
$Fcmd_{H1}$	23.73	22.69	19.33
$Fcmd_{EU}$	16.22	12.35	4.05

better is the assignment of the units to the  $C$  clusters. For each setting, we considered 100 random restarts and set the maximum number of iterations to 1000.

To assess the impact of the fuzziness parameter, we ran  $Fcmd_{ITS}$  by varying  $p \in \{0.05, 0.10, 0.15, 0.40\}$ . It is worth noting that the effect of  $p$  in terms of degree of fuzziness of the partition also depends on the scaling of the dissimilarity matrix. In these simulations, the benchmark methods are somewhat favoured because they have a larger scale thus leading to a less blurred partition while keeping the same value of  $p$ .

To take this into account and to provide comparable results as far as possible, we consider  $p \in \{0.10, 0.20, 0.30, 0.80\}$  for all the benchmark methods whose ratio between its mean distance and that of the proposed clustering method, over 100 simulations, was above 10 (see Table 1 for the details).

To evaluate the performance, we compared the 100 obtained partitions with the true one by means of the Fuzzy Adjusted Rand Index (ARI, Campello 2007), an external validation criterion that lies in  $[-1, 1]$ : the higher is its value, the higher is the agreement between the compared partitions.

The simulation results for all clustering methods are summarised in Table 2: for each scenario and for increasing values of  $p$ , the table reports the mean and standard error of the Fuzzy ARI index as well as the number of times, based on the FS value, the method leads to choose  $C^* = 3$  over 100 simulated data set (the column named *Rate*). Then, in Figs. 4, 5 and 6 the Fuzzy ARI distribution associated with each clustering method and setting is also shown by means of the violin plots.

Notice that  $Fcmd_{ITS}$  performs very well in all scenarios, including the second and the third, where the minima and maxima time series have different widths, sometimes overlapping. This positive evidence comes from considering that it shows a very good performance both in terms of ARI and the choice of the right  $C$ . In fact, a correct assessment of the quality of performance has to be based on the combination of both measures.

Moreover, in the third scenario, the role of  $p$  is more evident: the higher its value, the lower the performance, since we have compared a crisp partition with a fuzzy one (the low value of the Fuzzy ARI in the third scenario is only due to the huge effect that  $p = 0.40$  has on the membership degree).

The comparison with the other benchmarks leads to the following considerations. For the first scenario, based on the results in Table 2 and also looking at the distribution of the fuzzy ARI in Fig. 4,  $Fcmd_{H1}$  can be considered as a valid competitor, although

**Table 2** Simulation results: mean, standard error of the Fuzzy ARI index and the number of times, based on the FS value, the method leads to choose  $C^* = 3$  over 100 simulations for all clustering methods and all scenarios

	p	Scenario I			Scenario II			Scenario III		
		Mean	s.e.	Rate	Mean	s.e.	Rate	Mean	s.e.	Rate
<i>Fcmd<sub>ITS</sub></i>	<i>p</i> <sub>1</sub>	0.999	0.001	100	0.997	0.001	100	0.998	0.000	100
	<i>p</i> <sub>2</sub>	0.999	0.001	100	0.996	0.001	100	0.985	0.001	100
	<i>p</i> <sub>3</sub>	0.998	0.001	100	0.995	0.001	100	0.952	0.002	100
	<i>p</i> <sub>4</sub>	0.915	0.001	100	0.913	0.007	100	0.640	0.006	100
<i>Fcmd<sub>IY2</sub></i>	<i>p</i> <sub>1</sub>	0.640	0.033	40	0.997	0.001	100	0.564	0.007	6
	<i>p</i> <sub>2</sub>	0.535	0.038	39	0.991	0.006	99	0.551	0.004	14
	<i>p</i> <sub>3</sub>	0.443	0.042	35	0.991	0.006	99	0.521	0.005	47
	<i>p</i> <sub>4</sub>	0.219	0.034	20	0.954	0.014	95	0.436	0.004	99
<i>Fcmd<sub>IY1</sub></i>	<i>p</i> <sub>1</sub>	0.974	0.011	96	0.997	0.001	100	0.763	0.021	56
	<i>p</i> <sub>2</sub>	0.969	0.012	96	0.997	0.001	100	0.720	0.021	47
	<i>p</i> <sub>3</sub>	0.947	0.016	95	0.997	0.001	100	0.673	0.019	40
	<i>p</i> <sub>4</sub>	0.855	0.024	93	0.996	0.002	100	0.532	0.011	77
<i>Fcmd<sub>H2</sub></i>	<i>p</i> <sub>1</sub>	0.985	0.011	98	0.997	0.001	100	0.781	0.023	69
	<i>p</i> <sub>2</sub>	0.980	0.012	98	0.992	0.005	99	0.713	0.023	60
	<i>p</i> <sub>3</sub>	0.994	0.004	99	0.987	0.007	99	0.644	0.022	56
	<i>p</i> <sub>4</sub>	0.964	0.017	97	0.952	0.018	96	0.530	0.018	68
<i>Fcmd<sub>H1</sub></i>	<i>p</i> <sub>1</sub>	0.999	0.001	100	0.997	0.001	100	0.938	0.011	97
	<i>p</i> <sub>2</sub>	0.999	0.001	100	0.997	0.001	100	0.914	0.015	92
	<i>p</i> <sub>3</sub>	0.999	0.001	100	0.997	0.001	100	0.870	0.019	84
	<i>p</i> <sub>4</sub>	0.999	0.001	100	0.996	0.001	100	0.635	0.023	56
<i>Fcmd<sub>EU</sub></i>	<i>p</i> <sub>1</sub>	0.956	0.013	90	0.937	0.019	91	0.716	0.020	50
	<i>p</i> <sub>2</sub>	0.933	0.019	87	0.943	0.018	92	0.650	0.018	39
	<i>p</i> <sub>3</sub>	0.932	0.019	87	0.941	0.019	92	0.613	0.016	38
	<i>p</i> <sub>4</sub>	0.927	0.020	90	0.855	0.029	84	0.505	0.010	79

it is favoured by the scaling of its dissimilarity matrix (see Table 1), which reduces the influence of  $p$  with respect to *Fcmd<sub>ITS</sub>*.

For the second scenario, looking at the Table 2 and Fig. 5, the performance of *Fcmd<sub>ITS</sub>* is fairly comparable in particular with that of *Fcmd<sub>Y1</sub>* and *Fcmd<sub>H1</sub>* respectively, although both favoured again by the scaling of the dissimilarity.

The third scenario is the one for which the best performance of our proposal is evident (see Table 2 and Fig. 6). The only slightly comparable clustering method is again the one based on the Hausdorff distance, i.e. *Fcmd<sub>H1</sub>*, because it is the only



one that achieves good results both in terms of Fuzzy ARI and identification of the true number of clusters.<sup>2</sup>

Furthermore, we point out that, unlike point-to-point dissimilarity measures, our clustering method is the only one able to deal with time series of different lengths, thus broadening its applicability. Based on these very good results, in the next section, we apply the proposed clustering method to a real data set and we highlight that the choice of the best value of the fuzziness parameter  $p$  strictly depends on the scaling of the dissimilarity used and the degree of separation among groups; thus in practical applications, we recommend taking into account all these issues and selecting the best combinations of  $C$  and  $p$  based on some internal validity criterion.

## 6 An application to FTSE-MIB components

The proposed clustering approach has been applied to study the performance over time of the stocks that currently compose the FTSE-MIB index, which is listed on the Italian Stock Exchange owned by the London Stock Exchange.

The stocks composing the index can vary over time, based on their market capitalization and liquidity; moreover any stock may never account for more than 15% of the index.

We argue that, in the case of financial time series, the analysis of point-valued prices does not allow capturing the dynamics of prices' volatility, which is proxied by price range (Parkinson 1980; Chou et al. 2010). In this study, hence, we considered the monthly minimum and maximum prices of the FTSE-MIB components, spanning from October 2018 to October 2022, shown in Fig. 7.

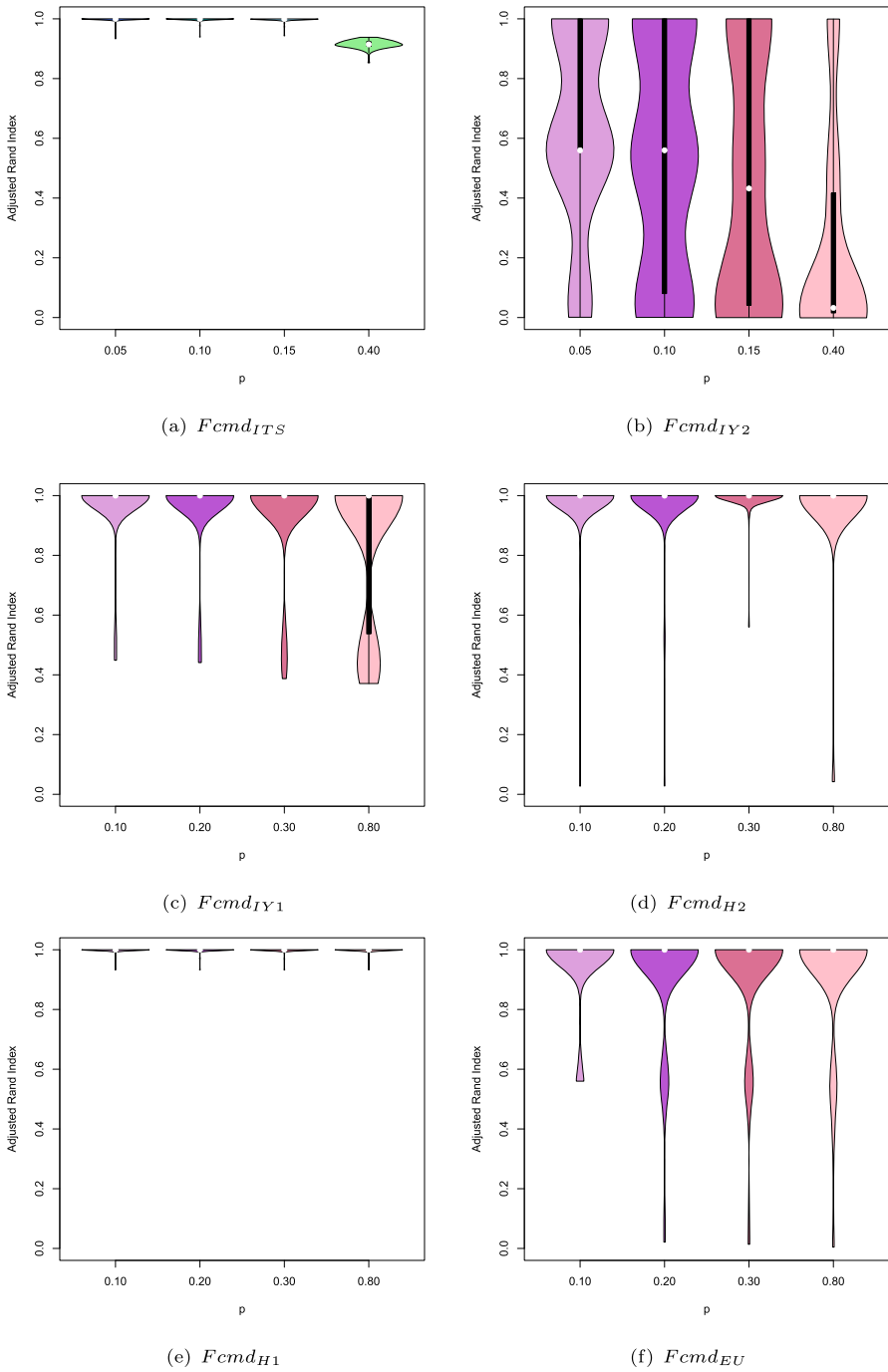
In this regard, we can highlight that almost all the stocks experimented large reduction in the prices around the first pandemic period. However, not all the stocks' prices reverted to their pre-shock values, with some that incremented their values (e.g. AMP.MI) and others that observed huge losses (e.g. BPE.MI). Some heterogeneity can be also observed considering the magnitude of the prices' changes due to the COVID-19 shock. Therefore, we use cluster analysis to deeply insight about common patterns in these stocks.

The ITS have been normalized using the proposed procedure in the *remark 2*. The best solution, i.e. the optimal number of clusters  $C^*$ , has been chosen based on the combination of  $C$  and the  $p$  that maximizes the Fuzzy Silhouette index. Considering  $C \in \{2, \dots, 10\}$  and  $p \in \{0.05, 0.08, 0.10, 0.12\}$ , we computed the Fuzzy Silhouette index accordingly.

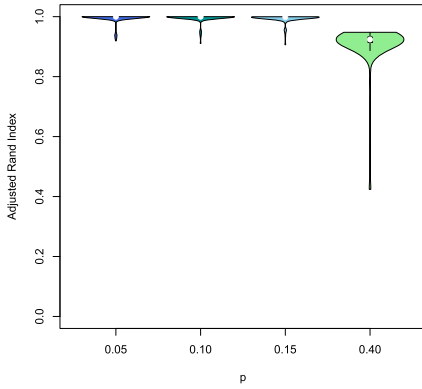
As can be seen from Fig. 8, regardless of the value of  $p$ , the best partition is for  $C^* = 2$ . Moreover, both the partition and the medoid units are stable as the value of  $p$  increases. Therefore, we focus only on the first "less fuzzy" partition, i.e. that based on  $p = 0.05$ , reported in Table 3.

The last column shows the corresponding crisp partition obtained by fixing a cut-off value for the  $u_{ic}$  equal to 0.7. The medoid units are those with the  $u_{ic}$  in bold, therefore

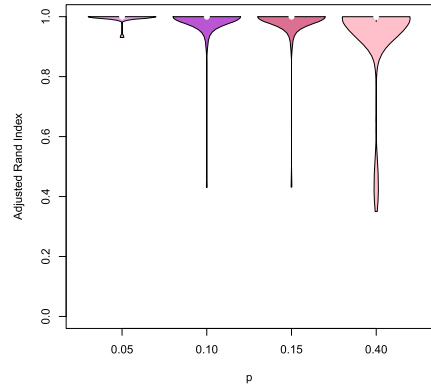
<sup>2</sup> In this respect, in the same scenario, the value of *Rate* equal to 0.99 for *Fcmdy2* leads to a misleading conclusion if it is not compared with the corresponding value of Fuzzy ARI, which is very low, thus meaning that the partition in 3 groups is not well identified.



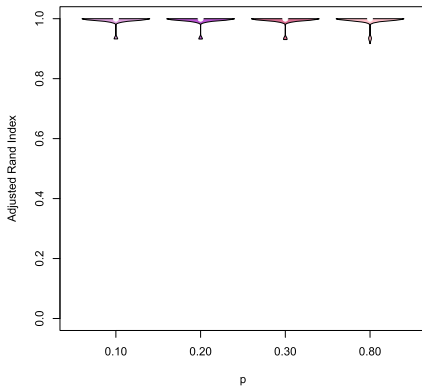
**Fig. 4** First scenario: violin plots of the Fuzzy ARI for all the clustering methods according to the different values of the  $p$  parameter



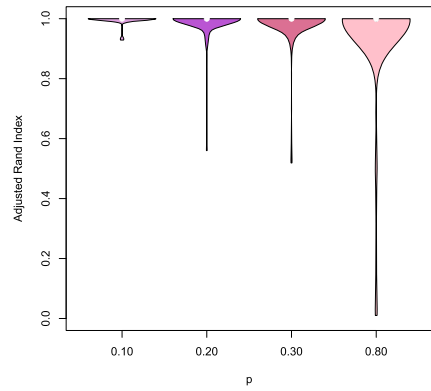
(a)  $Fcmd_{ITS}$



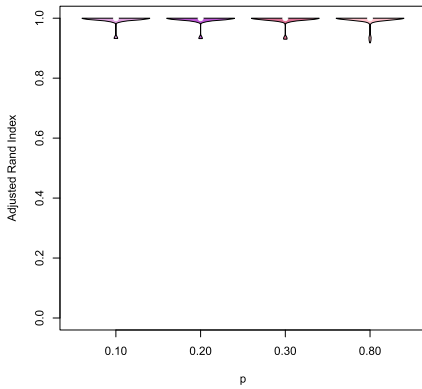
(b)  $Fcmd_{IY2}$



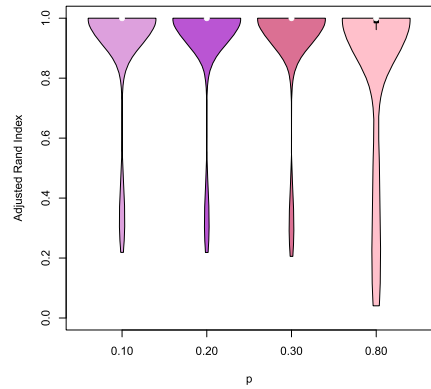
(c)  $Fcmd_{IY1}$



(d)  $Fcmd_{H2}$

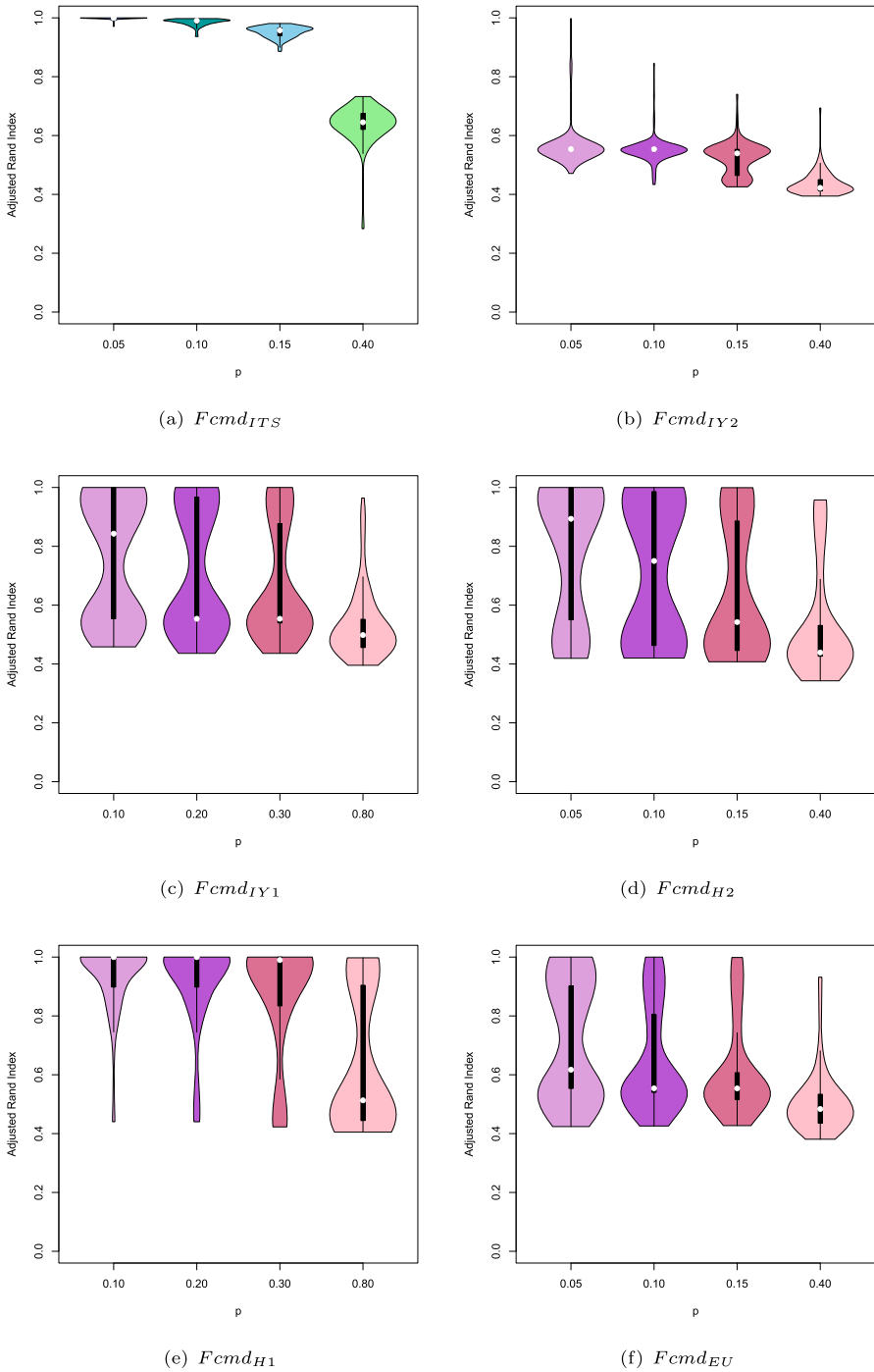


(e)  $Fcmd_{H1}$



(f)  $Fcmd_{EU}$

**Fig. 5** Second scenario: violin plots of the Fuzzy ARI for all the clustering methods according to the different values of the  $p$  parameter



**Fig. 6** Third scenario: violin plots of the Fuzzy ARI for all the clustering methods according to the different values of the  $p$  parameter



**Fig. 7** Monthly minimum and maximum prices of the FTSE-MIB components, spanning from October 2018 to October 2022. Each plot header is the stock’s acronym whose corresponding name can be found in Table 3—second column

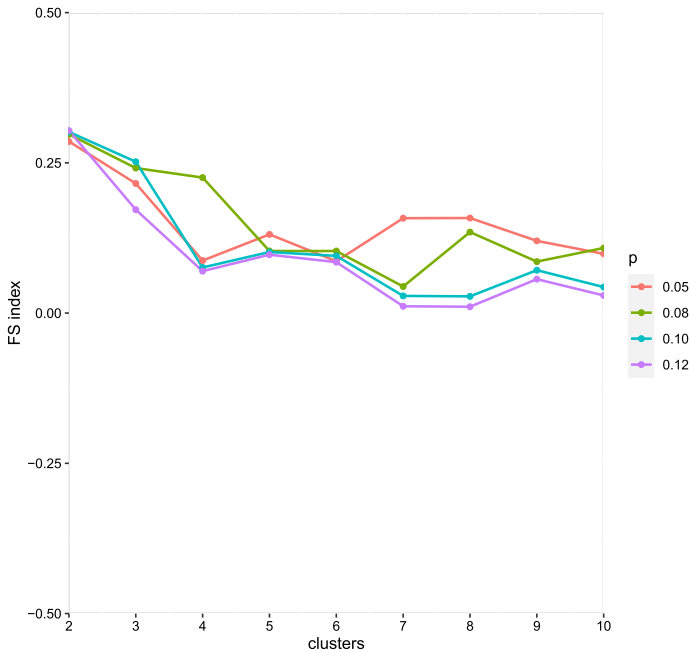
Snam (SRG.MI) and Generali (GI.MI) whose corresponding time series are plotted in Fig. 9.

Snam is Europe’s leading operator in natural gas transport and storage, with an infrastructure enabling the energy transition. It ranks among the top ten Italian listed companies by market capitalization and has been a public company, since 2001, while Generali is one of the largest global insurance and asset management providers, active in 50 countries in the world.

The medoid of the first group is characterized by a rapid increase in both minimum and maximum prices after the slump in March 2020, thus quickly reverted to the pre-shock values. Conversely, the medoid of the second group took many months to recover pre-shock values. This suggests that this medoid is characterized by a stronger persistency than the one of the first group for the months after the shock. Another interesting difference between the two cluster medoids is the magnitude of the price changes in March 2020, as the price reduction is larger for the second cluster medoid than for the first one.

The same differences between the two medoids can be retrieved in the two groups as we can see by looking at Fig. 10 that shows the crisp partition of all ITS with cut-off value 0.7,  $C = 2$  and  $p = 0.05$ .

Overall, we find in the Cluster 2 stocks characterized by a slower prices’ recovery rate after the shock and stocks that did not recover completely. In the Cluster 1,



**Fig. 8** Fuzzy silhouette index according to  $C \in \{2, \dots, 10\}$  and  $p \in \{0.05, 0.08, 0.10, 0.12\}$

differently, we find stocks whose shock impact has been lower than those placed in Cluster 2 and/or with very quick recovery rates. Moreover, in terms of trend, Cluster 1 includes stocks with positive trends pre-shock that were not affected negatively by the COVID-19 pandemic (e.g. PRY.MI, REC.MI or TRN.MI). Another interesting feature of clusters' composition, is that stocks in Cluster 1 show larger average price range compared with those in Cluster 2. In particular, the former have an average price range equal to 4.81 while the latter to 1.08. This means that, on average, Cluster 1 groups together more volatile stocks.

## 7 Concluding remarks

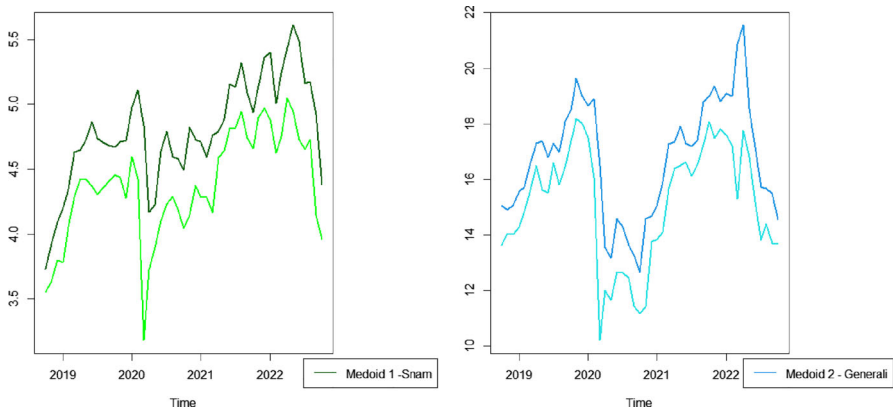
The interest in clustering complex data, such as Interval-valued time series, is today a need rather than only an opportunity.

The increasing availability of this type of data must be seen as a resource that may lead research towards more advanced statistical challenges. The use of summaries as the mean or median can produce a loss of information in terms of intra-series variability. The evolution of the range of variation of a variable over time is particularly relevant and could not be assumed to be ignored unless in some specific cases. Handling such symbolic data in the clustering process has not yet been thoroughly studied, indeed few works deal with it.

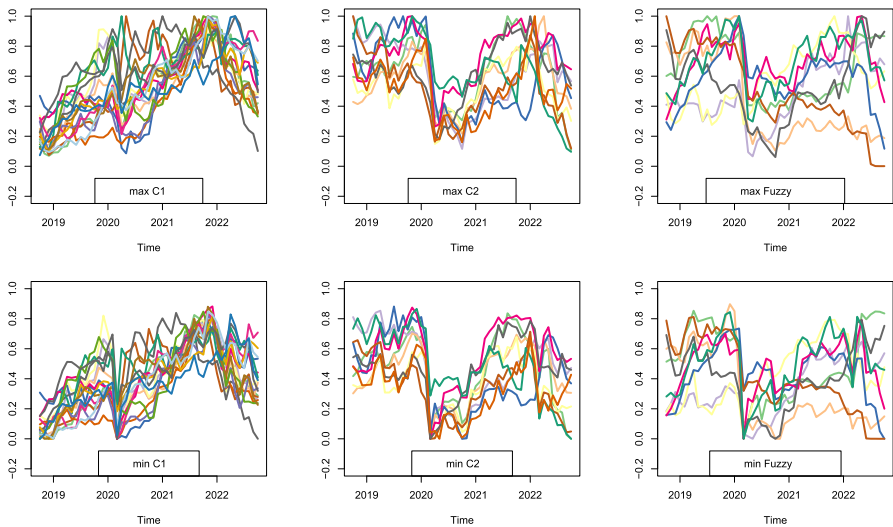
**Table 3** Fuzzy partition with  $C = 2$  and  $p = 0.05$ 

Stock acronym	Stock name	p=0.05		Crisp partition
		Cluster 1	Cluster 2	
A2A.MI	A2A	0.12	0.88	2
AMP.MI	Amplifon	1.00	0.00	1
ATL.MI	Atlantia	0.57	0.43	Fuzzy
AZM.MI	Azimut	0.89	0.11	1
BAMI.MI	Banco BPM	0.50	0.50	Fuzzy
BGN.MI	Banca Generali	0.97	0.03	1
BMED.MI	Banca Mediolanum	0.84	0.16	1
BPE.MI	BPER Banca	0.67	0.33	Fuzzy
CNHI.MI	CNH Industrials N.V.	0.93	0.07	1
CPR.MI	Campari	0.95	0.05	1
DIA.MI	Diasorin	1.00	0.00	1
ENEL.MI	Enel	0.94	0.06	1
ENI.MI	Eni	0.05	0.95	2
FBK.MI	FinecoBank	0.68	0.32	Fuzzy
<b>G.MI</b>	<b>Generali</b>	<b>0.00</b>	<b>1.00</b>	<b>2</b>
HER.MI	Hera	0.34	0.66	Fuzzy
IG.MI	Italgas	0.59	0.41	Fuzzy
INW.MI	INWIT	0.90	0.10	1
IP.MI	Interpump Group	1.00	0.00	1
ISP.MI	Intesa Sanpaolo	0.02	0.98	2
LDO.MI	Leonardo	0.03	0.97	2
MB.MI	Mediobanca	0.18	0.82	2
MONC.MI	Moncler	1.00	0.00	1
PIRC.MI	Pirelli	0.17	0.83	2
PRY.MI	Prysmian	1.00	0.00	1
PST.MI	Poste italiane	0.85	0.15	1
RACE.MI	Ferrari	1.00	0.00	1
REC.MI	Recordati	1.00	0.00	1
SPM.MI	Saipem	0.60	0.40	Fuzzy
<b>SRG.MI</b>	<b>Snam</b>	<b>1.00</b>	<b>0.00</b>	<b>1</b>
STLA.MI	Stellantis	0.16	0.84	2
STM.MI	STMicroelectronics	1.00	0.00	1
TEN.MI	Tenaris	0.63	0.37	Fuzzy
TIT.MI	Telecom Italia	0.02	0.98	2
TRN.MI	Terna	0.99	0.01	1
UCG.MI	UniCredit	0.02	0.98	2
UNI.MI	Unipol	0.39	0.61	Fuzzy

Last column reports the crisp partition when using the cut-off 0.7



**Fig. 9** The medoids units



**Fig. 10** Crisp partition with cut-off value 0.7,  $C = 2$  and  $p = 0.05$

We propose a new fuzzy clustering method suitable for ITS that not only fills a gap in the literature but also, as shown by simulations, overcomes some well-known dissimilarity measures for interval-valued data. Furthermore, unlike point-to-point dissimilarity measures, our clustering method can be used for time series of different lengths, broadening its applicability. Therefore, we propose a new fuzzy clustering method suitable for ITS to enrich the existing literature. Simulation results have been very promising and the application has revealed the goodness of our proposed clustering technique even with real data.

As a further development of this work, we will extend our methodological proposal by considering new dissimilarity measures as well as the possibility to define a metric



robust against outliers or noisy data. Another interesting research perspective to be explored is the extension of the proposed fuzzy methods to other types of complex structures of data such as count (de Nailly et al. 2023; Roick et al. 2021) and categorical (López-Oriona et al. 2023) time series.

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

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

**Human rights** This article does not contain any studies with human participants or animals performed by any of the authors.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alonso AM, Maharaj EA (2006) Comparison of time series using subsampling. *Comput Stat Data Anal* 50(10):2589–2599
- Alonso AM, D'Urso P, Gamboa C et al (2021) Cophenetic-based fuzzy clustering of time series by linear dependency. *Int J Approx Reason* 137:114–136
- Berndt D (1994) Using dynamic time warping to find patterns in time series. In: AAAI-94 Workshop on knowledge discovery in databases
- Billard L, Diday E (2006) Symbolic data analysis. Conceptual statistics and data mining. Wiley, Chichester
- Caiado J, Crato N (2010) Identifying common dynamic features in stock returns. *Quant Finance* 10(7):797–807
- Caiado J, Crato N, Peña D (2006) A periodogram-based metric for time series classification. *Comput Stat Data Anal* 50(10):2668–2684
- Caiado J, Crato N, Peña D (2009) Comparison of times series with unequal length in the frequency domain. *Commun Stat Simul Comput* 38(3):527–540
- Caiado J, Maharaj EA, D'Urso P (2015) Time-series clustering. In: Hennig C, Meila M, Murtagh F et al (eds) *Handbook of cluster analysis*, vol 12. Chapman and Hall/CRC
- Campello RJ (2007) A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern Recogn Lett* 28(7):833–841
- Campello JR, Hruschka ER (2006) A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets Syst* 157(21):2858–2875
- Cerqueti R, D'Urso P, De Giovanni L et al (2022) Ingarch-based fuzzy clustering of count time series with a football application. *Mach Learn Appl* 10(100):417
- Chou RY, Chou H, Liu N (2010) Range volatility models and their applications in finance. Springer, Berlin
- Coppi R, D'Urso P (2002) Fuzzy k-means clustering models for triangular fuzzy time trajectories. *Stat Methods Appl* 11:21–40
- Coppi R, D'Urso P (2003) Three-way fuzzy clustering models for LR fuzzy time trajectories. *Comput Stat Data Anal* 43(2):149–177

- de Carvalho FdA, Simões EC (2017) Fuzzy clustering of interval-valued data with city-block and Hausdorff distances. *Neurocomputing* 266:659–673
- De Carvalho FdA, Tenório CP (2010) Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets Syst* 161(23):2978–2999
- De Carvalho FdA, Brito P, Bock HH (2006a) Dynamic clustering for interval data based on  $l_2$  distance. *Comput Stat* 21(2):231–250
- De Carvalho FdA, De Souza RM, Chavent M et al (2006b) Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recogn Lett* 27(3):167–179
- De Luca G, Zuccolotto P (2011) A tail dependence-based dissimilarity measure for financial time series clustering. *Adv Data Anal Classif* 5(4):323–340
- De Luca G, Zuccolotto P (2017) Dynamic tail dependence clustering of financial time series. *Stat Pap* 58(3):641–657
- de Nailly P, Côme E, Oukhellou L et al (2023) Multivariate count time series segmentation with “sums and shares” and Poisson lognormal mixture models: a comparative study using pedestrian flows within a multimodal transport hub. *Adv Data Anal Classif* 1–37
- Disegna M, D’Urso P, Durante F (2017) Copula-based fuzzy clustering of spatial time series. *Spat Stat* 21:209–225
- Durante F, Pappadà R, Torelli N (2014) Clustering of financial time series in risky scenarios. *Adv Data Anal Classif* 8:359–376
- Durante F, Pappadà R, Torelli N (2015) Clustering of time series via non-parametric tail dependence estimation. *Stat Pap* 56(3):701–721
- D’Urso P (2005) Fuzzy clustering for data time arrays with inlier and outlier time trajectories. *IEEE Trans Fuzzy Syst* 13(5):583–604
- D’Urso P, Maharaj EA (2009) Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets Syst* 160(24):3565–3589
- D’Urso P, Maharaj EA (2012) Wavelets-based clustering of multivariate time series. *Fuzzy Sets Syst* 193:33–61
- D’Urso P, Cappelli C, Di Lallo D et al (2013) Clustering of financial time series. *Physica A* 392(9):2114–2129
- D’Urso P, De Giovanni L, Massari R (2015a) Time series clustering by a robust autoregressive metric with application to air pollution. *Chemom Intell Lab Syst* 141:107–124
- D’Urso P, De Giovanni L, Massari R (2015b) Trimmed fuzzy clustering for interval-valued data. *Adv Data Anal Classif* 9(1):21–40
- D’Urso P, De Giovanni L, Massari R (2016) Garch-based robust clustering of time series. *Fuzzy Sets Syst* 305:1–28
- D’Urso P, Maharaj EA, Alonso AM (2017a) Fuzzy clustering of time series using extremes. *Fuzzy Sets Syst* 318:56–79
- D’Urso P, Massari R, De Giovanni L et al (2017b) Exponential distance-based fuzzy clustering for interval-valued data. *Fuzzy Optim Decis Mak* 16(1):51–70
- D’Urso P, De Giovanni L, Massari R (2018) Robust fuzzy clustering of multivariate time trajectories. *Int J Approx Reason* 99:12–38
- D’Urso P, De Giovanni L, Massari R (2021a) Trimmed fuzzy clustering of financial time series based on dynamic time warping. *Ann Oper Res* 299(1):1379–1395
- D’Urso P, García-Escudero LA, De Giovanni L et al (2021b) Robust fuzzy clustering of time series based on b-splines. *Int J Approx Reason* 136:223–246
- D’Urso P, De Giovanni L, Maharaj EA et al (2023) Wavelet-based fuzzy clustering of interval time series. *Int J Approx Reason* 152:136–159
- Everitt SBS, Landau Leese M (2001) Cluster analysis. Arnold Press, London
- García-Escudero LA, Gordaliza A (2005) A proposal for robust curve clustering. *J Classif* 22(2):185–201
- Hwang H, DeSarbo WS, Takane Y (2007) Fuzzy clusterwise generalized structured component analysis. *Psychometrika* 72(2):181–198
- Ichino M, Yaguchi H (1994) Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Trans Syst Man Cybern* 24(4):698–708. <https://doi.org/10.1109/21.286391>
- Kejžar N, Korenjak-Černe S, Batagelj V (2021) Clustering of modal-valued symbolic data. *Adv Data Anal Classif* 15(2):513–541
- Krishnapuram R, Joshi A, Yi L (1999) A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. International fuzzy systems conference (FUZZIEEE99). IEEE, Seoul, pp 1281–1286

- Krishnapuram R, Joshi A, Nasraoui O et al (2001) Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Trans Fuzzy Syst* 9(4):595–607
- Lafuente-Rego B, D'Urso P, Vilar JA (2020) Robust fuzzy clustering based on quantile autocovariances. *Stat Pap* 61(6):2393–2448
- Li R, Mukaidono M (1995) A maximum entropy approach to fuzzy clustering. In: Proceedings of the fourth IEEE conference on fuzzy systems (FUZZ-IEEE/IFES '95), pp 2227–2232
- Li RP, Mukaidono M (1999) Gaussian clustering method based on maximum-fuzzy-entropy interpretation. *Fuzzy Sets Syst* 102(2):253–258
- López-Oriona A, D'Urso P, Vilar JA et al (2022a) Quantile-based fuzzy c-means clustering of multivariate time series: robust techniques. *Int J Approx Reason* 150:55–82
- López-Oriona A, D'Urso P, Vilar JA et al (2022b) Spatial weighted robust clustering of multivariate time series based on quantile dependence with an application to mobility during covid-19 pandemic. *IEEE Trans Fuzzy Syst* 30(9):3990–4004. <https://doi.org/10.1109/TFUZZ.2021.3136005>
- López-Oriona A, Vilar JA, D'Urso P (2022c) Quantile-based fuzzy clustering of multivariate time series in the frequency domain. *Fuzzy Sets Syst* 443:115–154. From Learning to Modeling and Control
- López-Oriona Á, Vilar JA, D'Urso P (2023) Hard and soft clustering of categorical time series based on two novel distances with an application to biological sequences. *Inf Sci* 624:467–492
- Maharaj AE, D'Urso P (2011) Fuzzy clustering of time series in the frequency domain. *Inf Sci* 181(7):1187–1211
- Maharaj AE, D'Urso P, Galagedera DU (2010) Wavelet-based fuzzy clustering of time series. *J Classif* 27(2):231–275
- Maharaj EA, Teles P, Brito P (2019) Clustering of interval time series. *Stat Comput* 29(5):1011–1034
- Miyamoto S, Mukaidono M (1997) Fuzzy c-means as a regularization and maximum entropy approach. In: Proc. of 7th international fuzzy systems association world congress (IFSA'97), II, pp 86–92
- Montanari A, Calò DG (2013) Model-based clustering of probability density functions. *Adv Data Anal Classif* 7:301–319
- Noirhomme-Fraiture M, Brito P (2011) Far beyond the classical data models: symbolic data analysis. *Stat Anal Data Min ASA Data Sci J* 4(2):157–170
- Otranto E (2008) Clustering heteroskedastic time series by model-based procedures. *Comput Stat Data Anal* 52(10):4685–4698
- Otranto E (2010) Identifying financial time series with similar dynamic conditional correlation. *Comput Stat Data Anal* 54(1):1–15
- Otranto E, Mucciardi M (2019) Clustering space-time series: Fstar as a flexible star approach. *Adv Data Anal Classif* 13:175–199
- Parkinson M (1980) The extreme value method for estimating the variance of the rate of return. *J Bus* 61–65
- Piccolo D (1990) A distance measure for classifying ARIMA models. *J Time Ser Anal* 11(2):153–164
- Roick T, Karlis D, McNicholas PD (2021) Clustering discrete-valued time series. *Adv Data Anal Classif* 15:209–229
- Umbleja K, Ichino M, Yaguchi H (2021) Hierarchical conceptual clustering based on quantile method for identifying microscopic details in distributional data. *Adv Data Anal Classif* 15:407–436
- Velichko V, Zagoruyko N (1970) Automatic recognition of 200 words. *Int J Man Mach Stud* 2:223–234
- Vilar JA, Lafuente-Rego B, D'Urso P (2018) Quantile autocovariances: a powerful tool for hard and soft partitioning of time series. *Fuzzy Sets Syst* 340:38–72
- Xiong Y, Yeung DY (2004) Time series clustering with ARMA mixtures. *Pattern Recogn* 37(8):1675–1689