ORIGINAL PAPER

A hierarchical clustering method for random intervals based on a similarity measure

Ana Belén Ramos-Guajardo[1]

Abstract
A new clustering method for random intervals that are measured in the same units over the same group of individuals is provided. It takes into account the similarity degree between the expected values of the random intervals that can be analyzed by means of a two-sample similarity bootstrap test. Thus, the expectations of each pair of random intervals are compared through that test and a p-value matrix is finally obtained. The suggested clustering algorithm considers such a matrix where each p-value can be seen at the same time as a kind of similarity between the random intervals. The algorithm is iterative and includes an objective stopping criterion that leads to statistically similar clusters that are different from each other. Some simulations to show the empirical performance of the proposal are developed and the approach is applied to two real-life situations.

Keywords Random intervals · Similarity measure · Bootstrap test · p-values matrix · Clustering approach

1 Introduction

Interval-valued observations appear in many real-life situations as those involving fluctuations, subjective perceptions, ranges, censored or grouped data, to mention only a few (Billard and Diday 2003; D'Urso and Giordani 2004; Hudgens 2005; Horowitz and Manski 2006; Kreinovich et al. 2007; Magnac and Maurin 2008; Ramos-Guajardo and Grzegorzewski 2016).

Random intervals (RIs for short) have been shown to model and handle suitably such kind of data in different settings as, for instance, regression analysis, time series, hypothesis testing and clustering (Blanco-Fernández et al. 2013; Cappelli et al. 2013;

✉ Ana Belén Ramos-Guajardo
 ramosana@uniovi.es

1 Department of Statistics, O.R. and D.M., University of Oviedo, Federico García Lorca 18, 33007 Oviedo, Spain

 ⓐ Springer

De Carvalho et al. 2006a; D'Urso et al. 2015; González-Rodríguez et al. 2009; Ramos-Guajardo and Blanco-Fernández 2017; Ramos-Guajardo et al. 2020; Sinova et al. 2012). This work is focused on those situations in which a random process generating interval-valued outcomes is considered, and classical statistics are applied on them. Therefore, the approach proposed here is soundly applicable when the intervals generated by the random process either represent precise objects or they are imprecise descriptions or perceptions provided by an observer. Thus, intervals themselves are considered, without losing any information delivered by the interval structure.

Diverse (crisp) hierarchical and non-hierarchical clustering methods for grouping interval-valued data have been developed in the literature (see, for instance, (Hajjar and Hamdan 2013; Park et al. 2020; Rodriguez and De Carvalho 2019) and the references therein). Most of these methods have been elaborated in the context of symbolic data analysis. Concerning hierarchical methods, agglomerative and divisive methods have been introduced in Chavent (1998), Gowda and Ravi (1995), Xu et al. (2018). Besides, in the partitioning clustering framework several methods have been proposed based on different techniques, including dynamic clustering, city-block distances and self-organizing maps, to name a few (see Chavent and Lechevallier 2002; De Carvalho et al. 2006a, b; Hajjar and Hamdan 2013; Ralambondrainy 1995). Note that all mentioned methods are useful for classifying individuals but not variables.

Most of the approaches to the clustering of real variables found in the literature are of hierarchical type (Nicolau and Bacelar-Nicolau 1998) and vary according to either the nature of the variables under study or to the choice of the similarity measure. The usefulness of a hierarchical clustering method for variables focuses, for example, on being able to replace a set of variables that are closely related by a single representative variable or even by a synthetic variable by means of a principal component analysis approach (El-Shaarawi and Piegorsch 2002). Such a clustering method also allows arranging variables into homogeneous clusters in order to get meaningful structures. Two examples of the clustering of real variables can be found, for instance, in Chavent et al. (2012). The first one entails the classification of 10 sports performed by 41 athletes (according to their performances), and the second one is devoted to the classification of 31 variables describing some properties of 21 French wines.

In the real framework, a frequently used technique to cluster a set of variables consists of calculating similarities between these variables by means of correlation measures. However, the space of intervals is not linear but semilinear due to the lack of symmetric element with respect to the Minkowski addition, and the extension of the concept of covariance (and therefore of correlation) is not straightforward derived. Some alternatives to the employment of correlation measures for clustering real variables have been proposed, for instance, in Kojadinovic (2004), Maharaj (1996). Maharaj (1996) showed the use of the p-values obtained from a two-sample test to apply later a hierarchical clustering in the framework of time series. This method inspired the one developed by González-Rodríguez et al. (2009) that, as far as we know, is the only hierarchical approach in the literature for clustering RIs measured in the same units over the same group of individuals. Specifically, their approach aims to classify fuzzy random variables (FRVs for short), which are those variables taking fuzzy data as responses, although RIs can be seen as a particular case.

The goal of the method developed in González-Rodríguez et al. (2009) in the context of RIs is to group *s* features when the available data is a matrix of intervals obtained when observing *s* independent RIs on *n* independent statistical units. This objective is achieved by taking into account the *p*-values obtained by applying a hypothesis test for checking the equality of expectations of RIs proposed in González-Rodríguez et al. (2006). Thus, the higher the *p*-value attained, the more similar are the expected values of two RIs, since each *p*-value can be seen as a measure of the similarity between two RIs measured in the same group of individuals. However, it should be noted that the multi-sample test in González-Rodríguez et al. (2006) is only valid for simple RIs, i.e, those taking on a finite quantity of values.

In this work, the method in González-Rodríguez et al. (2009) is proposed to be extended for general (not only simple) RIs measured in the same units over the same group of individuals by using a bootstrap test procedure based on a similarity measure that is able to check not only the strict equality between expectations but also a degree of similarity among them ranging from 0 and 1, where 1 means that the intervals are exactly equal [which is analogous to the method in González-Rodríguez et al. (2009)] and 0 means there is not intersection between them. Note that the concept of similarity proposed here is a geometric one in the sense that two intervals are considered more similar whenever the ratio between the part they have in common and the union of both is higher. Thus, if that ratio is equal to 1 then both intervals are completely equal. Another possibility could be to consider a distance between intervals but then the concept of geometric similarity would be lost, and both the analysis and the interpretability will be different. Then, the idea is to relax the concept of strict equality between expected values by using a measure of geometric similarity between intervals that was initially introduced by Jaccard (1901), and which is defined as a ratio of the Lebesgue measures of the intersection and the union intervals (Shawe-Taylor and Cristianini 2004). To do that, a two-sample similarity test for the expected value of random intervals suggested by Ramos-Guajardo and Blanco-Fernández (2017) can be applied to obtain the *p*-values matrix instead of the usual two-sample test for the equality of expectations of RIs. The main advantages of the new approach are the following: (1) it can be used for comparing all types of RIs and not only simple RIs, (2) two groups of RIs can be linked providing their expected values present a high degree of similarity (according to the Jaccard-based similarity coefficient), not being necessary the strict equality between such expectations.

In this framework a hierarchical clustering method for RIs will be developed allowing specifying the degree of similarity that is intended to exist at least between the expectations of the clusters that are joined together in every step. Thus, once we have prefixed a similarity grade, the idea is to detect clusters of RIs such that the RIs in each group have not necessarily equal but similar population expected values at certain extent.

The paper is organized as follows. Section 2 includes two cases study involving real-life RIs in order to motivate the proposal. In Sect. 3 some preliminaries on RIs are given. Section 4 contains the clustering criterion to be applied based on the two-sample similarity test for RIs. The suggested hierarchical clustering algorithm for RIs is provided in Sect. 5. The results of a simulation study are discussed in Sect. 6 where the empirical performance of the proposal is investigated, also in comparison

with the method in González-Rodríguez et al. (2009). The application of the approach to the real-life cases described in Sect. 2 is analyzed in Sect. 7 also in contrast to other methods by taking into account different similarity grades to set up the clusters. Finally, some concluding remarks are provided in Sect. 8.

2 Case-studies

The method to be developed in the next sections will allow us to classify random intervals measured in the same units over the same group of individuals by considering a measure of similarity between those random intervals. Below are presented two real situations in which the procedure outlined in this work can be subsequently applied.

2.1 Mathematics related beliefs questionnaire

As a first real life situation, a questionnaire including 17 statements has been proposed to a group of 117 students attending the second course of the Degree in Primary Education of the University of Cantabria (Spain). The statements are related to their mathematics related beliefs and are described below:

- *Q1* I think what I learn in math class is interesting.
- *Q2* I like what I learn in math class.
- *Q3* I am very interested in mathematics.
- *Q4* I like to do math stuff.
- *Q5* In math, I prefer challenging tasks to learn new things.
- *Q6* I hope to do well on math homework and on math tests.
- *Q7* I prefer mathematics in which I must strive to find the solution than those that do not require much effort.
- *Q8* I have found that if I have patience, I can do difficult math problems.
- *Q9* I am sure I can learn how to solve the most difficult math problem.
- *Q10* I do not have to try too hard to understand math.
- *Q11* Compared to other classmates I think I am good at math.
- *Q12* I think I will do well in math this course.
- *Q13* I understand everything we have done in math this year.
- *Q14* I can usually do math problems that take a long time to solve.
- *Q15* I can understand even the most difficult subject to teach me in mathematics.
- *Q16* By doing my best in math I try to show my teacher that I am better than other classmates.
- *Q17* I put a lot of effort into math to show the teacher and my classmates how good I am.

The survey respondents used interval data in a scale ranging from 0 to 10 for answering the questionnaire, where 0 represents a totally disagree to each statement whereas 10 represents a totally agree to it. Thus, each response in terms of an interval corresponds to the set of values that the student considers compatible with his/her opinion at some extent, i.e., the student considers that his/her opinion cannot be outside of this set. In this context, the interval responses of the 117 students about his/her

opinion about each statement i, for $i \in \{1, \ldots, 17\}$ are given in the same units and can be modeled through an RI **QI**. The purpose is, therefore, to classify the 17 RIs based on a measure of similarity between their expected values, as it will be seen in the following sections.

2.2 Daily concentration of PM10 particles in Spanish cities

The second study addresses the analysis of the daily concentration of PM10 particles from June to September 2019 (122 days) in 18 Spanish autonomous cities. The Spanish Ministry of the Environment defines the PM10 particles such as dust particles, ash, soot, metal, cement or pollen, scattered in the atmosphere, whose diameter varies between 2.5 and 10 μm. Nowadays, scientists consider such particles are the most severe environmental pollution problem, due to their serious affections to the respiratory tract and the lung.

The hourly data on the concentration of PM10 particles from June to September of 2019 in the different cities have been extracted from the website https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/datos/Datos_oficiales_2019.aspx. The cities analyzed were Vitoria, Palma de Mallorca, Santiago de Compostela, Logroño, Madrid, Murcia, Pamplona, Oviedo, Las Palmas de Gran Canaria, Santa Cruz de Tenerife, Santander, Sevilla, Toledo, Valencia, Valladolid, Zaragoza, Ceuta and Barcelona. The measurement of the daily concentration of these particles in each city i, for $i \in \{1, \ldots, 18\}$ is modeled through an RI X_i whose data $X_{ij} = [X_{ij}^m, X_{ij}^M]$ are determined by the minimum and maximum concentration of PM10 particles reached during day j (X_{ij}^m and X_{ij}^M, respectively), for $i \in \{1, \ldots, 122\}$. Thus, each interval reflects the daily variability of the pollution by PM10 particles in the corresponding city. Moreover, each of the cities can be identified with an RI and the aim is to be able to classify the 18 Spanish cities according to their daily PM10 concentration levels.

3 Preliminaries

Let $\mathcal{K}_c(\mathbb{R})$ be the family of non-empty closed and bounded intervals of $\mathbb{R}$. An interval $A \in \mathcal{K}_c(\mathbb{R})$ can be characterized by either its mid /spr representation or its inf/sup representation. The first one, which is so that $A = [\text{mid } A \pm \text{spr } A]$ with mid $A \in \mathbb{R}$ being the mid-point or center and spr $A \geq 0$ being the spread or radius of A, has been shown to be more operative than the second one, given by $A = [\inf A, \sup A]$, and a valuable tool for different statistical purposes (see, for instance, Blanco-Fernández et al. 2011; D'Urso and Giordani 2004; Sinova et al. 2012).

The usual arithmetic between intervals is based on the Minkowski's addition and the product by a scalar. In terms of the (mid , spr) representation it is given by

$$A_1 + \lambda A_2 = [(\text{mid } A_1 + \lambda \text{mid } A_2) \pm (\text{spr } A_1 + |\lambda| \text{spr } A_2)],$$

where $A_1, A_2 \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. It should be remarked that the space $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear but semilinear due to the lack of symmetric element with respect to the Minkowski addition. Thus, in general, $A + (-1)A \neq 0$ unless $A = \{a\}$ is a singleton. Moreover, the Minkowski difference between two intervals $A + (-1)B$ does not coincide with the natural difference $A - B$, since it does not fulfill the addition/subtraction simplification, i.e. $(A + (-1)B) + B \neq A$ (see Blanco-Fernández et al. 2013 for more details).

On the other hand, the Lebesgue measure (or length) of $A \in \mathcal{K}_c(\mathbb{R})$ so that $A \neq \emptyset$ is given by $\lambda(A) = 2\mathrm{spr}\, A$ whereas the Lebesgue measure of the empty set is clearly $\lambda(\emptyset) = 0$. In addition, according to (Shawe-Taylor and Cristianini 2004), if $A, B \in \mathcal{K}_c(\mathbb{R})$ the Lebesgue measure of the intersection between A and B can be expressed as follows:

$$\lambda(A \cap B) = \max \left\{ 0, \min \left\{ 2\mathrm{spr}\, A, 2\mathrm{spr}\, B, \mathrm{spr}\, A + \mathrm{spr}\, B - |\mathrm{mid}\, A - \mathrm{mid}\, B| \right\} \right\}. \tag{1}$$

A *random interval* (RI for short) is a random variable modelling those situations in which intervals on $\mathcal{K}_c(\mathbb{R})$ are provided as outcomes. Mathematically, given a probability space $(\Omega, \mathcal{A}, P)$, an RI is a Borel measurable mapping $X : \Omega \to \mathcal{K}_c(\mathbb{R})$ w.r.t. the well-known Hausdorff metric on $\mathcal{K}_c(\mathbb{R})$ (Matheron 1975). Equivalently, X is an RI if both $\mathrm{mid}\, X, \mathrm{spr}\, X : \Omega \to \mathbb{R}$ are real-valued random variables and $\mathrm{spr}\, X \geq 0$ a.s.-$[P]$.

The expected value of X introduced by Aumann (1965) is expressed as the interval $E([\mathrm{mid}\, X \pm \mathrm{spr}\, X]) = [E(\mathrm{mid}\, X) \pm E(\mathrm{spr}\, X)]$ whenever $\mathrm{mid}\, X, \mathrm{spr}\, X \in L^1(\Omega, \mathcal{A}, P)$, i.e., whenever $\mathrm{mid}\, X, \mathrm{spr}\, X$ are integrable functions in the probability space $(\Omega, \mathcal{A}, P)$. Given a simple random sample of X, $\{X_i\}_{i=1}^n$, the corresponding sample expectation of X is defined coherently in terms of the interval arithmetic as $\overline{X} = (1/n) \sum_{i=1}^n X_i$, and it fulfils $\overline{X} = [\overline{\mathrm{mid}\, X} \pm \overline{\mathrm{spr}\, X}]$. It should be noted that $\overline{\mathrm{mid}\, X}$ and $\overline{\mathrm{spr}\, X}$ are the sample mean of the real-valued random variables $\mathrm{mid}\, X$ and $\mathrm{spr}\, X$, respectively, i.e., $\overline{\mathrm{mid}\, X} = (1/n) \sum_{i=1}^n \mathrm{mid}\, X_i = \mathrm{mid}\, \overline{X}$ and $\overline{\mathrm{spr}\, X} = (1/n) \sum_{i=1}^n \mathrm{spr}\, X_i = \mathrm{spr}\, \overline{X}$.

On the basis of the measure defined in (1), a measure of the degree of similarity between two intervals $A, B \in \mathcal{K}_c(\mathbb{R})$ can be defined conforming to the Jaccard coefficient (Jaccard 1901) as

$$S(A, B) = \frac{\lambda(A \cap B)}{\lambda(A \cup B)}, \tag{2}$$

where $\lambda(A \cup B) = \lambda(A) + \lambda(B) - \lambda(A \cap B)$, and both A and B are presumed not to be reduced to a singleton. Undoubtedly, since $\lambda(A \cap B) \leq \lambda(A \cup B)$ it is fulfilled that $0 \leq S(A, B) \leq 1$. Moreover, $S(A, B) = 0$ iff $A \cap B = \emptyset$, $S(A, B) = 1$ if $A = B$, and $S(A, B) \in (0, 1)$ iff $A \cap B \neq \emptyset$ and $A \neq B$. Some situations considering various possible similarity degrees are gathered in Fig. 1.

Notice, for instance, that the similarity measure of intervals $A = [0, 2]$ and $B = [1, 2]$ is $1/2$, and so it is the similarity measures of A and $C = [0, 1]$, of A and $D = [0, 4]$, and of A and $E = [-2, 2]$.

Fig. 1 Different representations for the similarity degree between two intervals A and B

S(A,B)=1   S(A,B)=1/2   S(A,B)=1/4
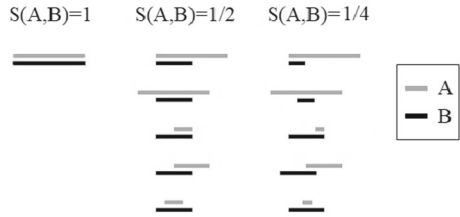
4 Clustering criterion: two-sample similarity test for RIs

In this paper we aim at developing a method to cluster k independent RIs by paying attention to their expected values.

Given a probability space $(\Omega, \mathcal{A}, P)$ and a set of s independent simple RIs $X_i : \Omega \to \mathcal{K}_c(\mathbb{R})$, for $i = 1, \ldots, s$, the purpose is to get different groups $G_k \subset \{1, \ldots, s\}$ so that the RIs belonging to the same group can be considered to be similar at some extent.

Concerning the generation process of the data, a simple random sample $\{X_{ij}\}_{j=1}^{n}$ distributed as the RI X_i is drawn for each $i \in \{i, \ldots, s\}$. Thus, in practice the data collected is gathered in an interval data matrix $[x_{ij}]_{i=1,\ldots,s}^{j=1,\ldots,n}$ where each $x_{ij} \in \mathcal{K}_c(\mathbb{R})$.

In order to test if the expected values of two RIs can be considered to be similar at some extent, a version for different sample sizes of the two-sample similarity test for the expected value of RIs proposed in Ramos-Guajardo and Blanco-Fernández (2017) is suggested. It should be noted that the version of the test for different sample sizes is necessary to be able to establish such a comparison between clusters that include different number of RIs and, therefore, different sample sizes.

First, the study developed for equal sample sizes is recalled. Let $X, Y : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ be two independent RIs such that $\text{spr}\, E(X) > 0$ and $\text{spr}\, E(Y) > 0$, and belonging to the class

$$\mathcal{P} = \left\{ X : \Omega \to \mathcal{K}_c(\mathbb{R}) \,|\, \sigma_{\text{mid}\, X}^2 < \infty \,,\, 0 < \sigma_{\text{spr}\, X}^2 < \infty \,\wedge\, \sigma_{\text{mid}\, X,\, \text{spr}\, X}^2 \neq \sigma_{\text{mid}\, X}^2 \sigma_{\text{spr}\, X}^2 \right\},$$

in order to assure the non-singularity of the covariance matrices that relate mid X and spr X, and mid Y and spr Y. Note that in the expression of $\mathcal{P}$, $\sigma_{\text{mid}\, X}^2$ and $\sigma_{\text{spr}\, X}^2$ denote the variances of the real variables mid X and spr X, whereas $\sigma_{\text{mid}\, X,\, \text{spr}\, X}^2$ is the square of the covariance between the real variables mid X and spr X.

To test the null hypothesis $H_0 : S(E(X), E(Y)) \geq d$, where $d \in [0, 1]$ is the degree of similarity between $E(X)$ and $E(Y)$, is necessary to compute the probability that the empirical version of the Jaccard-based similarity coefficient introduced in (2) is less than d, whenever the null hypothesis is satisfied. However, such a computation is not immediate, since it is necessary to consider the measure of the intersection between two intervals is given by the expression in (1). Thus, based on this intersection measure, the test can also be expressed as follows by considering the mid/spr characterization of

intervals (note that there has also been a change in the inequalities of the hypotheses):

$$H_0 : \max \Big\{ d \operatorname{spr} E(Y) - \operatorname{spr} E(X), d \operatorname{spr} E(X) - \operatorname{spr} E(Y),$$
$$(1 + d) \left| \operatorname{mid} E(X) - \operatorname{mid} E(Y) \right| + (d - 1) \left(\operatorname{spr} E(X) + \operatorname{spr} E(Y) \right) \Big\} \leq 0;$$
$$H_1 : \max \Big\{ d \operatorname{spr} E(Y) - \operatorname{spr} E(X), d \operatorname{spr} E(X) - \operatorname{spr} E(Y),$$
$$(1 + d) \left| \operatorname{mid} E(X) - \operatorname{mid} E(Y) \right| + (d - 1) \left(\operatorname{spr} E(X) + \operatorname{spr} E(Y) \right) \Big\} > 0. \tag{3}$$

Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be two simple random samples drawn from X and Y, respectively. The test statistic is defined as

$$T = \sqrt{n} \max \Big\{ d \operatorname{spr} \overline{Y_n} - \operatorname{spr} \overline{X_n}, d \operatorname{spr} \overline{X_n} - \operatorname{spr} \overline{Y_n},$$
$$(1 + d) \left| \operatorname{mid} \overline{X_n} - \operatorname{mid} \overline{Y_n} \right| + (d - 1) \left(\operatorname{spr} \overline{X_n} + \operatorname{spr} \overline{Y_n} \right) \Big\} . \tag{4}$$

Considerably large values of the statistic above would refer to a similarity degree between the expected values of X and Y lower than d and far from d and, as a consequence, a small p-value, so the null hypothesis should be rejected. The asymptotic distribution of the statistic T under H_0 is discussed in Lemma 1 and its proof is provided in Ramos-Guajardo and Blanco-Fernández (2017). The corresponding limit distribution is based on the bivariate normal distributions $Z = (z_1, z_2)' \equiv \mathcal{N}_2 \Big(\mathbf{0}, \Sigma_1 \Big)$ and $U = (u_1, u_2)' \equiv \mathcal{N}_2 \Big(\mathbf{0}, \Sigma_2 \Big)$, where Σ_1 is the covariance matrix for the random vector $(\operatorname{mid} X, \operatorname{spr} X)$ and Σ_2 is the corresponding one for $(\operatorname{mid} Y, \operatorname{spr} Y)$.

Lemma 1 (Ramos-Guajardo and Blanco-Fernández 2017) *For $n \in \mathbb{N}$, let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be two random samples independent and equally distributed from X and Y, respectively, and defined on the probability space $(\Omega, \mathcal{A}, P)$. Let T_n be defined as in (4). If $X, Y \in \mathcal{P}$, then:*

(a) If $\operatorname{spr} E(X) = d \operatorname{spr} E(Y)$ and $\operatorname{mid} E(X) - \operatorname{mid} E(Y) = (1 - d)\operatorname{spr} E(Y)$, then it is fulfilled that

$$T \xrightarrow{\mathcal{L}} \max\{du_2 - z_2, (1 + d)(z_1 - u_1) + (d - 1)(z_2 + u_2)\}.$$

(b) If $\operatorname{spr} E(X) = d \operatorname{spr} E(Y)$ and $-\operatorname{mid} E(X) + \operatorname{mid} E(Y) = (1 - d)\operatorname{spr} E(Y)$, then it is fulfilled that

$$T \xrightarrow{\mathcal{L}} \max\{du_2 - z_2, (1 + d)(u_1 - z_1) + (d - 1)(z_2 + u_2)\}.$$

(c) If $d \operatorname{spr} E(X) = \operatorname{spr} E(Y)$ and $\operatorname{mid} E(X) - \operatorname{mid} E(Y) = \dfrac{(1 - d)}{d} \operatorname{spr} E(Y)$, then it is fulfilled that

$$T \xrightarrow{\mathcal{L}} \max\{dz_2 - u_2, (1 + d)(z_1 - u_1) + (d - 1)(z_2 + u_2)\}.$$

(d) If d spr $E(X) = $ spr $E(Y)$ and $-$mid $E(X) +$ mid $E(Y) = \dfrac{(1-d)}{d}$ spr $E(Y)$, then it is fulfilled that

$$T \xrightarrow{\mathcal{L}} \max\{dz_2 - u_2, (1+d)(u_1 - z_1) + (d-1)(z_2 + u_2)\}.$$

It is also recalled that other situations under H_0 different than the ones shown in the lemma above imply a weak convergence of T to a limit distribution stochastically bounded by one of those provided in the lemma. In addition, since the limit distribution of T depends on X, an alternative X-dependent statistic, called T', based on the ideas described in Ramos-Guajardo (2015) is proposed, and the consistency and power of the test are also studied (see Lemma 2 in Ramos-Guajardo and Blanco-Fernández (2017) for more details).

The asymptotic distribution of the proposed statistic is not effective to compute critical values, so a bootstrap procedure is recommended. Specifically, a residual-type bootstrap based on the studies of Shao and Tu in Shao and Tu (1995) has been employed. Thus, if we consider bootstrap samples from X and Y, i.e. $\{X_i^*\}_{i=1}^n$ and $\{Y_i^*\}_{i=1}^n$ being chosen randomly and with replacement from $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively, the bootstrap statistic is defined as follows:

$$
\begin{aligned}
T^* = \max\Big\{ & \sqrt{n}\left(d\left(\text{spr}\,\overline{Y_n^*} - \text{spr}\,\overline{Y_n}\right) + \text{spr}\,\overline{X_n} - \text{spr}\,\overline{X_n^*}\right) + \min\left(0, n^{1/4}(\text{spr}\,\overline{Y_n} - \text{spr}\,\overline{X_n})\right), \\
& \sqrt{n}\left(d\left(\text{spr}\,\overline{X_n^*} - \text{spr}\,\overline{X_n}\right) + \text{spr}\,\overline{Y_n} - \text{spr}\,\overline{Y_n^*}\right) + \min\left(0, n^{1/4}(\text{spr}\,\overline{X_n} - \text{spr}\,\overline{Y_n})\right), \\
& \sqrt{n}\Big((1+d)\left(\text{mid}\,\overline{X_n^*} - \text{mid}\,\overline{X_n} + \text{mid}\,\overline{Y_n} - \text{mid}\,\overline{Y_n^*}\right) \\
& + (d-1)\left(\text{spr}\,\overline{X_n^*} - \text{spr}\,\overline{X_n} + \text{spr}\,\overline{Y_n^*} - \text{spr}\,\overline{Y_n}\right)\Big) \\
& + \min\left(0, n^{1/4}(\text{mid}\,\overline{X_n} - \text{mid}\,\overline{Y_n})\right), \\
& \sqrt{n}\Big((1+d)\left(\text{mid}\,\overline{X_n} - \text{mid}\,\overline{X_n^*} + \text{mid}\,\overline{Y_n^*} - \text{mid}\,\overline{Y_n}\right) \\
& + (d-1)\left(\text{spr}\,\overline{X_n^*} - \text{spr}\,\overline{X_n} + \text{mid}\,\overline{Y_n^*} - \text{mid}\,\overline{Y_n}\right)\Big) \\
& + \min\left(0, n^{1/4}(\text{mid}\,\overline{Y_n} - \text{mid}\,\overline{X_n})\right)\Big\}.
\end{aligned}
$$

As it has been described in Ramos-Guajardo and Blanco-Fernández (2017), the minima-terms included in T^* are useful to determine the parts on its expression which have the influence to the maximum depending on the situation under H_0. It has been also shown that T^* converges weekly (almost sure) to the same distributions as the ones provided in Lemma 1 under the same requisites. In addition, the bootstrap approach is also consistent and, in practice, the distribution of T^* is approximated by considering the Monte Carlo method. The p-value of the bootstrap test would be the probability that $T^* > T$. In practice, the bootstrap procedure consists of getting a large sample B of values of the statistic, $T_1^*, \ldots, T_B^*$, close enough to the population distribution, and the p-value is approximated by the proportion of values in $T_1^*, \ldots, T_B^*$ which are greater than or equal to the value of the statistic T.

Remark 1 Note that the results above are valid for equal sample sizes. Nevertheless, if the size of the samples drawn from X and Y differs, the same statistics T and T^* are valid as long as the corresponding sample sizes, say n and m, satisfy that $n/(n+m) \to p_1 \in (0, 1)$ and $m/(n+m) \to (1 - p_1) \in (0, 1)$, as n and m tends to ∞.

5 Hierarchical clustering method for RIs

The two sample similarity test for RIs given in Sect. 4 can be used to get a matrix of similarities to apply later a hierarchical clustering for interval-valued data following the ideas in González-Rodríguez and others (González-Rodríguez et al. 2009). In that paper the authors start from a matrix **P** holding the p-values obtained by using a multi-sample test for random fuzzy sets to all the pairs of random fuzzy sets. The procedure to get the p-values to cluster RIs is provided in the next lines.

Assume that $X_1 \ldots, X_s$ are s independent RIs taking values in $\mathcal{K}_c(\mathbb{R})$. For each $j \in \{1, \ldots, s\}$, let $\{X_{jl}\}_{l=1}^{n_j}$ be a simple random sample obtained from the RI X_j, and let $N \in \mathbb{N}$ be the overall sample size.

Maharaj (1996) showed the use of the p-value of a two-sample test in order to get a matrix of p-values, which can be interpreted as a matrix of similarities, to apply later a hierarchical clustering in the framework of time series, whenever the measure employed is symmetric and non-negative. The same procedure can be imitated by considering the similarity measure proposed in (2) due to its symmetry and no negativity. Thus, the two-sample similarity bootstrap test described in Sect. 4 can be applied to each pair $i_1, i_2 \in 1, \ldots, s$ in order to get a matrix of p-values $\mathbf{P} = [p_{i_1, i_2}]$, where p_{i_1, i_2} is the p-value of the test $H_0 : S(E(X_{i_1}), E(X_{i_2})) \geq d$. Of course, it is necessary to fix in advance a value $d \in [0, 1]$ as a threshold so that two RIs are grouped together whenever H_0 is not rejected at the usual significance levels.

It is well-known that a p-value close to zero suggests rejecting the null hypothesis H_0. Hence, the closer to zero the p-value is, the lower the similarity between $E(X_{j_1})$ and $E(X_{j_2})$ with respect to d is. On the other hand, one can assume that if a p-value is high, the similarity between $E(X_{j_1})$ and $E(X_{j_2})$ cannot be far below d, otherwise the test would be rejected. In fact, the higher the p-value is, the more similar $E(X_{j_1})$ and $E(X_{j_2})$ are and, in particular, if $j_1 = j_2$ the p-value should be equal to 1. In this way the obtained matrix **P** can be seen as a kind of similarity matrix. In addition, consistently with (González-Rodríguez et al. 2009), $d(j_1, j_2) = 1 - p_{j_1, j_2}$ is a distance that quantifies how similar $E(X_{j_1})$ and $E(X_{j_2})$ are with respect to d, so that the lower such a distance is, the more reason $E(X_{j_1})$ and $E(X_{j_2})$ must belong to the same cluster. Those distances can be used to obtain a hierarchical clustering method with any of the usual linkage criteria.

The proposed hierarchical clustering procedure starts from the singleton clusters, $C_l = X_l$ for $l \in \{1, \ldots, s\}$, and computes the matrix of p-values **P** resulting from the corresponding two-sample similarity bootstrap tests $H_0 : S(E(X_{i_1}), E(X_{i_2})) \geq d$ for $i_1, i_2 \in 1, \ldots, s$. For that, the bootstrap procedure is applied to each pair of RIs by resampling from the individuals composing each variable to form a bootstrap sampling and computing the bootstrap statistic based on the midpoints and spreads

of the resampling. At each stage, the procedure merges the two RIs with the highest
p-value, whenever it is greater than a fixed significance level α. The two RIs involved
in the new cluster form a unique RI (with double sample size) and the resampling in the
next step is made over the individuals that make up such a joint RI. Then, the measures
needed to start again the process are computed by considering the new cluster. Finally,
the stopping criterion of the procedure is determined on the basis of the significance
level α, which usually ranges from 0.01 to 0.1.

The reported method requires the computation of the bootstrap p-values for the
two-sample similarity tests associated to all pairs of clusters at every step, which
implies a moderate computational cost. However, one of the main advantages of this
method in contrast to the one proposed in González-Rodríguez et al. (2009) is that the
former is more flexible in the sense that the clustering of RIs is based on their degree
of similarity and not on their exact equality. In this way, two random intervals could be
grouped whenever their expectations present a high degree of similarity (according to
the Jaccard-based similarity coefficient), and it is not necessary that both expectations
are exactly equal. Besides, the multi-sample test employed in González-Rodríguez
et al. (2009) is valid for fuzzy random variables or random intervals taking on a finite
quantity of values, whereas the method proposed in the manuscript allows to classify
all types of random intervals.

The clustering algorithm is displayed below.

Clustering algorithm

Step 1 Fix a similarity degree d, a significance level α and the number of bootstrap
replications B. Let $C_i = X_i$ be the singleton clusters for $i \in \{1, \ldots, s\}$.

Step 2 For each $i \in \{1, \ldots, s\}$, obtain a large number B of bootstrap samples
$\{X_{i1}^{*b}, \ldots, X_{in_i}^{*b}\}_{b=1}^{B}$ drawn from $\{X_{i1}, \ldots, X_{in_i}\}$.

Step 3 For each cluster C_i, compute the values mid $\overline{X_i}$ and spr $\overline{X_i}$.

Step 4 For each cluster C_i and each value $b \in \{1, \ldots, B\}$, compute the mid and
spread of the sample mean of the values of each resample b, i.e., mid $\overline{X_i^{*b}}$ and
spr $\overline{X_i^{*b}}$.

Step 5 For each pair of clusters C_1 and C_2 compute the associate p-value p_{C_1, C_2} by
using Monte Carlo method as follows:

  5.1 Initialize a counter $W = 0$ and compute the value of the two-sample similarity
  test statistic

$$
\begin{aligned}
T = \sqrt{n_1} \max \Big\{ & D \operatorname{spr} \overline{X_{C_2}} - \operatorname{spr} \overline{X_{C_1}}, d \operatorname{spr} \overline{X_{C_1}} - \operatorname{spr} \overline{X_{C_2}}, \\
& (1 + d) \big| \operatorname{mid} \overline{X_{C_1}} - \operatorname{mid} \overline{X_{C_2}} \big| \\
& + (d - 1) \big(\operatorname{spr} \overline{X_{C_1}} + \operatorname{spr} \overline{X_{C_2}} \big) \Big\} .
\end{aligned}
$$

  5.2 For each value $b \in \{1, \ldots, B\}$, compute the value of the bootstrap statistic

$$
\begin{aligned}
T^{*b} = \max \Big\{ & \sqrt{n_1} \Big(d \Big(\operatorname{spr} \overline{X_{C_2}^{*b}} - \operatorname{spr} \overline{X_{C_2}} \Big) + \operatorname{spr} \overline{X_{C_1}} - \operatorname{spr} \overline{X_{C_1}^{*b}} \Big) \\
& + \min \Big(0, n_1^{1/4} (\operatorname{spr} \overline{X_{C_2}} - \operatorname{spr} \overline{X_{C_1}}) \Big),
\end{aligned}
$$

$$\sqrt{n_1}\left(d\left(\mathrm{spr}\,\overline{X_{C_1}^{*b}} - \mathrm{spr}\,\overline{X_{C_1}}\right) + \mathrm{spr}\,\overline{X_{C_2}} - \mathrm{spr}\,\overline{X_{C_2}^{*b}}\right)$$

$$+ \min\left(0, n_1^{1/4}(\mathrm{spr}\,\overline{X_{C_1}} - \mathrm{spr}\,\overline{X_{C_2}})\right),$$

$$\sqrt{n_1}\left((1+d)\left(\mathrm{mid}\,\overline{X_{C_1}^{*b}} - \mathrm{mid}\,\overline{X_{C_1}} + \mathrm{mid}\,\overline{X_{C_2}} - \mathrm{mid}\,\overline{X_{C_2}^{*b}}\right)\right.$$

$$+ (d-1)\left(\mathrm{spr}\,\overline{X_{C_1}^{*b}} - \mathrm{spr}\,\overline{X_{C_1}} + \mathrm{spr}\,\overline{X_{C_2}^{*b}} - \mathrm{spr}\,\overline{X_{C_2}}\right)\right)$$

$$+ \min\left(0, n_1^{1/4}(\mathrm{mid}\,\overline{X_{C_1}} - \mathrm{mid}\,\overline{X_{C_2}})\right),$$

$$\sqrt{n_1}\left((1+d)\left(\mathrm{mid}\,\overline{X_{C_1}} - \mathrm{mid}\,\overline{X_{C_1}^{*b}} + \mathrm{mid}\,\overline{X_{C_2}^{*b}} - \mathrm{mid}\,\overline{X_{C_2}}\right)\right.$$

$$+ (d-1)\left(\mathrm{spr}\,\overline{X_{C_1}^{*b}} - \mathrm{spr}\,\overline{X_{C_1}} + \mathrm{mid}\,\overline{X_{C_2}^{*b}} - \mathrm{mid}\,\overline{X_{C_2}}\right)\right)$$

$$+ \min\left(0, n_1^{1/4}(\mathrm{mid}\,\overline{X_{C_2}} - \mathrm{mid}\,\overline{X_{C_1}})\right)\Bigg\},$$

and IF $T^{*b} > T$, THEN W=W+1.

5.3 Compute the bootstrap p-value as $p_{C_1,C_2} = W/B$.

Step 6 Find the maximum value of the matrix of p-values **P** obtained in the previous step, say **max(P)**, and label by $C_{\mathbf{P}_1}$ and $C_{\mathbf{P}_2}$ one of the pairs of clusters with this p-value.

Step 7 IF

$$\mathbf{max(P)} > \alpha$$

THEN
join the clusters $C_{\mathbf{P}_1}$ and $C_{\mathbf{P}_2}$ in a new cluster $C = C_{\mathbf{P}_1} \cup C_{\mathbf{P}_2}$ and repeat Steps 3 and 4 for this new cluster by computing the mid and spread of the sample mean of the elements in C, and of the mean of the elements of each resampling $b \in \{1, \ldots, B\}$. Compute the p-values between the new joint cluster and the other ones as in Step 5.1.-5.3. and go to Step 6.
ELSE
Stop.

Theorem 1 demonstrate that the clusters obtained by applying the approach described in the previous algorithm are statistically pairwise different. The result is supported by next proposition.

Proposition 1 *Let $C = \{X_1\} \cup \cdots \cup \{X_{m_1}\}$ be a cluster and let Y be a random interval so that $S(E(X_i), E(Y)) \geq d$ for a fixed similarity degree $d \in [0, 1]$ and $i \in \{1, \ldots, m_1\}$. Then, $S(E(C), E(Y)) \geq d$.*

Proof As $S(E(X_i), E(Y)) \geq d$ for $i \in \{1, \ldots, m_1\}$, then, H_0 in (3) is fulfilled for each pair $\{X_i, Y\}$. In addition,

$$\min\{\mathrm{spr}\,E(X_i)\}_{i=1}^{m_1} \leq \mathrm{spr}\,E(C) \leq \max\{\mathrm{spr}\,E(X_i)\}_{i=1}^{m_1},$$
$$\min\{\mathrm{mid}\,E(X_i)\}_{i=1}^{m_1} \leq \mathrm{mid}\,E(C) \leq \max\{\mathrm{mid}\,E(X_i)\}_{i=1}^{m_1}.$$

These conditions imply that

$$\xi_1 = d\,\mathrm{spr}\,E(Y) - \mathrm{spr}\,E(C) \le d\,\mathrm{spr}\,E(Y) - \min\{\mathrm{spr}\,E(X_i)\}_{i=1}^{m_1} \le 0;$$
$$\xi_2 = d\,\mathrm{spr}\,E(C) - \mathrm{spr}\,E(Y) \le d\,\max\{\mathrm{spr}\,E(X_i)\}_{i=1}^{m_1} - \mathrm{spr}\,E(Y) \le 0.$$

In addition, the same conditions together with the triangular inequality entails that

$$m_1 \cdot \xi_3 = m_1 \cdot \Big((1+d)|\mathrm{mid}\,E(C) - \mathrm{mid}\,E(Y)| + (d-1)(\mathrm{spr}\,E(C) + \mathrm{spr}\,E(Y))\Big)$$
$$\le m_1\Big((1+d)\Big|\sum_{i=1}^{m_1}(\mathrm{mid}\,E(X_i) - \mathrm{mid}\,E(Y))\Big| + (d-1)\sum_{i=1}^{m_1}(\mathrm{spr}\,E(X_i) + \mathrm{spr}\,E(Y))\Big)$$
$$\le 0$$

Therefore, $\xi_3 \le 0$ and $\max\{\xi_1, \xi_2, \xi_3\} \le 0$, which means that H_0 in (3) is fulfilled for $\{C, Y\}$, so the result is given. □

Theorem 1 *Given a similarity degree $d \in [0,1]$ and a fixed significance level α, the clustering algorithm provided in this section produces clusters $\{C_i\}_{i=1}^{k}$ such that for all $k_1 \ne k_2 \in \{1, \ldots, k\}$ there exist $i_1 \in C_{k_1}$ and $i_2 \in C_{k_2}$ satisfying that $S(E(X_{i_1}), E(X_{i_2})) < d$ at the significance level α.*

Proof Let C_{k_1} and C_{k_2} be two clusters obtained by following the algorithm above such that $C_{k_1} = \{X_1\} \cup \ldots \cup \{X_{m_1}\}$ and $C_{k_2} = \{Y_1\} \cup \ldots \cup \{Y_{m_2}\}$.

Suppose that $S(E(X_i), E(Y_j)) \ge d$ for $i \in \{1, \ldots, m_1\}$ and $j \in \{1, \ldots, m_2\}$. Then, H_0 in (3) is fulfilled for each pair $\{X_i, Y_j\}$. By Proposition 1, the inequality $S(E(C_{k_1}), E(Y_j)) \ge d$ is given for all $j \in \{1, \ldots, m_2\}$. In the same way, by applying again Proposition 1 to the random interval C_{k_1} and $\{Y_j\}_{j=1}^{m_2}$ it is satisfied that $S(E(C_{k_1}), E(C_{k_2})) \ge d$, and the algorithm would have continued by joining C_{k_1} and C_{k_2}. Thus, there must exist at least $i_1 \in C_{k_1}$ and $i_2 \in C_{k_2}$ so that $S(E(X_{i_1}), E(X_{i_2})) < d$. □

From Theorem 1 it can be concluded that the clusters are different from each other. Besides, the procedure developed here is more flexible that the one in González-Rodríguez et al. (2009) since a degree of similarity between the expected values of the RIs can be chosen to compare with and not only exact equalities are allowed to classify the RIs. Moreover, in this case it is not required neither to prefix the number of clusters nor to solve any other optimization problem since the design of the clusters is given automatically by considering the probability of type I error that can be assumed.

6 Simulation studies

In this section some simulation studies are carried out in order to evaluate the performance of the proposal. In addition, a comparison between the method developed by González-Rodríguez et al. in González-Rodríguez et al. (2009) using the ANOVA

test introduced in González-Rodríguez et al. (2012) particularizing it to RIs and the method proposed in this work is also established. From now on, the new proposal will be called ClustSim whereas the method in González-Rodríguez et al. (2009) will be called ClustANOVA. Note that the clustering algorithm used in ANOVA is the same as in González-Rodríguez et al. (2009) but considering the test and bootstrap statistics T and T^* proposed in González-Rodríguez et al. (2012) for the ANOVA test instead of the ones proposed in González-Rodríguez et al. (2006) for the multi-sample test for simple RIs.

As a first study, an inspection of the behavior of the type I error of the procedure ClustSim is developed. Thus, k populations belonging to the same cluster (i.e., whose similarity between the expectations of each pair of variables can be assumed to be greater than or equal to d) have been simulated for $d \in \{0.8, 0.9, 1\}$. Moreover, the case $d = 1$ (equality of expectations) is also compared with method ClustANOVA in terms of type I error. RIs are randomly generated based on two parameters: their midpoint and their spreads. Therefore, any real continuous random variable will be valid for modeling the behavior of the midpoints, while for modeling the behavior of spreads it is necessary to consider a real continuous and positive random variable.

As a first attempt the following distributions have been chosen without loss of generality for $j \in \{1, \ldots, k\}$:

$$\begin{aligned} \text{mid } X_j &\hookrightarrow U(0 + (j-1)(1-d)/(k-1), 1) \\ \text{spr } X_j &\hookrightarrow U(0, 1 - (j-1)(1-d)/(k-1)) \end{aligned} \tag{5}$$

As an example of the generation process, the case $d = 0.8$ and $k = 3$ is being inspected, which is so that

X_1 is so that mid $X_1 \hookrightarrow U(0, 1)$ and spr $X_1 \hookrightarrow U(0, 1)$,
X_2 is so that mid $X_2 \hookrightarrow U(0.1, 1)$ and spr $X_2 \hookrightarrow U(0, 0.9)$,
X_3 is so that mid $X_3 \hookrightarrow U(0.2, 1)$ and spr $X_3 \hookrightarrow U(0, 0.8)$.

Thus, $E(X_1) = [0, 1]$, $E(X_2) = [0.1, 1]$ and $E(X_3) = [0.2, 1]$, so we can assure that the degree of similarity between those expectations is at least 0.8 since $S(E(X_1), E(X_2)) = S(E(X_2), E(X_3)) = 0.9$ and $S(E(X_1), E(X_3)) = 0.8$.

As it has been shown that the empirical sizes of the bootstrap similarity test and the ANOVA test are quite close to the expected nominal significance levels for moderate sample sizes (see Ramos-Guajardo 2015, and González-Rodríguez et al. 2012), then we will fix $n = 100$. In addition, the algorithm has been run for 1000 simulations with 500 bootstrap replications of the bootstrap test each by considering a significance level $\alpha = 0.05$ by considering 3, 5 and 8 populations belonging to one unique cluster. The proportion of times in which the correct number of clusters is obtained in every situation whenever uniform distributions are considered for mid /spr has been collected in Table 1.

An analogous study has been carried out given alternative distributions for the midpoints and spreads. Specifically, variations from normal/beta distributions and Student's/chi-square distributions for midpoints/spreads have been considered by applying a variable generation process equivalent to the one described in Sect. 5.

Table 1 Proportion of times in which the correct number of clusters is obtained when uniform distributions are considered for mid /spr

mid /spr $\approx$ uniform/uniform				
	Method	$k = 3$	$k = 5$	$k = 8$
$d = 0.8$	ClustSim	.9712	.9754	.9782
$d = 0.9$	ClustSim	.9591	.9604	.9548
$d = 1$	ClustSim	.9035	.9084	.8951
	ClustANOVA	.8987	.9041	.9027

Table 2 Proportion of times in which the correct number of clusters is obtained when normal/beta distributions are considered for mid /spr

mid /spr $\approx$ normal/beta				
	Method	$k = 3$	$k = 5$	$k = 8$
$d = 0.8$	ClustSim	.9678	.9649	.9699
$d = 0.9$	ClustSim	.9628	.954	.9509
$d = 1$	ClustSim	.9011	.8989	.8923
	ClustANOVA	.8934	.8991	.8976

Table 3 Proportion of times in which the correct number of clusters is obtained when student's/chi-square distributions are considered for mid /spr

mid /spr $\approx$ Student's/chi-square				
	Method	$k = 3$	$k = 5$	$k = 8$
$d = 0.8$	ClustSim	.9698	.9641	.971
$d = 0.9$	ClustSim	.9601	.9576	.9633
$d = 1$	ClustSim	.8934	.8887	.8857
	ClustANOVA	.8901	.8932	.8822

The proportion of times in which the correct number of clusters is obtained in both situations is gathered in Tables 2 and 3.

From Tables 1, 2 and 3 it can be concluded that the procedure has a success rate close to the empirical size of the bootstrap method $(1 - \alpha)$ for values $d = 0.8$ and $d = 0.9$. However, it can be observed that in the case $d = 1$ the results obtained remain below $1 - \alpha$. Although it is true that the method proposed in González-Rodríguez et al. (2009) based on the multi-sample test introduced in González-Rodríguez et al. (2006) for simple RIs presents better results, it should be noted that in that case, apart from being an appropriate procedure only for a very specific case of RIs, the bootstrap test statistic includes the variability of the RIs in its denominator, while in the ClustSim and ClustANOVA methods such variability is not included in the bootstrap statistics (see Ramos-Guajardo 2015; González-Rodríguez et al. 2012). Since the inclusion of variability of RIs is not a straightforward problem, it is left as an open problem to be addressed in the near future.

On the other hand, as a second scenario different sample sizes are considered, being them $n \in \{50, 80, 100\}$, as well as $s = 10$ RIs and three similarity grades $d \in \{0.8, 0.9, 1\}$. A structure of 3 clusters is considered so that 2 RIs belong to Cluster 1, 3 RIs belong to Cluster 2 and 5 RIs belong to Cluster 3. Regarding the distinction of clusters three different situations are considered:

– *Case 1* The clusters are distinguished with respect to the midpoints of the intervals.
– *Case 2* The clusters are distinguished with respect to the midpoints and the spreads of the intervals.
– *Case 3* The clusters are distinguished with respect to the spreads of the intervals.

In addition, two levels of separation for the clusters are taken into account:

– *Case A* small separation between clusters.
– *Case B* large separation between clusters.

The details of the generation process for the midpoints and spreads of the RIs in cases $d = 0.8$, $d = 0.9$ and $d = 1$ whenever uniform distributions are considered for mid /spr are gathered on Tables 4, 5 and 6.

An anologous generation process has been carried out for the situations in which normal/beta distributions and Student's/chi-square distributions have been considered for mid /spr. However, the specific details of such a generation process have been omitted to make the section easier to follow by the reader and due to extension reasons.

As an example, the RIs obtained through the generation process for case 2.A (clusters distinguished with respect to the midpoints and spreads of the intervals and small separation between clusters) when $d = 0.8$ proposed in Table 4 fulfil the following conditions:

– The RIs involved in the first cluster (say X_1 and X_2) satisfy that $E(X_1) = [-1, 0]$ and $E(X_2) = [-1, -0.2]$, so $S(E(X_1), E(X_2)) = 0.8$.
– The RIs involved in the second cluster (say X_3, X_4 and X_5) satisfy that $E(X_3) = [-0.5, 0.5]$, $E(X_4) = [-0.45, 0.45]$ and $E(X_5) = [-0.4, 0.4]$, so the lowest similarity grade among each pair of random intervals in cluster 2 is $S(E(X_3), E(X_5)) = 0.8$.
– The RIs involved in the third cluster (say X_6, X_7, X_8, X_9 and X_{10}) satisfy that $E(X_6) = [0, 1]$, $E(X_7) = [0.05, 1]$, $E(X_8) = [0.1, 1]$, $E(X_9) = [0.15, 1]$ and $E(X_{10}) = [0.2, 1]$, so the lowest similarity grade among each pair of random intervals in cluster 3 is $S(E(X_6), E(X_{10})) = 0.8$.
– The highest similarity grade between RIs of different clusters is given by $S(E(X_1), E(X_3)) = S(E(X_3), E(X_6)) = 1/3$.

In order to gather the similarity degree between each pair of expected values of the variables involved in the different situations displayed in Tables 7, 8 and 9, the corresponding similarity degree matrices S_{case}^d are shown in Appendix A, for $d \in \{0.8, 0.9, 1\}$ and $case \in \{1.A, 2.A, 3.A, 1.B, 2.B, 3.B\}$. Note that the similarity degree matrices provided do not depend on the distributions chosen for midpoints and spreads, since the three types of RIs generation processes considered (uniform/uniform, normal/beta and Student's/chi-square) led to the same expected value for each RI, i.e., the expected value obtained for X_i, for $i \in \{1, \ldots, 10\}$, when applying an RI generation process based on uniform distributions for midpoints and spreads coincided with the one obtained when applying the alternatives normal/beta and Student's/chi-square in the different frameworks analyzed.

For every case and every sample size chosen, $r = 300$ random samples have been generated and, for each data set, $B = 500$ bootstrap replications of the two-sample similarity bootstrap test has been carried out in order to generate the p-values of the

Table 4 Generation of mids and spreads in each case for $d = 0.8$

Case	Cluster	mid /spr	Case	Cluster	mid /spr
1.A	1	U(−1, 0)/U(0,1)	1.B	1	U(−1, 0)/U(0,1)
		U(−1.11, −0.11)/U(0,1)			U(-1.11,-0.11)/U(0,1)
	2	U(−0.5, 0.5)/U(0,1)		2	U(1,2)/U(0,1)
		U(−0.4474, 0.5526)/U(0,1)			U(0.944,1.944)/U(0,1)
		U(−0.5526, 0.4474)/U(0,1)			U(1.056,2.056)/U(0,1)
	3	U(0,1)/U(0,1)		3	U(3,4)/U(0,1)
		U(0.028,1.028)/U(0,1)			U(3.028,4.028)/U(0,1)
		U(0.056,1.056)/U(0,1)			U(3.056,4.056)/U(0,1)
		U(0.083,1.083)/U(0,1)			U(3.083,4.083)/U(0,1)
		U(0.11,1.11)/U(0,1)			U(3.11,4.11)/U(0,1)
2.A	1	U(−1, 0)/U(0,1)	2.B	1	U(−1, 0)/U(0,1)
		U(−1, −0.2)/U(0,0.8)			U(−1, −0.2)/U(0,0.8)
	2	U(−0.5, 0.5)/U(0,1)		2	U(1,2)/U(0,1)
		U(−0.45, 0.45)/U(0,0.9)			U(1.05,1.95)/U(0,0.9)
		U(−0.4, 0.4)/U(0,0.8)			U(1.1,1.9)/U(0,0.8)
	3	U(0,1)/U(0,1)		3	U(3,4)/U(0,1)
		U(0.05,1)/U(0,95)			U(3.05,4)/U(0,0.95)
		U(0.1,1)/U(0,0.9)			U(3.1,4)/U(0,0.9)
		U(0.15,1)/U(0,0.85)			U(3.15,4)/U(0,0.85)
		U(0.2,1)/U(0,0.8)			U(3.2,4)/U(0,0.8)
3.A	1	U(0,1)/U(0,2.5)	3.B	1	U(0,1)/U(0,3.5)
		U(0,1)/U(0,2)			U(0,1)/U(0,2.8)
	2	U(0,1)/U(0,1)		2	U(0,1)/U(0,2)
		U(0,1)/U(0,0.9)			U(0,1)/U(0,1.8)
		U(0,1)/U(0,0.8)			U(0,1)/U(0,1.6)
	3	U(0,1)/U(0,0.25)		3	U(0,1)/U(0,25)
		U(0,1)/U(0,2375)			U(0,1)/U(0,0.2375)
		U(0,1)/U(0,225)			U(0,1)/U(0,0.225)
		U(0,1)/U(0,0.2125)			U(0,1)/U(0,0.2125)
		U(0,1)/U(0,0.2)			U(0,1)/U(0,0.2)

similarity matrix **P**. Tables 7, 8 and 9 collect the percentage of times in which the algorithm suggests 3 clusters containing the initial RIs of the generation process with respect to all the different cases proposed in Tables 4, 5 and 6 and the three types of distributions considered for mid /spr. Again, the case $d = 1$ of method ClustSim has been compared with method ClustANOVA in this framework.

From the results gathered in Tables 7, 8 and 9 we can conclude that the procedure proposed in this work determines the correct number of clusters (as well as the RIs belonging to each cluster) in most of the suggested situations. However, once again it can be seen that in the case $d = 1$ the correct detection of clusters is not so suit-

Table 5 Generation of mids and spreads in each case for $d = 0.9$

Case	Cluster	mid /spr	Case	Cluster	mid /spr
1.A	1	U(−1, 0)/U(0,1)	1.B	1	U(−1, 0)/U(0,1)
		U(−1.053, −0.053)/U(0,1)			U(−1.053, −0.053)/U(0,1)
	2	U(−0.5, 0.5)/U(0,1)		2	U(1,2)/U(0,1)
		U(−0.526, 0.474)/U(0,1)			U(0.974,1.974)/U(0,1)
		U(−0.474, 0.526)/U(0,1)			U(1.026,2.026)/U(0,1)
	3	U(0,1)/U(0,1)		3	U(3,4)/U(0,1)
		U(0.013,1.013)/U(0,1)			U(3.013,4.013)/U(0,1)
		U(0.026,1.026)/U(0,1)			U(3.026,4.026)/U(0,1)
		U(0.04,1.04)/U(0,1)			U(3.04,4.04)/U(0,1)
		U(0.053,1.053)/U(0,1)			U(3.053,4.053)/U(0,1)
2.A	1	U(−1, 0)/U(0,1)	2.B	1	U(−1, 0)/U(0,1)
		U(−1, −0.1)/U(0,0.9)			U(−1, −0.1)/U(0,0.9)
	2	U(−0.5, 0.5)/U(0,1)		2	U(1,2)/U(0,1)
		U(−0.475, 0.475)/U(0,0.95)			U(1.025,0.975)/U(0,0.95)
		U(−0.45, 0.45)/U(0,0.9)			U(1.05,1.95)/U(0,0.9)
	3	U(0,1)/U(0,1)		3	U(3,4)/U(0,1)
		U(0.025,1)/U(0,975)			U(3.025,4)/U(0,0.975)
		U(0.05,1)/U(0,0.95)			U(3.05,4)/U(0,0.95)
		U(0.075,1)/U(0,0.925)			U(3.075,4)/U(0,0.925)
		U(0.1,1)/U(0,0.9)			U(3.1,4)/U(0,0.9)
3.A	1	U(0,1)/U(0,2)	3.B	1	U(0,1)/U(0,4)
		U(0,1)/U(0,1.8)			U(0,1)/U(0,3.6)
	2	U(0,1)/U(0,1)		2	U(0,1)/U(0,2)
		U(0,1)/U(0,0.95)			U(0,1)/U(0,1.9)
		U(0,1)/U(0,0.9)			U(0,1)/U(0,1.8)
	3	U(0,1)/U(0,0.4)		3	U(0,1)/U(0,0.4)
		U(0,1)/U(0,0.39)			U(0,1)/U(0,0.39)
		U(0,1)/U(0,0.38)			U(0,1)/U(0,0.38)
		U(0,1)/U(0,0.37)			U(0,1)/U(0,0.37)
		U(0,1)/U(0,0.36)			U(0,1)/U(0,0.36)

able, which may be due to the lack of a variability-related term in the corresponding test statistics. In all cases it should be remarked that the results are better when the sample size is larger. This is because both the two-sample similarity bootstrap test and the ANOVA test perform better for moderate/large sample sizes. Finally, there are no major differences in the performance of methods ClustSim and ClustANOVA when considering one type of distributions or another for mid /spr as well as when considering differences in midpoints, in spreads, or in both.

Table 6 Generation of mids and spreads in each case for $d = 1$

Case	Cluster	mid /spr	Case	Cluster	mid /spr
1.A	1	U(−1, 0)/U(0,1)	1.B	1	U(−1, 0)/U(0,1)
	2	U(−0.5, 0.5)/U(0,1)		2	U(1,2)/U(0,1)
	3	U(0,1)/U(0,1)		3	U(3,4)/U(0,1)
2.A	1	U(−1, 0)/U(0,1)	2.B	1	U(−1, 0)/U(0,1)
	2	U(−0.4, 0.4)/U(0,0.8)		2	U(1.1,1.9)/U(0,0.8)
	3	U(0.1,1)/U(0,0.9)		3	U(3.1,4)/U(0,0.9)
3.A	1	U(0,1)/U(0,2)	3.B	1	U(0,1)/U(0,4)
	2	U(0,1)/U(0,1)		2	U(0,1)/U(0,2)
	3	U(0,1)/U(0,0.25)		3	U(0,1)/U(0,0.25)

Table 7 Percentage of times in which the algorithm suggests the true composition of clusters when uniform distributions are considered for mid /spr

mid /spr $\approx$ uniform/uniform

Case	d	Method	$n = 50$	$n = 80$	$n = 100$
1.A.	0.8	ClustSim	.9154	.9394	.9498
	0.9	ClustSim	.8912	.929	.9513
	1	ClustSim	.8456	.8765	.9028
		ClustANOVA	.8438	.8755	.903
2.A.	0.8	ClustSim	.9428	.9606	.9479
	0.9	ClustSim	.9134	.9391	.9562
	1	ClustSim	.8567	.8812	.9055
		ClustANOVA	.858	.8802	.9041
3.A.	0.8	ClustSim	.937	.9424	.9475
	0.9	ClustSim	.9213	.9399	.9495
	1	ClustSim	.8568	.8767	.9105
		ClustANOVA	.8629	.8805	.9011
1.B.	0.8	ClusSim	.9205	.9432	.9537
	0.9	ClustSim	.9145	.9392	.9509
	1	ClustSim	.8598	.8795	.8985
		ClustANOVA	.8561	.8812	.9014
2.B.	0.8	ClustSim	.9387	.9583	.9524
	0.9	ClustSim	.9271	.9365	.9542
	1	ClustSim	.8598	.8799	.9019
		ClustANOVA	.8579	.8808	.9025
3.B.	0.8	ClustSim	.9305	.9436	.9487
	0.9	ClustSim	.9209	.9312	.9471
	1	ClustSim	.8533	.8744	.8969
		ClustANOVA	.8519	.8713	.9003

Table 8 Percentage of times in which the algorithm suggests the true composition of clusters when normal/beta distributions are considered for mid /spr

mid /spr ≈ normal/beta					
Case	d	Method	$n = 50$	$n = 80$	$n = 100$
1.A.	0.8	ClustSim	.9111	.9323	.9486
	0.9	ClustSim	.8903	.9291	.9509
	1	ClustSim	.8407	.8742	.8978
		ClustANOVA	.8419	.871	.8966
2.A.	0.8	ClustSim	.9401	.9689	.9468
	0.9	ClustSim	.9106	.9377	.953
	1	ClustSim	.8533	.8792	.9028
		ClustANOVA	.8555	.8781	.9006
3.A.	0.8	ClustSim	.9356	.9409	.949
	0.9	ClustSim	.9289	.9387	.95
	1	ClustSim	.8581	.8776	.9055
		ClustANOVA	.8602	.8777	.9023
1.B.	0.8	ClusSim	.9201	.9419	.9502
	0.9	ClustSim	.9139	.9362	.9488
	1	ClustSim	.8549	.878	.8966
		ClustANOVA	.8511	.8764	.9004
2.B.	0.8	ClustSim	.9355	.9554	.9476
	0.9	ClustSim	.9219	.9333	.9502
	1	ClustSim	.8581	.8704	.8993
		ClustANOVA	.8575	.8731	.8984
3.B.	0.8	ClustSim	.9291	.9484	.9509
	0.9	ClustSim	.9199	.9289	.9452
	1	ClustSim	.8513	.8704	.8939
		ClustANOVA	.8521	.8691	.8977

7 Real-life applications

The clustering method proposed in this work (ClustSim) is applied to the real cases presented in Sect. 2 by considering several similarity grades from which the different variables will be grouped. Situations $d = 1$, $d = 0.9$ and $d = 0.8$ are explored in both scenarios and $B = 5000$ bootstrap replications have been considered in all cases. The p-value matrices computed by following the procedure in the different cases are included in Appendices B and C. Besides, a comparison between case $d = 1$ of ClustSim, the procedure ClustANOVA developed in González-Rodríguez et al. (2009) and the classical hierarchical clustering method just using the midpoints is also provided.

Table 9 Percentage of times in which the algorithm suggests the true composition of clusters when Student's/chi-square distributions are considered for mid /spr

mid /spr $\approx$ Student's/chi-squared

Case	d	Method	$n = 50$	$n = 80$	$n = 100$
1.A.	0.8	ClustSim	.9107	.9298	.9477
	0.9	ClustSim	.8893	.9278	.9489
	1	ClustSim	.8402	.8733	.8961
		ClustANOVA	.8408	.869	.8963
2.A.	0.8	ClustSim	.9391	.9669	.9459
	0.9	ClustSim	.9081	.9367	.95
	1	ClustSim	.8512	.8787	.9005
		ClustANOVA	.8545	.8778	.8996
3.A.	0.8	ClustSim	.9365	.9418	.9472
	0.9	ClustSim	.9266	.938	.9479
	1	ClustSim	.8555	.8763	.9117
		ClustANOVA	.8562	.8707	.9006
1.B.	0.8	ClusSim	.9198	.9412	.9487
	0.9	ClustSim	.9131	.9354	.9481
	1	ClustSim	.853	.8767	.8962
		ClustANOVA	.8507	.8747	.8985
2.B.	0.8	ClustSim	.9347	.9506	.948
	0.9	ClustSim	.9212	.9323	.9492
	1	ClustSim	.8566	.8698	.8984
		ClustANOVA	.8569	.8719	.8981
3.B.	0.8	ClustSim	.9288	.9478	.949
	0.9	ClustSim	.9211	.933	.9479
	1	ClustSim	.8527	.8688	.894
		ClustANOVA	.8515	.8733	.8991

7.1 Application 1: mathematics related beliefs questionnaire

The aim is to classify the RIs associated with the statements $Q1 - Q17$ about mathematics related beliefs according to the opinion given by 117 students. The representation of the sample means of the answers to each statement (which is also an interval) is gathered Fig. 2. It should be noticed that the spreads of the expected values considered are very similar (around 1 in all cases) which implies that the classification will be given based on the differences in midpoints in this case.

Case $d = 1$:

A p-value lower than the significance level $\alpha = 0.05$ in the p-value matrix $\mathbf{P}_1$ (see Appendix B) implies that the the mean values of the rating of two statements involved in the questionnaire can be considered to be not completely equal. The optimal solution when applying the clustering procedure proposed in this work to the matrix $\mathbf{P}_1$ is composed by seven clusters: Cluster 1 = {Q2,Q14}, Cluster 2 = {Q3,Q10,Q15}, Cluster 3 = {Q4,Q7,Q9,Q11,Q5}, Cluster 4 = {Q1,Q8,Q13}, Cluster 5 = {Q16,Q17}, Cluster 6 = {Q6} and Cluster 7 = {Q12}.
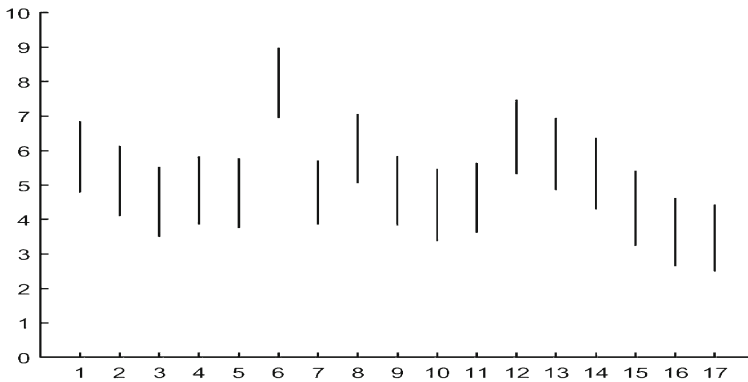
**Fig. 2** Representation of the sample mean of RIs Q1–Q17

Cluster 1 connects the statement related with the taste for mathematical learning and the one related to the perseverance when solving math problems; in both cases the average midpoints are around 5.2 points. Statements in Cluster 2 are those whose average midpoints are around 4.4 points and they talk about personal interest in math and self-abilities in understanding math. Cluster 3 includes those statements with average midpoints around 4.75 points which are related with general math liking, effort capacity and self-concept in math compared to other students. On the other hand, statements in Cluster 4 (average midpoints around 5.9 points) refer to learning interest, patience when solving math and self-expectations in understanding math the current course. Cluster 5 contains two statements regarding competitiveness and insecurity when dealing with mathematics with average midpoints around 3.5 points and, finally, the statement making up the sixth cluster has an average midpoint around 8 points and corresponds to the own general expectations in math marks, whereas the one making up the seventh cluster (whit an average midpoint around 6.4 points) entails self-concept about math achievements the current course.

The procedure ClustANOVA has also been applied leading to the same statements classification than ClustSim for $d = 1$. Besides, if the classical hierarchical algorithm for the mean of the midpoints of the intervals is applied, the dendrogram displayed in Fig. 3 is obtained.

Figure 3 shows that similar clusters are given when an appropriate cut of the dendrogram is considered. However, while an objective stopping criterion leads automatically to six groups when applying the other two procedures, further analyses are required to find a suitable cut for this classical approach. In addition, the classical hierarchical clustering approach does not consider the inherent imprecision in the students' answers, producing a loss of information that can be avoided by using the whole intervals and not only their midpoints while it is true that this loss of information is not well reflected in this example due to the similarity in the spreads of the different expected values.

Case $d = 0.9$:

In this case, a p-value lower than $\alpha = 0.05$ in the p-value matrix $\mathbf{P}_2$ (see Appendix B) implies that the similarity degree between the mean values of the rating of two state-

Classical Hierarchical Clustering
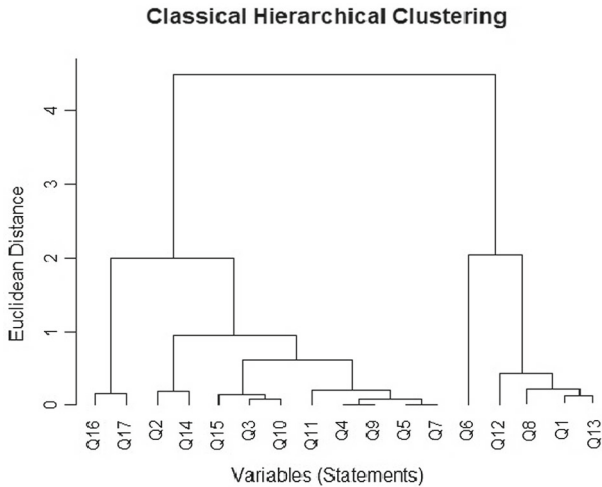
Fig. 3 Dendrogram of the classical hierarchical algorithm for the mean of the midpoints of the intervals corresponding to RIs Q1–Q17

ments involved in the questionnaire can be considered to be lower than 0.9. For instance, the p-value equal to .004 for statements Q1 and Q5 means that such statements are perceived in a very different way by the sample of students.

The method is applied to matrix $\mathbf{P}_2$, and five clusters have been obtained, mainly, Cluster 1={Q2,Q14}, Cluster 2={Q3,Q4,Q5,Q7,Q9,Q10,Q11,Q15}, Cluster 3={Q1,Q8,Q12,Q13}, Cluster 4={Q16,Q17} and Cluster 5={Q6}.

Clusters 1, 4 and 5 coincide with Clusters 1, 5 and 6 in case $d = 1$. On the other hand, Cluster 2 is obtained by linking Clusters 2 and 3 in case $d = 1$, having statements with average midpoints between 4 and 5 points. Thus, it includes statements related with general math liking, personal interest in math, self-abilities in understanding math, effort capacity in math and self-concept in math compared to other students. Statements included in Cluster 3 are the result of joining Clusters 4 and 7 in case $d = 1$ and refers to learning interest, patience when solving math, self-expectations in math understanding and math achievements during the current course. The statements included in new Cluster 3 have average midpoints between 5.5 and 6.5 points.

Case $d = 0.8$:

The p-value matrix is $\mathbf{P}_3$ (see Appendix B) and, in this case, four clusters are obtained when applying the clustering procedure proposed in this work. The first cluster contains all elements of Clusters 1 and 2 for case $d = 0.9$ concerning math liking, math interest, self-concept in math compared to other students, effort capacity and self-expectations in the current course, perseverance and self-abilities in understanding math. All those statements have average midpoints moving betwen 4 and 5.5 points. Finally, Clusters 2, 3 and 4 coincide with clusters 3, 4 and 5 in case $d = 0.9$.
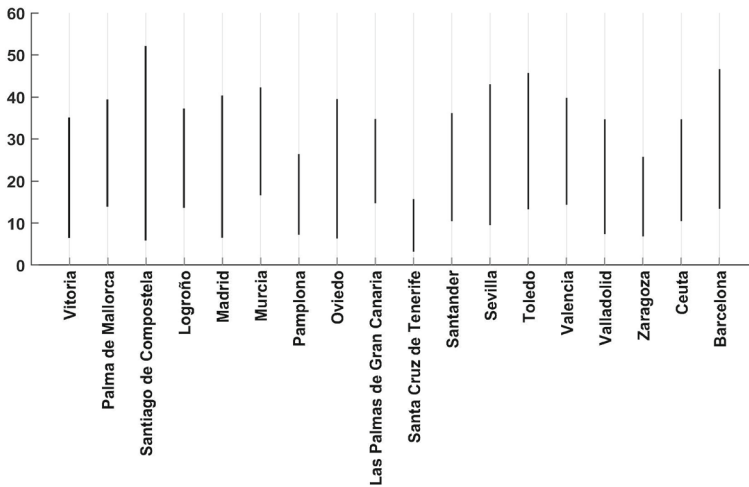
Fig. 4 Representation of the sample mean of the daily concentration of PM10 particles in the 18 Spanish cities

7.2 Application 2: daily concentration of PM10 particles in Spanish cities

Now the aim is to classify the RIs associated with 18 Spanish cities according to their concentration of PM10 particles during 122 days of the central months of 2019. Figure 4 shows the representation of the intervals corresponding to the sample means of the daily concentration of PM10 particles in the 18 cities.

Case $d = 1$

A p-value lower than $\alpha = 0.05$ in the p-value matrix $\mathbf{R}_1$ (see Appendix C) indicates that the mean values of the daily concentration of PM10 particles in the two cities compared are not equal. The solution obtained after applying the proposed procedure leads to nine clusters, which are the following: Cluster 1 = {Vitoria, Valladolid}, Cluster 2 = {Oviedo,Madrid}, Cluster 3 = {Palma de Mallorca, Sevilla, Valencia}, Cluster 4 = {Logroño, Las Palmas de Gran Canaria}, Cluster 5 = {Santiago de Compostela, Barcelona, Toledo.}, Cluster 6 = {Pamplona, Zaragoza}, Cluster 7 = {Santander, Ceuta}, Cluster 8 = {Murcia} and Cluster 9 = {Santa Cruz de Tenerife}.

It can be concluded Santa Cruz de Tenerife is, by far, the city that presents the least pollution by PM10 particles in mean in the central months of 2019 followed by the cities composing Cluster 6, that are Pamplona and Zaragoza. On the other hand, cities in Cluster 5 have the highest maximums, although the average intervals are quite wide because the variability of the daily concentration of PM10 particles during these months has been greater. The rest of the clusters contain cities that have had a moderate variability in terms of the concentration of PM10 particles, and said variability for the cities that make up each cluster can be considered to be the same.

The procedure ClusANOVA has also been applied given the analogies of both procedures, and the optimal solution coincides with the one above. Once again, the results obtained in the study of simulations are corroborated in relation to the similar performance of both methods.

Classical Hierarchical Clustering
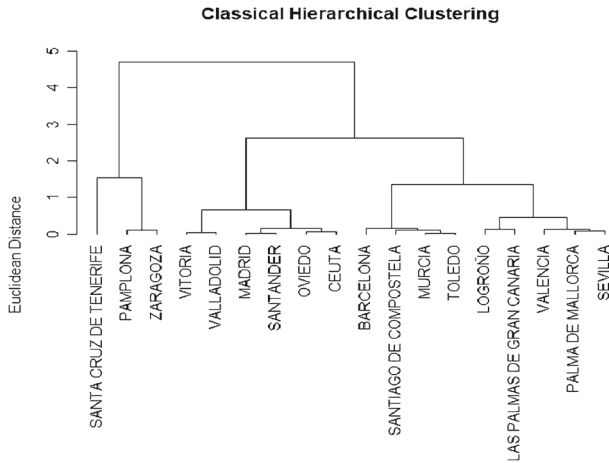
Fig. 5 Dendrogram of the classical hierarchical algorithm for the mean of the midpoints of the intervals corresponding to the daily concentration of PM10 particles in the 18 Spanish cities

On the other hand, the classical hierarchical algorithm for the mean of the midpoints of the intervals has been implemented leading to the classification shown in the dendrogram gathered in Fig. 5.

Figure 5 evidences that most of clusters are given when an appropriate cut of the dendrogram is considered. Nevertheless, in this case it should be noted that the imprecision and variability collected by an interval is lost when the information is summarized by a single value. This is the case, for example, of the city of Murcia, which according to the method proposed in this work does not belong to the same cluster as Santiago de Compostela, Toledo and Barcelona since the variability presented by the mean daily concentration of PM10 particles in the 4 cities cannot be considered to be the same.

Case $d = 0.9$

Now a p-value lower than $\alpha = 0.05$ in the p-value matrix $\mathbf{R}_2$ (see Appendix C) implies that the similarity degree between the mean values of the daily concentration of PM10 particles in the two cities compared can be considered to be lower than 0.9.

In this case, the clustering approach is applied to matrix $\mathbf{R}_2$, and five clusters have been obtained, mainly, Cluster 1 = {Vitoria, Valladolid, Oviedo, Madrid, Santander, Ceuta}, Cluster 2 = {Palma de Mallorca, Logroño, Las Palmas de Gran Canaria, Valencia}, Cluster 3 = {Santiago de Compostela, Murcia, Sevilla, Toledo, Barcelona}, Cluster 4 = {Pamplona, Zaragoza} and Cluster 5 = {Santa Cruz de Tenerife}.

Clusters 4 and 5 are the same that Clusters 6 and 9 in the case $d = 1$, and comprehend the cities with lower minimum pollution and lower variability of such pollution during the analyzed period of 2019. Cluster 1 comes from joining Clusters 1, 2 and 7 in the case $d = 1$, concerning cities that present moderate variabilities in terms of pollution with low minimums. On the other hand, Cluster 2 links Clusters 3 and 4 in case $d = 1$ except for the city of Sevilla which now is added to the previous Cluster 5 to form Cluster 3 (jointly with Murcia) in the new cluster composition. It can be

concluded that Sevilla and Murcia are now linked with the cities that have the highest maximum values in terms of pollution, while the cities that include Cluster 2 have more moderate minimums and maximums, as well as a moderate variability. Therefore, it can be observed that if d reduces from 1, the cluster is obtained such that a degree of overlapping of two intervals are rightly considered even if the midpoints are slightly different.

Case $d = 0.8$**:**

The p-value matrix obtained in this case is $\mathbf{R}_3$ (see Appendix C) and the application of the approach leads to four clusters: Clusters 1, 3 and 4 coincide with Clusters 1, 4 and 5 in case $d = 0.9$, referred to those cities with lower minimum daily concentration of PM10 particles and lower/moderate variability of such concentration. The second cluster is the result of linking Clusters 2 and 3 in case $d = 0.9$, corresponding to cities with higher maximum values and higher variability in terms of pollution.

8 Final remarks

In this contribution a novel approach for clustering RIs has been introduced. In contrast to the method developed in González-Rodríguez et al. (2009) for clustering simple fuzzy random variables (where RIs can be seen as a particular case of fuzzy random variables) based on the equality of expectations, the linkage criteria suggested here considers the similarity degree of expected values of the RIs belonging to the same cluster. It allows us to previously specify a degree of similarity in $(0, 1)$ between expected values that we are willing to admit. Thus, a two-sample test for the degree of similarity of RIs is applied and a p-value matrix is obtained, that indicates somehow how similar are each pair of RIs and serves as a starting point of the clustering algorithm. The simulation studies to observe the performance of the approach along with the real data applications have shown the benefits of the proposed strategy of analysis, also compared to the method in González-Rodríguez et al. (2009) when the ANOVA test for all types RIs introduced in González-Rodríguez et al. (2012) is considered.

As future directions, the partition of RIs based on fuzzy clustering procedures and the similarity measure included here could be analyzed. The inclusion of the variability of the RIs both in the suggested statistic as well as in the ANOVA statistic in González-Rodríguez et al. (2012) can also be studied. The cluster analysis of more complex kind of variables (as, for instance, fuzzy random variables) considering such similarity measure can also be carried out. In addition, a distance between random intervals instead of a similarity index could be also considered for developing different clustering approaches. Finally, the classification of variables whose outputs are more complex objects in the context of simbolic data analysis based on a similarity grade could also be addressed.

Declarations

Conflict of interest The corresponding author states that there is no conflict of interest.

Appendix A Similarity degree matrices for the situations considered in Tables 7, 8 and 9

$$\mathbf{S}_{1.A}^{0.8} = \begin{pmatrix}
1 & .8 & .33 & .288 & .382 & 0 & 0 & 0 & 0 & 0 \\
.8 & 1 & .242 & .203 & .284 & 0 & 0 & 0 & 0 & 0 \\
.33 & .242 & 1 & .9 & .8 & .33 & .309 & .285 & .263 & .242 \\
.288 & .203 & .9 & 1 & .9 & .382 & .355 & .33 & .307 & .284 \\
.382 & .284 & .8 & .9 & 1 & .288 & .265 & .243 & .223 & .203 \\
0 & 0 & .33 & .382 & .288 & 1 & .95 & .9 & .85 & .8 \\
0 & 0 & .309 & .355 & .265 & .95 & 1 & .95 & .9 & .85 \\
0 & 0 & .285 & .33 & .243 & .9 & .95 & 1 & .95 & .9 \\
0 & 0 & .263 & .307 & .223 & .85 & .9 & .95 & 1 & .95 \\
0 & 0 & .242 & .284 & .203 & .8 & .85 & .9 & .95 & 1
\end{pmatrix}$$

$$\mathbf{S}_{2.A}^{0.8} = \begin{pmatrix}
1 & .8 & .33 & .31 & .286 & 0 & 0 & 0 & 0 & 0 \\
.8 & 1 & .2 & .172 & .143 & 0 & 0 & 0 & 0 & 0 \\
.33 & .2 & 1 & .9 & .8 & .33 & .3 & .267 & .233 & .2 \\
.31 & .172 & .9 & 1 & .9 & .31 & .355 & .33 & .307 & .284 \\
.286 & .143 & .8 & .9 & 1 & .288 & .276 & .241 & .207 & .172 \\
0 & 0 & .33 & .31 & .288 & 1 & .95 & .9 & .85 & .8 \\
0 & 0 & .3 & .255 & .276 & .95 & 1 & .95 & .9 & .85 \\
0 & 0 & .267 & .33 & .241 & .9 & .95 & 1 & .95 & .9 \\
0 & 0 & .233 & .307 & .207 & .85 & .9 & .95 & 1 & .95 \\
0 & 0 & .2 & .284 & .172 & .8 & .85 & .9 & .95 & 1
\end{pmatrix}$$

$$\mathbf{S}_{3.A}^{0.8} = \begin{pmatrix}
1 & .8 & .4 & .36 & .32 & .1 & .095 & .09 & .085 & .08 \\
.8 & 1 & .5 & .45 & .4 & .125 & .119 & .113 & .106 & .1 \\
.4 & .5 & 1 & .9 & .8 & .25 & .237 & .225 & .212 & .2 \\
.36 & .45 & .9 & 1 & .89 & .278 & .264 & .25 & .236 & .222 \\
.32 & .4 & .8 & .89 & 1 & .312 & .297 & .281 & .266 & .25 \\
.1 & .125 & .25 & .278 & .312 & 1 & .95 & .9 & .85 & .8 \\
.095 & .119 & .237 & .264 & .297 & .95 & 1 & .947 & .895 & .842 \\
.09 & .113 & .225 & .25 & .281 & .9 & .947 & 1 & .944 & .889 \\
.085 & .106 & .212 & .236 & .266 & .85 & .895 & .944 & 1 & .941 \\
.08 & .1 & .2 & .222 & .25 & .8 & .842 & .889 & .941 & 1
\end{pmatrix}$$

$$\mathbf{S}_{1.B}^{0.8} = \mathbf{S}_{2.B}^{0.8} = \begin{pmatrix}
1 & .8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.8 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & .9 & .8 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .9 & 1 & .9 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .8 & .9 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & .95 & .9 & .85 & .8 \\
0 & 0 & 0 & 0 & 0 & .95 & 1 & .95 & .9 & .85 \\
0 & 0 & 0 & 0 & 0 & .9 & .95 & 1 & .95 & .9 \\
0 & 0 & 0 & 0 & 0 & .85 & .9 & .95 & 1 & .95 \\
0 & 0 & 0 & 0 & 0 & .8 & .85 & .9 & .95 & 1
\end{pmatrix}$$

$$\mathbf{S}_{3.B}^{0.8} = \begin{pmatrix}
1 & .8 & .571 & .514 & .457 & .071 & .068 & .064 & .061 & .057 \\
.8 & 1 & .714 & .643 & .571 & .089 & .085 & .08 & .076 & .071 \\
.571 & .714 & 1 & .9 & .8 & .125 & .119 & .113 & .106 & .1 \\
.514 & .643 & .9 & 1 & .889 & .139 & .132 & .125 & .118 & .111 \\
.457 & .571 & .8 & .889 & 1 & .156 & .148 & .141 & .133 & .125 \\
.071 & & .089 & .125 & .139 & .156 & 1 & .95.9 & .85 & .8 \\
.068 & .085 & .119 & .132 & .148 & .95 & 1 & .947 & .895 & .842 \\
.064 & .08 & .113 & .125 & .141 & .9 & .947 & 1 & .944 & .889 \\
.061 & & .076 & .106 & .118 & .133 & .85 & .895.944 & 1 & .941 \\
.057 & & .071 & .1 & .111 & .125 & .8 & .842.889 & .941 & 1
\end{pmatrix}$$

$$\mathbf{S}_{1.A}^{0.9} = \begin{pmatrix}
1 & .9 & .33 & .357 & .311 & 0 & 0 & 0 & 0 & 0 \\
.9 & 1 & .288 & .31 & .267 & 0 & 0 & 0 & 0 & 0 \\
.33 & .288 & 1 & .95 & .9 & .33 & .322 & .311 & .299 & .288 \\
.357 & .31 & .95 & 1 & .95 & .311 & .3 & .289 & .277 & .267 \\
.311 & .267 & .9 & .95 & 1 & .357 & .345 & .333 & .321 & .31 \\
0 & 0 & .33 & .311 & .357 & 1 & .975 & .95 & .925 & .9 \\
0 & 0 & .322 & .3 & .345 & .975 & 1 & .975 & .95 & .85 \\
0 & 0 & .311 & .389 & .333 & .95 & .975 & 1 & .975 & .95 \\
0 & 0 & .299 & .277 & .321 & .925 & .95 & .975 & 1 & .975 \\
0 & 0 & .288 & .267 & .31 & .9 & .925 & .95 & .975 & 1
\end{pmatrix}$$

$$\mathbf{S}_{2.A}^{0.9} = \begin{pmatrix}
1 & .9 & .33 & .322 & .31 & 0 & 0 & 0 & 0 & 0 \\
.9 & 1 & .267 & .254 & .241 & 0 & 0 & 0 & 0 & 0 \\
.33 & .267 & 1 & .95 & 9 & .33 & .317 & .3 & .283 & .267 \\
.322 & .254 & .95 & 1 & .95 & .322 & .305 & .288 & .271 & .254 \\
.31 & .241 & .9 & .95 & 1 & .31 & .293 & .276 & .259 & .241 \\
0 & 0 & .33 & .322 & .31 & 1 & .975 & .95 & .925 & .9 \\
0 & 0 & .317 & .305 & .293 & .975 & 1 & .975 & .95 & .925 \\
0 & 0 & .3 & .288 & .276 & .95 & .975 & 1 & .975 & .9.5 \\
0 & 0 & .283 & .271 & .259 & .925 & .95 & .975 & 1 & .975 \\
0 & 0 & .267 & .254 & .241 & .9 & .925 & .95 & .975 & 1
\end{pmatrix}$$

$$\mathbf{S}_{3.A}^{0.9} = \begin{pmatrix}
1 & .9 & .5 & .47 & .45 & .2 & .195 & .19 & .185 & .18 \\
.9 & 1 & .556 & .528 & .5 & .222 & .217 & .211 & .206 & .2 \\
.5 & .556 & 1 & .95 & .9 & .4 & .39 & .38 & .37 & .36 \\
.475 & .528 & .95 & 1 & .9474 & .421 & .411 & .4 & .39 & .379 \\
.45 & .5 & .9 & .9474 & 1 & .444 & .433 & .422 & .411 & .4 \\
.2 & .222 & .4 & .421 & .444 & 1 & .975 & .95 & .925 & .9 \\
.195 & .217 & .39 & .411 & .433 & .975 & 1 & .974 & .949 & .923 \\
.19 & .211 & .38 & .4 & .422 & .95 & .974 & 1 & .974 & .947 \\
.185 & .206 & .37 & .39 & .411 & .925 & .949 & .974 & 1 & .973 \\
.18 & .2 & .36 & .379 & .4 & .9 & .923 & .947 & .973 & 1
\end{pmatrix}$$

$$\mathbf{S}_{1.B}^{0.9} = \mathbf{S}_{2.B}^{0.9} = \begin{pmatrix}
1 & .9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & .95 & .9 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .95 & 1 & .95 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .9 & .95 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & .975 & .95 & .925 & .9 \\
0 & 0 & 0 & 0 & 0 & .975 & 1 & .975 & .95 & .925 \\
0 & 0 & 0 & 0 & 0 & .95 & .975 & 1 & .975 & .95 \\
0 & 0 & 0 & 0 & 0 & .925 & .95 & .975 & 1 & .975 \\
0 & 0 & 0 & 0 & 0 & .9 & .925 & .95 & .975 & 1
\end{pmatrix}$$

$$\mathbf{S}_{3.B}^{0.9} = \begin{pmatrix}
1 & .9 & .5 & .475 & .45 & .1 & .098 & .095 & .093 & .09 \\
.9 & 1 & .556 & .528 & .5 & .111 & .108 & .106 & .103 & .1 \\
.5 & .556 & 1 & .95 & .9 & .2 & .195 & .19 & .185 & .18 \\
.475 & .528 & .95 & 1 & .947 & .211 & .205 & .2 & .195 & .189 \\
.45 & .5 & .9 & .947 & 1 & .222 & .217 & .211 & .206 & .2 \\
.1 & .111 & .2 & .211 & .222 & 1 & .975 & .95 & .925 & .9 \\
.098 & .108 & .195 & .205 & .217 & .975 & 1 & .974 & .949 & .923 \\
.095 & .106 & .19 & .2 & .211 & .95 & .974 & 1 & .974 & .947 \\
.093 & .103 & .185 & .195 & .206 & .925 & .949 & .974 & 1 & .973 \\
.09 & .1 & .18 & .189 & .2 & .9 & .923 & .947 & .973 & 1
\end{pmatrix}$$

$$\mathbf{S}^1_{1.A} = \mathbf{S}^1_{2.A} = \mathbf{S}^1_{3.A} = \mathbf{S}^1_{1.B} = \mathbf{S}^1_{2.B} = \mathbf{S}^1_{3.B} = \begin{pmatrix} 1&1&1&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0 \\ 0&0&1&1&1&0&0&0&0&0 \\ 0&0&1&1&1&0&0&0&0&0 \\ 0&0&1&1&1&0&0&0&0&0 \\ 0&0&0&0&0&1&1&1&1&1 \\ 0&0&0&0&0&1&1&1&1&1 \\ 0&0&0&0&0&1&1&1&1&1 \\ 0&0&0&0&0&1&1&1&1&1 \\ 0&0&0&0&0&1&1&1&1&1 \end{pmatrix}$$

Appendix B p-value matrices generated in the real-life Application 1

$$\mathbf{P}_1 = \begin{pmatrix}
1 & .024 & 0 & .001 & .005 & 0 & .005 & .418 & .001 & 0 & 0 & .033 & .802 & .09 & 0 & 0 & 0 \\
.024 & 0 & .102 & .431 & .329 & 0 & .339 & .002 & .399 & .055 & .155 & 0 & .020 & .522 & .0191 & 0 & 0 \\
0 & .102 & 1 & .372 & .479 & 0 & .433 & 0 & .314 & .821 & .751 & 0 & 0 & .011 & .579 & .01 & .001 \\
.001 & .431 & .372 & 1 & .833 & 0 & .877 & 0 & .995 & .259 & .578 & 0 & .004 & .119 & .131 & 0 & 0 \\
.005 & .329 & .479 & .833 & 1 & 0 & .942 & 0 & .879 & .33 & .687 & 0 & .001 & .08 & .193 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.005 & .339 & .433 & .877 & .942 & 0 & 1 & 0 & .912 & .323 & .664 & 0 & .001 & .099 & .169 & .001 & .001 \\
.418 & .002 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & .276 & .609 & .01 & 0 & 0 & 0 \\
.001 & .399 & .313 & .994 & .879 & 0 & .912 & 0 & 1 & .274 & .545 & 0 & .002 & .118 & .13 & .001 & 0 \\
0 & .055 & .821 & .259 & .33 & 0 & .323 & 0 & .274 & 1 & .577 & 0 & 0 & .01 & .794 & .02 & .002 \\
0 & .155 & .751 & .578 & .687 & 0 & .664 & 0 & .545 & .576 & 1 & 0 & 0 & .0318 & .365 & .002 & 0 \\
.033 & 0 & 0 & 0 & 0 & 0 & 0 & .276 & 0 & 0 & 0 & 1 & .098 & 0 & 0 & 0 & 0 \\
.802 & .02 & 0 & .004 & .001 & 0 & .001 & .609 & .002 & 0 & 0 & .098 & 1 & .063 & 0 & 0 & 0 \\
.09 & .522 & .011 & .119 & .08 & 0 & .099 & .011 & .118 & .01 & .032 & 0 & .063 & 1 & 0 & 0 & 0 \\
0 & .002 & .579 & .131 & .193 & 0 & .169 & 0 & .129 & .794 & .345 & 0 & 0 & 0 & 1 & .029 & .008 \\
0 & 0 & .01 & 0 & 0 & 0 & .001 & 0 & .001 & .02 & .002 & 0 & 0 & 0 & .029 & 1 & .607 \\
0 & 0 & .001 & 0 & 0 & 0 & .001 & 0 & 0 & .002 & 0 & 0 & 0 & 0 & .008 & .607 & 0
\end{pmatrix}$$

$$\mathbf{P}_2 = \begin{pmatrix}
1 & .067 & 0 & .009 & .004 & 0 & .004 & .644 & .005 & 0 & 0 & .094 & 1 & .202 & 0 & 0 & 0 \\
.067 & 1 & .152 & .624 & .503 & 0 & .488 & .001 & .609 & .106 & .277 & 0 & .041 & .747 & .035 & 0 & 0 \\
0 & .152 & 1 & .531 & .644 & 0 & .631 & 0 & .541 & 1 & .99 & 0 & 0 & .032 & .859 & .019 & .003 \\
.009 & .624 & .531 & 1 & 0 & 1 & .003 & 1 & .379 & .771 & 0 & .235 & .228 & 0 & 0 \\
.004 & .503 & .644 & 1 & 1 & 0 & 1 & 0 & 1 & .511 & .943 & 0 & .004 & .164 & .32 & .004 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.004 & .488 & .631 & 1 & 1 & 0 & 1 & 0 & 1 & .48 & .907 & 0 & .004 & .166 & .284 & .004 & 0 \\
.644 & .014 & 0 & .003 & 0 & 0 & 0 & 1 & .001 & 0 & 0 & .43 & .896 & .044 & 0 & 0 & 0 \\
.005 & .609 & .541 & 1 & 1 & 0 & 1 & 0 & 1 & .392 & .799 & 0 & .002 & .214 & .229 & .002 & 0 \\
0 & .106 & 1 & .379 & .511 & 0 & .48 & 0 & .392 & 1 & .782 & 0 & 0 & .023 & 1 & .039 & .021 \\
0 & .277 & .99 & .771 & .943 & 0 & .907 & 0 & .799 & .782 & 1 & 0 & .002 & .076 & .592 & .011 & .002 \\
.094 & 0 & 0 & 0 & 0 & 0 & 0 & .43 & 0 & 0 & 0 & 1 & .176 & .001 & 0 & 0 & 0 \\
1 & .041 & 0 & .006 & .004 & 0 & .004 & .896 & .002 & 0 & .002 & .176 & 1 & .12 & 0 & 0 & 0 \\
.202 & .747 & .032 & .235 & .164 & 0 & .166 & .044 & .214 & .023 & .076 & .001 & .12 & 1 & .009 & 0 & 0 \\
0 & .035 & .859 & .228 & .32 & 0 & .284 & 0 & .229 & 1 & .592 & 0 & 0 & .009 & 1 & .065 & .026 \\
0 & 0 & .019 & 0 & .004 & 0 & .004 & 0 & .002 & .039 & .011 & 0 & 0 & 0 & .065 & 1 & .863 \\
0 & 0 & .003 & 0 & 0 & 0 & 0 & 0 & 0 & .021 & .002 & 0 & 0 & 0 & .026 & .863 & 1
\end{pmatrix}$$

$$
P_3 = \begin{pmatrix}
1 & .14 & .001 & .018 & .01 & 0 & .009 & .984 & .011 & .001 & .008 & .192 & 1 & .374 & 0 & 0 & 0 \\
.140 & 1 & .265 & .89 & .714 & 0 & .73 & .032 & .868 & .221 & .437 & .001 & .09 & 1 & .076 & 0 & 0 \\
.001 & .265 & 1 & .782 & .959 & 0 & .905 & .001 & .767 & 1 & 1 & 0 & .077 & 1 & .051 & .011 & \\
.018 & .89 & .782 & 1 & 1 & 0 & 1 & .003 & 1 & .599 & 1 & 0 & .024 & .434 & .405 & .002 & 0 \\
.01 & .714 & .959 & 1 & 1 & 0 & 1 & 0 & 1 & .746 & 1 & 0 & .007 & .31 & .551 & .004 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.009 & .73 & .905 & 1 & 1 & 0 & 1 & .001 & 1 & .699 & 1 & 0 & .004 & .312 & .506 & .008 & .001 \\
.984 & .032 & .001 & .003 & 0 & 0 & .001 & 1 & .003 & 0 & 0 & .75 & 1 & .106 & 0 & 0 & 0 \\
.011 & .868 & .767 & 1 & 1 & 0 & 1 & .003 & 1 & .605 & 1 & 0 & .01 & .382 & .417 & .002 & 0 \\
.001 & .221 & 1 & .599 & .746 & 0 & .699 & 0 & .605 & 1 & 1 & 0 & 0 & .037 & 1 & .1 & .035 \\
.008 & .437 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & .001 & .15 & .861 & .023 & .004 \\
.192 & .001 & 0 & 0 & 0 & 0 & 0 & .75 & 0 & 0 & 0 & 1 & .392 & .002 & 0 & 0 & 0 \\
1.001 & .09 & .003 & .024 & .007 & 0 & .004 & 1 & .01 & 0 & .001 & .392 & 1 & .275 & 0 & 0 & 0 \\
.374 & 1 & .077 & .434 & .31 & 0 & .312 & .106 & .382 & .037 & .15 & .002 & .275 & 1 & .006 & 0 & 0 \\
0 & .076 & 1 & .405 & .551 & 0 & .506 & 0 & .417 & 1 & .861 & 0 & 0 & .006 & 1 & .132 & .049 \\
0 & 0 & .051 & .002 & .004 & 0 & .008 & 0 & .002 & .1 & .023 & 0 & 0 & 0 & .132 & 1 & 1 \\
0 & 0 & .011 & 0 & 0 & 0 & .001 & 0 & 0 & .035 & .004 & 0 & 0 & 0 & .049 & 1 & 1
\end{pmatrix}
$$

Appendix C p-value matrices generated in the real-life Application 2

$$
R_1 = \begin{pmatrix}
1 & 0 & .006 & 0 & .18 & 0 & .004 & .128 & .007 & 0 & .077 & 0 & 0 & 0 & .873 & .007 & .187 & 0 \\
0 & 1 & .051 & .319 & .042 & .02 & 0 & 0 & .112 & 0 & .006 & .106 & .059 & .78 & 0 & 0 & 0 & .048 \\
.006 & .051 & 1 & .026 & .066 & .073 & 0 & .02 & .011 & 0 & .017 & .255 & .22 & .057 & .004 & 0 & .02 & .279 \\
0 & .319 & .026 & 1 & .128 & .002 & 0 & .035 & .426 & 0 & .061 & .061 & .022 & .187 & 0 & 0 & .007 & .004 \\
.18 & .042 & .066 & .128 & 1 & 0 & 0 & .76 & .02 & 0 & .202 & .121 & 0 & .039 & .189 & 0 & .132 & .002 \\
0 & .02 & .073 & .002 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & .031 & .288 & .081 & 0 & 0 & 0 & .277 \\
.004 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .692 & 0 & 0 \\
.128 & 0 & .02 & .035 & .76 & 0 & 0 & 1 & .007 & 0 & .084 & .011 & 0 & .004 & .147 & 0 & .039 & 0 \\
.007 & .112 & .011 & .426 & .02 & 0 & 0 & .007 & 1 & 0 & .209 & .017 & .009 & .042 & .009 & 0 & .046 & .006 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.077 & .006 & .017 & .061 & .202 & 0 & 0 & .084 & .209 & 0 & 1 & .011 & 0 & .007 & .062 & 0 & .475 & 0 \\
0 & .106 & .255 & .061 & .121 & .031 & 0 & .011 & .017 & 0 & .011 & 1 & .061 & .132 & 0 & 0 & 0 & .044 \\
0 & .059 & .22 & .022 & 0 & .288 & 0 & 0 & .009 & 0 & 0 & .061 & 1 & .172 & 0 & 0 & 0 & .815 \\
0 & .78 & .057 & .187 & .039 & .081 & 0 & .004 & .042 & 0 & .007 & .132 & .172 & 1 & 0 & 0 & .002 & .101 \\
.873 & 0 & .004 & 0 & .189 & 0 & 0 & .147 & .009 & 0 & .062 & 0 & 0 & 0 & 1 & 0 & .198 & 0 \\
.007 & 0 & 0 & 0 & 0 & 0 & .692 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
.187 & 0 & .02 & .007 & .132 & 0 & 0 & .039 & .046 & 0 & .475 & 0 & 0 & .002 & .198 & 0 & 1 & 0 \\
0 & .048 & .279 & .004 & .002 & .277 & 0 & 0 & .006 & 0 & 0 & .044 & .815 & .101 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

$$
R_2 = \begin{pmatrix}
1 & .002 & .017 & .033 & .623 & 0 & .061 & .788 & .07 & 0 & .455 & .018 & 0 & .004 & 1 & .048 & .818 & 0 \\
.002 & 1 & .126 & 1 & .305 & .243 & 0 & .049 & .516 & 0 & .071 & .362 & .392 & 1 & .002 & 0 & .013 & .263 \\
.017 & .126 & 1 & .08 & .247 & .133 & 0 & .157 & .048 & 0 & .124 & .463 & .439 & .157 & .027 & .001 & .077 & .459 \\
.033 & 1 & .08 & 1 & .286 & .026 & 0 & .164 & .817 & 0 & .476 & .218 & .097 & .833 & .022 & 0 & .162 & .057 \\
.623 & .305 & .247 & .286 & 1 & .008 & .002 & 1 & .109 & 0 & .447 & .543 & .033 & .21 & .645 & .002 & .301 & .02 \\
0 & .243 & .133 & .026 & .008 & 1 & 0 & 0 & .002 & 0 & 0 & .228 & .57 & .447 & 0 & 0 & 0 & .512 \\
.061 & 0 & 0 & 0 & .002 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .017 & 1 & 0 & 0 \\
.788 & .049 & .157 & .164 & 1 & 0 & 0 & 1 & .023 & 0 & .336 & .227 & .002 & .038 & .639 & 0 & .174 & .002 \\
.07 & .516 & .048 & .817 & .109 & .002 & 0 & .023 & 1 & 0 & .476 & .043 & .04 & .384 & .039 & 0 & .395 & .015 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.455 & .071 & .124 & .476 & .447 & 0 & 0 & .336 & .476 & 0 & 1 & .24 & .003 & .05 & .467 & 0 & 1 & .002 \\
.018 & .362 & .463 & .218 & .543 & .228 & 0 & .227 & .043 & 0 & .24 & 1 & .364 & .382 & .008 & 0 & .045 & .263 \\
0 & .392 & .439 & .097 & .033 & .57 & 0 & .002 & .04 & 0 & .003 & .364 & 1 & .549 & 0 & 0 & .002 & 1 \\
.004 & 1 & .157 & .833 & .21 & .447 & 0 & .038 & .384 & 0 & .05 & .382 & .549 & 1 & 0 & 0 & .01 & .43 \\
1 & .002 & .027 & .022 & .645 & 0 & .017 & .639 & .039 & 0 & .467 & .008 & 0 & 0 & 1 & .017 & .95 & 0 \\
.048 & 0 & .001 & 0 & .002 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .017 & 1 & 0 & 0 \\
.818 & .013 & .077 & .162 & .301 & 0 & 0 & .174 & .395 & 0 & 1 & .045 & .002 & .01 & .95 & 0 & 1 & 0 \\
0 & .263 & .459 & .057 & .02 & .512 & 0 & .002 & .015 & 0 & .002 & .263 & 1 & .43 & 0 & 0 & 0 & 1
\end{pmatrix}
$$

$$\mathbf{R}_3 = \begin{pmatrix}
1 & .047 & .161 & .247 & 1 & .001 & .366 & 1 & .399 & 0 & 1 & .207 & .001 & .037 & 1 & .27 & 1 & .003 \\
.047 & 1 & .281 & 1 & .843 & 1 & 0 & .739 & .978 & 0 & .673 & .88 & .98 & 1 & .037 & 0 & .233 & .906 \\
.161 & .281 & 1 & .202 & .744 & .296 & .002 & .589 & .08 & 0 & .278 & .757 & .678 & .263 & .156 & .002 & .219 & .734 \\
.247 & 1 & .202 & 1 & .59 & .304 & 0 & .487 & 1 & 0 & 1 & .541 & .571 & 1 & .218 & 0 & .944 & .381 \\
1 & .843 & .744 & .59 & 1 & .106 & .02 & 1 & .219 & 0 & .852 & 1 & .214 & .851 & .997 & .023 & .636 & .153 \\
.001 & 1 & .296 & .304 & .106 & 1 & 0 & .01 & .07 & 0 & .006 & .892 & .986 & 1 & 0 & 0 & 0 & .933 \\
.366 & 0 & .002 & 0 & .02 & 0 & 1 & .005 & 0 & 0 & 0 & 0 & 0 & 0 & .202 & 1 & 0 & 0 \\
1 & .739 & .589 & .487 & 1 & .01 & .005 & 1 & .097 & 0 & .911 & 1 & .061 & .467 & 1 & .006 & .602 & .033 \\
.399 & .978 & .08 & 1 & .219 & .07 & 0 & .097 & 1 & 0 & .948 & .17 & .245 & .984 & .401 & 0 & 1 & .158 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & .673 & .278 & 1 & .852 & .006 & 0 & .911 & .948 & 0 & 1 & .892 & .053 & .443 & 1 & .001 & 1 & .036 \\
.207 & .88 & .757 & .541 & 1 & .892 & 0 & 1 & .17 & 0 & .892 & 1 & 1 & .868 & .203 & 0 & .58 & .998 \\
.001 & .98 & .678 & .571 & .214 & .986 & 0 & .061 & .245 & 0 & .053 & 1 & 1 & .98 & .004 & 0 & .019 & 1 \\
.037 & 1 & .263 & 1 & .851 & 1 & 0 & .467 & .984 & 0 & .443 & .868 & .98 & 1 & .02 & 0 & .141 & .911 \\
1 & .037 & .156 & .218 & .997 & 0 & .202 & 1 & .401 & 0 & 1 & .203 & .004 & .02 & 1 & .156 & 1 & .001 \\
.27 & 0 & .002 & 0 & .023 & 0 & 1 & .006 & 0 & 0 & .001 & 0 & 0 & 0 & .156 & 1 & .002 & 0 \\
1 & .233 & .219 & .944 & .636 & 0 & 0 & .602 & 1 & 0 & 1 & .58 & .019 & .141 & 1 & .002 & 1 & .006 \\
.003 & .906 & .734 & .381 & .153 & .933 & 0 & .033 & .158 & 0 & .036 & .998 & 1 & .911 & .001 & 0 & .006 & 1
\end{pmatrix}$$

References

Aumann RJ (1965) Integrals of set-valued functions. J Math Anal Appl 12:1–12

Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98:470–487

Blanco-Fernández A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput Stat Data Ann 55(9):2568–2578

Blanco-Fernández A, Colubi A, García-Bárzana M (2013) A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables. Inf Sci 247:109–122

Cappelli C, D'Urso P, Di Iorio F (2013) Change point analysis for imprecise time series. Fuzzy Sets Syst 225:23–38

Chavent M (1998) A monothetic clustering method. Pattern Recogn Lett 19(11):989–996

Chavent M, Lechevallier Y (2002) Dynamical clustering algorithm of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In: Sokolowski A, Bock HH (eds) Classification, Clustering and Data Analysis. Springer-Verlag, Berlin, Germany, pp 53–59

Chavent M, Kuentz-Simonet V, Liquet B, Saracco J (2012) ClustOfVar: an R package for the clustering of variables. J Stat Softw 50(13):1–16

D'Urso P, Giordani P (2004) A least squares approach to principal component analysis for interval valued data. Chemometr Intell Lab Syst 70(2):179–192

D'Urso P, De Giovanni L, Massari R (2015) Trimmed fuzzy clustering for interval-valued data. Adv Data Anal Class 9(1):21–40

De Carvalho FdAT, De Souza RM, Chavent M, Lechevallier Y (2006) Adaptive hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recogn Lett 27(3):167–179

De Carvalho FdAT, Brito P, Bock HH (2006) Dynamic clustering for interval data based on L2 distance. Comput Stat 21(2):231–250

El-Shaarawi AH, Piegorsch WW (2002) Encyclopedia of environmetrics. Wiley, New York

González-Rodríguez G, Montenegro M, Colubi A, Gil MA (2006) Bootstrap techniques and fuzzy random variables: synergy in hypothesis testing with fuzzy data. Fuzzy Sets Syst 157:2608–2613

González-Rodríguez G, Colubi A, D'Urso P, Montenegro M (2009) Multi-sample test-based clustering for fuzzy random variables. Int J Approx Reason 50:721–731

González-Rodríguez G, Colubi A, Gil MA (2012) Fuzzy data treated as functional data: a one-way ANOVA test approach. Comput Stat Data Anal 56:943–955

Gowda KC, Ravi T (1995) Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity. Pattern Recogn Lett 16(6):647–652

Hajjar C, Hamdan H (2013) Interval data clustering using self-organizing maps based on adaptive Mahalanobis distances. Neural Networks 46:124–132

Horowitz JL, Manski CF (2006) Identification and estimation of statistical functionals using incomplete data. J Econ 132:445–459

Hudgens MG (2005) On nonparametric maximum likelihood estimation with interval censoring and left truncation. J R Stat Soc Ser B 67:573–587

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37:547–579

Kojadinovic I (2004) Agglomerative hierarchical clustering of continuous variables based on mutual information. Comput Stat Data Anal 46:269–294

Kreinovich V, Nguyen HT, Wu B (2007) On-line algorithms for computing mean and variance of interval data, and their use in intelligent systems. Inf Sci 177(16):3228–3238

Magnac T, Maurin E (2008) Partial identification in monotone binary models: discrete regressors and interval data. Rev Econ Stud 75:835–864

Maharaj EA (1996) A significance test for classifying ARMA models. J Stat Comput Simul 54:305–331

Matheron G (1975) Random sets and integral geometry. Wiley, New York

Nicolau F, Bacelar-Nicolau H (1998) Some trends in the classi.cation of variables. In: Hayashi C, Ohsumi N, Yajima K, Tanaka Y, Bock H, Baba Y (eds) Data science, classiffication and related methods. Springer, Berlin, pp 89–98

Park S, Lim J, Choi H, Kwak M (2020) Clustering of longitudinal interval-valued data via mixture distribution under covariance separability. J Appl Stat 47(10):1739–1756

Ralambondrainy H (1995) A conceptual version of the K-means algorithm. Pattern Recogn Lett 16(11):1147–1157

Ramos-Guajardo AB (2015) Similarity Test for the Expectation of a Random Interval and a Fixed Interval. In: Grzegorzewski P, Gagolewski M, Hryniewicz O, Gil MA (eds) Strengthening links between data analysis and soft computing. Advances in Intelligent Systems and Computing vol 315, pp 175–182

Ramos-Guajardo AB, Grzegorzewski P (2016) Distance-based linear discriminant analysis for interval-valued data. Inf Sci 372:591–607

Ramos-Guajardo AB, Blanco-Fernández A (2017) Two-sample similarity test for the expected value of random intervals. In: Ferraro MB, Giordani P, Vantaggi B, Gagolewski M, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) Soft methods for data science. Advances in Intelligent Systems and Computing vol 456, pp 423–430

Ramos-Guajardo AB, González-Rodríguez G, Colubi A (2020) Testing the degree of overlap for the expected value of random intervals. Int J Approx Reason 119:1–19

Rodriguez SIR, De Carvalho FdAT (2019) Clustering interval-valued data with automatic variables weighting. In: Proceedings of the international joint conference on neural networks pp 1–8

Shao J, Tu D (1995) The Jackknife and Bootstrap. Springer-Verlag, New York

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

Sinova B, Colubi A, Gil MA, González-Rodríguez G (2012) Interval arithmetic-based linear regression between interval data: discussion and sensitivity analysis on the choice of the metric. Inf Sci 199:109–124 sensitivity analysis on the choice of the metric. Inf Sci 199:109–124

Xu X, Lu J, Wang W (2018) Hierarchical clustering of complex symbolic data and application for emitter identification. J Comput Sci Technol 33(4):807–822