International Journal of Approximate Reasoning 50 (2009) 721–731

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Multi-sample test-based clustering for fuzzy random variables

Gil González-Rodríguez [a], Ana Colubi [b,*], Pierpaolo D'Urso [c], Manuel Montenegro [b]

[a] *Research Unit on Intelligent Data Analysis and Graphical Models, European Centre for Soft Computing, 33600 Mieres, Spain*
[b] *Dpto. de Estadística e I.O. y D.M., Universidad de Oviedo, C/Calvo sotelo s/n, 33007 Oviedo, Spain*
[c] *Dipartimento di Teoria economica e metodi quantitativi per le scelte politiche, Sapienza Università di Roma, P.za Aldo Moro, 5-00185 Rome, Italy*

ARTICLE INFO

Article history:
Received 24 July 2008
Received in revised form 21 January 2009
Accepted 26 January 2009
Available online 4 February 2009

Keywords:
Clustering
Fuzzy random variable
Multi-sample test
Bootstrap hypothesis testing

ABSTRACT

A clustering method to group independent fuzzy random variables observed on a sample by focusing on their expected values is developed. The procedure is iterative and based on the *p*-value of a multi-sample bootstrap test. Thus, it simultaneously takes into account fuzziness and stochastic variability. Moreover, an objective stopping criterion leading to statistically equal groups different from each other is provided. Some simulations to show the performance of this inferential approach are included. The results are illustrated by means of a case study.

1. Introduction

Experimental data are often affected by several sources of uncertainty, namely, randomness, vagueness, imprecision, ambiguity, etc. Probability and statistics have proved to be sound theories to handle with uncertainty due to randomness, while fuzzy sets are increasingly used to deal with imprecision or vagueness of the data (see, for instance, [11,24,28]).

Fuzzy random variables (FRV) arose to model experiments in which both randomness and imprecision are present (see, for instance, [4,21,31]). They have been used in various areas like forestry, hydrology and economics (see, for instance, [1,5,10]).

The theory of FRVs in Puri and Ralescu's sense [31] is fully meaningful when the results of the random experiments are 'purely fuzzy', and the aim is to obtain conclusions regarding the fuzzy data. Nevertheless, there are other experiments in which the aim is focused on a real-valued random variable, although for any reason this variable cannot be precisely observed, and only a fuzzy perception is available. A common approach to handle these kinds of experiments is based on the extension principle [32]. Namely, the statistical procedures are developed from the corresponding ones for the underlying real-valued random variable, and the uncertainty due to the ill-observation is propagated through the extension principle (see, for instance [13]).

Statistical inference with fuzzy components has been tackled from different perspectives in the literature (see, for instance, [2,13,16,17,25–27,30]). In this paper, precise statistical models for fuzzy data (and not for any real-valued underlying variable) will be considered. Then Puri and Ralescu's concept of FRV [31] is employed (i.e., a FRV is identified with a random element whose values are fuzzy sets).

The aim here is to develop clustering methods for grouping features when the available data is a $k \times n$ matrix of fuzzy sets obtained from the observation of k independent FRVs on n independent statistical units.

* Corresponding author. Tel.: +34 985103190; fax: +34 985103354.
E-mail addresses: gil.gonzalez@softcomputing.es (G. González-Rodríguez), colubi@uniovi.es (A. Colubi), pierpaolo.durso@uniroma1.it (P. D'Urso), mmontenegro@uniovi.es (M. Montenegro).

0888-613X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2009.01.003

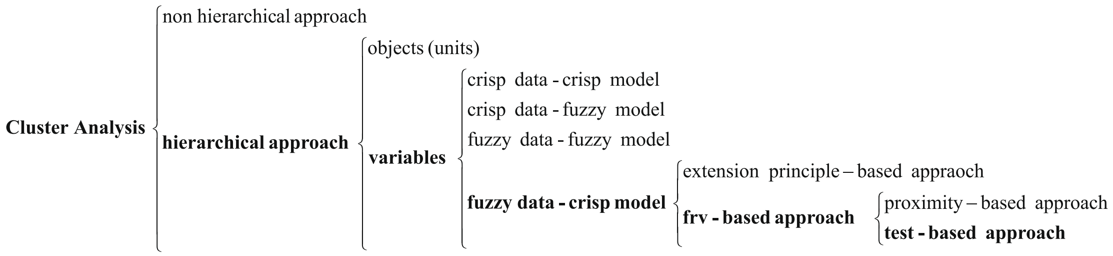

Fig. 1. Different types of clustering approaches.

The common purpose of any cluster analysis is to find a priori unknown groups. The assumption is that the elements of the dataset (objects or variables) within a cluster are in some sense more similar to each other than to objects or variables in other clusters. The position of our problem within a general scheme of different types of clustering approaches is shown in Fig. 1.

The focus here is on hierarchical clustering of variables, which is useful, for instance, to replace a class of closely related variables by a single representative, or a combination, for subsequent analysis [15]. The similarity between variables is typically measured by their correlation, although different proposals have been made in the literature depending on the nature of the considered variables (see, for instance, [19,3,12,23]).

In this work, we will find clusters of fuzzy features by comparing the sample means and by taking into account the stochastic variability. Several (fuzzy) clustering of fuzzy data have been developed (see, for instance, [14] and the references therein). A first attempt to develop a clustering method for random variables gathering both the imprecision and the statistical variability has been advanced in [9] and will be recalled in Section 4. However, the final solution of the approach in [9] had not statistical significance. The aim of this paper is to present a computationally feasible method that overcomes this problem.

Classical clustering procedures based on distances computed from the fuzzy data matrix do not take into account the stochastic variability. Thus, in order to consider the relative statistical variability we propose to develop a procedure inspired from the ideas in [23] based on the p-value of a multi-sample test for the expectations of FRVs (see [16]). This approach is interesting in different settings. For instance, in sociological surveys, people are often asked to answer questions about a topic using a discrete scale from 1 to 5, 1 meaning "total disagreement" (or "the worst option"), and 5 meaning "total agreement" (or "the best option"). However, if people are allowed to express the "degree of precision' of their opinions, the sample information is more expressive. This kind of sample information may be properly described by means of fuzzy sets as shown in the example of Section 6. The technique presented in Section 4 is useful to group either questions which produce similar mean responses, or people with similar mean perceptions about the topic.

The main advantage of the approach we propose is that an objective stopping criterion for the iterative clustering is obtained. Thus, it is not necessary to fix the number of clusters, and when the process concludes, statistically equal groups different from each other are obtained.

Some simulations to illustrate the performance of the clustering algorithm and a case study will be also shown.

The rest of the manuscript is organized as follows. In Section 2 some preliminary concepts and results concerning FRVs are recalled. In Section 3 the clustering criterion is established. Since it is based on the bootstrap multi-sample hypothesis test for the expectations FRVs, the testing procedure is briefly explained. In Section 4 we introduce the iterative procedure to get the clusters on the basis of the p-value of the test. Some simulation studies are shown in Section 5. In Section 6 the iterative procedure is illustrated by means of real-life experiment, and a comparison with some basic hierarchical clustering algorithms is carried out. Finally, some concluding remarks and open problems are gathered in Section 7.

2. Preliminaries

Let $\mathscr{F}_c(\mathbb{R})$ denote the class of *fuzzy numbers* $U : \mathbb{R} \to [0,1]$ whose β-levels U_β are nonempty compact intervals of $\mathbb{R}$, for all $\beta \in [0,1]$, where $U_\beta = \{x \in \mathbb{R} | U(x) \geqslant \beta\}$ for all $\beta \in (0,1]$, and U_0 is the closure of $\{x \in \mathbb{R} | U(x) > 0\}$. Zadeh's extension principle [31] allows us to endow the space $\mathscr{F}_c(\mathbb{R})$ with a sum and a product by a scalar satisfying

$$(U+V)_\beta = U_\beta + V_\beta = \{u + v | u \in U_\beta, v \in V_\beta\} \quad \text{and} \quad (\lambda U)_\beta = \lambda U_\beta = \{\lambda u | u \in U_\beta\}$$

for all $U, V \in \mathscr{F}_c(\mathbb{R})$, $\lambda \in \mathbb{R}$ and $\beta \in [0,1]$.

The multi-sample tests for FRVs in the literature (see, for instance, [25,16]) are established in terms of the (W,φ)-*distance* [6] defined by

$$D_W^\varphi(U,V) = \sqrt{\int_{[0,1]} \int_{[0,1]} [f_U(\beta,\lambda) - f_V(\beta,\lambda)]^2 dW(\lambda) d\varphi(\beta)}$$

for all $U, V \in \mathscr{F}_c(\mathbb{R})$, where $f_U(\beta,\lambda) = \lambda \sup U_\beta + (1-\lambda)\inf U_\beta$. The weighting measures W and φ can be formalized as probability measures on $([0,1], \mathscr{B}_{[0,1]})$ ($\mathscr{B}_{[0,1]}$ being the Borel σ-field on [0,1]), where W is assumed to be associated with a nondegenerate distribution, and φ is assumed to have a strictly increasing distribution function on [0,1] (although these assumptions do not entail in fact a stochastic meaning for W and φ). In [29] it is shown that D_W^φ is invariant to rigid motions if, and only if, the first moment of W is 1/2, and in this case, D_W^φ may be alternatively written as

$$D_W^\varphi(U,V) = \sqrt{\int_{[0,1]} (\text{mid}U_\alpha - \text{mid}V_\alpha)^2 + \theta_W(\text{spr}U_\alpha - \text{spr}V_\alpha)^2 d\varphi(\alpha)},$$

where mid stands for the center of any interval, spr for the half of the length, and $\theta_W = \int_{[0,1]}(2t-1)^2 dW(t)$. It can be shown that $0 < \theta_W \leqslant 1$. Thus, the measure W determines the relative importance of the squared distance between the spreads in relationship with the squared distance between the centers through the weight θ_W. Since $0 < \theta_W \leqslant 1$ the importance of the spreads should be no greater than that of the mid-points. If W is chosen to be the Lebesgue measure, then the relative importance of the spreads is $\theta_W = 1/3$. On the other hand, φ weights the importance of the level sets. For instance, the Lebesgue measure assigns the same importance to every α-level φ, however to give more mass to α-levels close to 1 a different distribution (usually a Beta one) may be chosen. Further discussions and a generalization of this metric for the multi-dimensional setting can be found in [29].

If $(\Omega, \mathscr{A}, P)$ is a probability space, a ($\mathscr{F}_c(\mathbb{R})$-valued) *fuzzy random variable* (FRV) [31] is a mapping $\mathscr{X} : \Omega \to \mathscr{F}_c(\mathbb{R})$ so that the set-valued β-level mappings $\mathscr{X}_\beta : \Omega \to \mathscr{K}_c(\mathbb{R})$ are random sets for all $\beta \in [0,1]$, that is, the $\mathscr{X}_\beta$ are Borel-measurable mappings when the Hausdorff metric is considered on $\mathscr{K}_c(\mathbb{R}) = \{\text{nonempty compact intervals of } \mathbb{R}\}$. Equivalently, a FRV can be defined as a Borel-measurable function w.r.t. the D_W^φ-metric (see [7,20]). We will say that a FRV is *simple* if the cardinality of $\mathscr{X}(\Omega)$ is finite.

The *expected value (or mean)* of an integrably bounded FRV $\mathscr{X}$ (that is, $\mathscr{X}$ satisfying $\max\{|\inf \mathscr{X}_0|, |\sup \mathscr{X}_0|\} \in L^1(\Omega, \mathscr{A}, P)$, where $\mathscr{X}_0$ stands for the 0-level or support of $\mathscr{X}$), is the unique $E(\mathscr{X}) \in \mathscr{F}_c(\mathbb{R})$ such that $(E(\mathscr{X}))_\beta$ is the Aumman integral of the random set $\mathscr{X}_\beta$ for all $\beta \in [0,1]$ (see [31]), that is,

$$\begin{aligned}(E(\mathscr{X}))_\beta &= \{E(f) | f : \Omega \to \mathbb{R}, f \in L^1, f \in \mathscr{X}_\beta, a.s.-[P]\} \\ &= [E(\inf \mathscr{X}_\beta), E(\sup \mathscr{X}_\beta)] \text{ for all, } \quad \beta \in [0,1].\end{aligned}$$

If $\mathscr{X}$ is a FRV so that $\max\{|\inf X_0|, |\sup X_0|\} \in L^2(\Omega, \mathscr{A}, P)$, the (W,φ)-*variance* of $\mathscr{X}$ (see [22,20]) is given by:

$$\text{Var}(\mathscr{X}) = E([D_W^\varphi(\mathscr{X}, E(\mathscr{X}))]^2).$$

In this work we will make use of the variance to quantify the variability of the fuzzy values of the FRV about its expected value. As it is mentioned in the introduction, the nature of the values of the random elements considered here is purely fuzzy, and the statistical aim refers to these values. Then, the distance-based variance is meaningful as a measure of the error of approximating the fuzzy values by their expected value. However, when the aim concerns an imprecisely observed underlying real-valued variable, one may also consider an imprecise variance (interval or fuzzy-valued) by propagating the uncertainty due to the imperfect observation process.

3. Clustering criterion: the multi-sample test

The goal in this paper is to develop a method to group k independent FRVs by focusing on the expected value. Formally, let $(\Omega, \mathscr{A}, P)$ be a probability space and let $\mathscr{X}_i : \Omega \to \mathscr{F}_c(\mathbb{R})$ with $i = 1, \ldots, k$ be k independent simple FRVs. The objective is to find groups $C_l \subset \{1, \ldots, k\}$ in such a way that the FRVs in each group have the same population expected value, that is, $i_1, \ldots i_{k_l} \in C_l$ whenever $E(\mathscr{X}_{i_1}) = \cdots = E(\mathscr{X}_{i_{k_l}})$.

Since it is unusual to know population expected values, some inferential technique should be applied to check the equality of population means from a sample. A sample of k FRVs $(\mathscr{X}_1, \ldots, \mathscr{X}_k)$ on n independent statistical units can be arranged in a fuzzy data matrix $\{\tilde{x}_{i,j}\}_{i=1,\ldots,k}^{j=1,\ldots,n}$ with $\tilde{x}_{i,j} \in \mathscr{F}_c(\mathbb{R})$. It should be noticed that this matrix is the result of a random generation process of fuzzy data. Formally, for each $i = 1, \ldots, k$ we consider a simple random sample $\mathscr{X}_{i,1}, \ldots, \mathscr{X}_{i,n}$, that is, a set of independent FRV distributed as $\mathscr{X}_i$.

In order to test whether the expected values of k_l FRVs are equal we can employ the multi-sample bootstrap test in [16] based on the decomposition of the variability. Let $C = \{i_1, \ldots, i_{k_C}\} \subset \{1, \ldots, k\}$ be a generic set of indices with cardinality k_C. The sample mean and the sample variance corresponding to each population are given by

$$\overline{\mathscr{X}}_i = \frac{1}{n}(\mathscr{X}_{i,1} + \cdots + \mathscr{X}_{i,n}) \quad \text{and} \quad \widehat{S}_i^2 = \frac{1}{n}\sum_{k=1}^n \left[D_W^\varphi(\mathscr{X}_{i,j}, \overline{\mathscr{X}}_i)\right]^2$$

and the overall mean corresponding to the group C is given by

$$\overline{\mathscr{X}_C} = \frac{1}{k_C}(\overline{\mathscr{X}}_{i_1} + \cdots + \overline{\mathscr{X}}_{i_{k_C}}).$$

Let α be the significance level. The basic statistic T to test the null hypothesis $H_0 : E(\mathscr{X}_{i_1}) = \cdots = E(\mathscr{X}_{i_{k_C}})$ is inspired by the classical analysis of variance (ANOVA) test statistic. To be precise, since the D_W^φ allows us to decompose the overall variability as the sum of the variability between the groups plus the variabilities within the groups, T is

$$T = \frac{\sum_{i \in C} n [D_W^\varphi(\overline{\mathscr{X}_i}, \overline{\mathscr{X}_C})]^2}{\sum_{i \in C} \sum_{j=1}^n [D_W^\varphi(\mathscr{X}_{i,j}, \overline{\mathscr{X}_i})]^2}.$$

Large values of this statistic would imply that the difference between the sample means of the groups is relatively too large and the null hypothesis should be rejected. In [16] it is shown that the asymptotic distribution of this statistic is not efficient to compute the critical values, and a bootstrap procedure is suggested.

The bootstrap statistic T^* is defined as follows. Let the bootstrap populations with a common fuzzy mean be defined for each $i \in C$ as a new FRV $\mathscr{Y}_i^*$ taking on values $\mathscr{X}_{i,1} + \overline{\mathscr{X}}_{-i}, \ldots, \mathscr{X}_{i,n} + \overline{\mathscr{X}}_{-i}$, where $\overline{\mathscr{X}}_{-i}$ is the (fuzzy) sum of all the available sample means in C but without the ith one. Then, we will resample from these new populations, that is, for any $i \in C$ we draw a sample of n independent observations $\mathscr{Y}_{i1}^*, \ldots, \mathscr{Y}_{in}^*$ from $\mathscr{Y}_i^*$. The bootstrap test statistic is given by

$$T^* = \frac{\sum_{i \in C} n [D_W^\varphi(\overline{\mathscr{Y}_i^*}, \overline{\mathscr{Y}_C^*})]^2}{\sum_{i \in C} \sum_{j=1}^n \left[D_W^\varphi(\mathscr{Y}_{i,j}^*, \overline{\mathscr{Y}_i^*})\right]^2}.$$

The p-value of the bootstrap test is the probability of $T^* > T$. The distribution of the bootstrap statistic can be approximated, as usual, by Monte Carlo simulation. The Monte Carlo approximation requires to simulate $(\mathscr{Y}_{i,1}^*, \ldots, \mathscr{Y}_{i,n}^*)$ from each bootstrap population $\mathscr{Y}_i^*$ a large numbers of times B in order to obtain a large sample of values of the statistic $\{T_1^*, \ldots, T_B^*\}$ close enough to the population distribution. The usual values of B range from 1000 to 10,000. Thus, the p-value is approximately the proportion of values in $T_1^*, \ldots, T_B^*$ greater than or equal to the value of the statistic T. The null hypothesis of equality of population means will be rejected whenever the bootstrap p-value is lower than α.

4. Clustering method

In the time series context, [23] shows the use of the p-value of a two-sample test in order to get a matrix of similarities to apply later a hierarchical clustering. Based on the same idea, we propose to use the p-value to cluster FRVs.

A first attempt to consider the p-values to cluster FRVs is described in [9] and can be summarized as follows. Let $\mathscr{X}_{i,1}, \ldots, \mathscr{X}_{i,n}$ be the sample of independent FRV distributed as $\mathscr{X}_i$ for each $i = 1, \ldots, k$. In these conditions we can apply pairwise the two-samples tests in [25] for each $i_1, i_2 \in \{1, \ldots, k\}$. Thus, we denote by p_{i_1,i_2} the p-value of the test $H_0 : E(\mathscr{X}_{i_1}) = E(\mathscr{X}_{i_2})$. According to [23] we use the similarity measure $d(i_1, i_2) = 1 - p_{i_1,i_2}$ to obtain a hierarchical clustering method with any of the usual linkage criteria. Since the p-value of the two-sample test can be interpreted as a kind of "relative distance" between the population expected values, this clustering procedure agrees with the original target of grouping fuzzy random variables with similar expected values.

Since both the variables and the individuals are independent, the procedure may be used to cluster either the variables or the individuals. Nevertheless, this preliminary approach only considers the inferential process in a first stage (to obtain the matrix of similarities to be used in all the hierarchical process), and it is not possible to guarantee that the elements clustered at any other stage are really statistically equal. For this reason, we introduce a new iterative method.

If the aim is clustering, for instance, the variables, once we have linked two variables, we have a group with a sample of size $2 \times n$ (2 data of n individuals). Then, we can apply again the two-sample test to compare the new cluster with other variable (for which we have a sample of size n), because the independence of the variables and the individuals assures that the available data for the new group is again a random sample. However, if we compare groups with very different sample sizes (as $2 \times n$ versus n), the two-sample test is not efficient (Beren–Fisher problem). In contrast, we will employ the multi-sample test which produces much more suitable results.

The procedure we propose here consists in starting with the singleton groups $C_l = \{l\}$ and computing the symmetric matrix of p-values obtained from the bootstrap tests $H_0 : E(\mathscr{X}_{i_1}) = E(\mathscr{X}_{i_2})$ for all $i_1, i_2 \in \{1, \ldots, k\}$. If there is any p-value greater than or equal to the fixed significance level α, then the corresponding groups should be linked. Otherwise, the groups would be pairwise different at this significance level, and the process should terminate. At each stage, we should link the two groups of variables with the greatest p-value (whenever it be greater than or equal to α) and start again the process by considering the new group.

The stopping criterion of the iterative clustering procedure is then determined by the significance level, that is, by the probability of type I error that we are willing to tolerate. The usual levels in Statistics that allow us to balance the probability of type I and II ranges from 0.01 to 0.1.

The naive implementation of this procedure implies a high computational cost, because it is necessary to approximate the distributions of the bootstrap statistics by Monte Carlo and it entails the computation of many distances between fuzzy sets (between/within variability). However, in Proposition 4.1 we will show an alternative expression of the "between variability within each group C" that will allow us to compute most of the necessary quantities in terms of some other ones computed at the beginning of the process. In order to introduce the decomposition some further notation is required.

Let $\mathscr{X}_i^{\top} = \mathscr{X}_{i,1} + \cdots + \mathscr{X}_{i,n}$ be the overall sum on each sample $i = 1, \ldots, k$, $\mathscr{X}_C^{\top} = \mathscr{X}_{i_1}^{\top} + \cdots + \mathscr{X}_{i_{k_C}}^{\top}$ be the overall sum in a group C and $S_C^{2\top} = \sum_{i \in C} \widehat{S}_i^2$ be the sample variance sum in the group C.

The aim of Proposition 4.1 is to decompose the "between variability within any group C" for each simulation considered for the Monte Carlo approximation of the distribution of the bootstrap statistic. Thus, consider a simulated sample $(\mathscr{X}_{i,1}^*, \ldots, \mathscr{X}_{i,n}^*)$ of i.i.d. FRVs from $\{\mathscr{X}_{i,1}, \ldots, \mathscr{X}_{i,n}\}$ for each $i \in C$. If we set $\mathscr{Y}_{i,j}^* = \mathscr{X}_{i,j}^* + \overline{\mathscr{X}}_{-i}$ for all $i \in C$ and $j = 1, \ldots, n$ then, $(\mathscr{Y}_{i,1}^*, \ldots, \mathscr{Y}_{i,n}^*)$ is a sample of i.i.d. random variables from the bootstrap population $\mathscr{Y}_i^*$ defined in Section 3. Let $\tilde{0}$ be the null fuzzy set (i.e., the characteristic set of the singleton $\{0\}$). The next result can be proved by using the expression of the metric D_W^{φ} in terms of the support functions (see [20]).

Proposition 4.1. The between variability within any group $C \subset \{1, \ldots, k\}$ can be decomposed as follows:

(i) $\sum_{i \in C} n \left[D_W^{\varphi}(\overline{\mathscr{X}_i}, \overline{\mathscr{X}_C})\right]^2 = \sum_{i \in C} \frac{1}{n} \left[D_W^{\varphi}(\mathscr{X}_i^{\top}, \tilde{0})\right]^2 - \frac{1}{nk_C} \left[D_W^{\varphi}(\mathscr{X}_C^{\top}, \tilde{0})\right]^2$,

(ii) $\sum_{i \in C} n \left[D_W^{\varphi}(\overline{\mathscr{Y}_i^*}, \overline{\mathscr{Y}_C^*})\right]^2 = \sum_{i \in C} \frac{1}{n} \left[D_W^{\varphi}(\mathscr{X}_i^{*\top}, \mathscr{X}_i^{\top})\right]^2 - \frac{1}{nk_C} \left[D_W^{\varphi}(\mathscr{X}_C^{*\top}, \mathscr{X}_C^{\top})\right]^2$.

Proposition 4.1 states that, for any group C, the between variability can be computed in terms of the overall sum in each of the populations in C of their distances to the fuzzy set 0 and a unique new distance. In the same way, the between variability for each bootstrap sample can be computed in terms of the distances of the overall sums of the sample and the resample and a new term. Then, if the resample process for all the Monte Carlo approximations is made at the beginning, we can compute in a first step the corresponding distances for the k populations and the computational effort of the naive approach is substantially improved.

On the basis of the previous comments and this proposition, the clustering algorithm can be applied in practice as follows. In a first step the significance level determining the stopping criterion and the initial solution is fixed. In a second step all the resamples to make the Monte Carlo approximation are drawn. Next, the basic quantities of the decomposition of the "between variability" stated in Proposition 4.1 for the k population (Step 3) and the corresponding bootstrap samples (Step 4) are respectively computed (note that the "between variability" corresponding to the bootstrap samples $\mathscr{Y}_i^*$ is simply written in Proposition 4.1 in terms of $\mathscr{X}_i^*$). Then, the p-values of the bootstrap tests are computed (Step 5). A p-value greater than the significance level implies that the involved groups are statistically equal and should be linked. In this case, two groups with the greatest p-value are linked and the process starts again with this new solution. Otherwise the algorithm ends. That is:

Clustering algorithm

Step 1. Fix a significance level α, the number of bootstrap replications B and consider the singleton clusters $\{i\}$ with $i = 1, \ldots, k$.

Step 2. Obtain B samples of i.i.d. random variables $(\mathscr{X}_{i,1}^{b*}, \ldots, \mathscr{X}_{i,n}^{b*})$ $(b = 1, \ldots, B)$ from $(\mathscr{X}_{i,1}, \ldots, \mathscr{X}_{i,n})$ for each $i = 1, \ldots, k$.

Step 3. For each cluster C compute $\mathscr{X}_C^{\top}$, $S_C^{2^{\top}}$ and $D_C = [D(\mathscr{X}_C^{\top}, \widetilde{0})]^2/n$.

Step 4. For each cluster C and each $b = 1, \ldots, B$, compute the overall sum and sample variance sum of each resample b, that is, $\mathscr{X}_C^{b*\top}$, $S_C^{2b*\top}$ and $D_C^{b*} = [D(\mathscr{X}_C^{b*\top}, \mathscr{X}_C^{\top})]^2/n$.

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:

1. Set $\text{count}_{C_1,C_2} = 0$ and compute the value of the k-sample test statistic

$$T = \frac{(D_{C_1} + D_{C_2}) - [D(\mathscr{X}_{C_1}^{\top} + \mathscr{X}_{C_2}^{\top}, \tilde{0})]^2/(n(k_{C_1} + k_{C_2}))}{S_{C_1}^{2\top} + S_{C_2}^{2\top}}.$$

2. For each $b = 1, \ldots, B$ compute the value of the bootstrap statistic

$$T^{b*} = \frac{(D_{C_1}^{b*} + D_{C_2}^{b*}) - \frac{D(\mathscr{X}_{C_1}^{b*\top} + \mathscr{X}_{C_2}^{b*\top}, \mathscr{X}_{C_1}^{\top} + \mathscr{X}_{C_2}^{\top})^2}{(n(k_{C_1} + k_{C_2}))}}{S_{C_1}^{2b*\top} + S_{C_2}^{2b*\top}}.$$

3. For each $b = 1, \ldots, B$, IF $T^{b*} > T$, THEN $\text{count}_{C_1,C_2} = \text{count}_{C_1,C_2} + 1$.
4. Compute the bootstrap p-value $p_{C_1,C_2} = \text{count}_{C_1,C_2}/B$.

Step 6. Find the greatest p-value Gp and label by $C_{\mathbb{A}}$ and $C_{\mathbb{B}}$ one of the couples with this p-value.

Step 7. IF $Gp \geqslant \alpha$, THEN link the groups $C_{\mathbb{A}}$ and $C_{\mathbb{B}}$ in a new group $C = C_{\mathbb{A}} \cup C_{\mathbb{B}}$ and compute the quantities in Steps 2 and 4 for this new group, that is,

$$\mathscr{X}_C^{\top} = \mathscr{X}_{C_{\mathbb{A}}}^{\top} + \mathscr{X}_{C_{\mathbb{B}}}^{\top}, S_C^{2\top} = S_{C_{\mathbb{A}}}^{2\top} + S_{C_{\mathbb{B}}}^{2\top}, \text{ and } D_{C^*} = [D(\mathscr{X}_C^{\top}, \tilde{0})]^2/n;$$

and for each $b = 1, \ldots B$,

$$\mathscr{X}_C^{b*\top}, S_C^{2b*\top}, \quad \text{and} \quad D_C^{b*} = [D(\mathscr{X}_C^{b*\top}, \mathscr{X}_C^{\top})]^2/n.$$

Finally, compute the p-values between the new group and the other ones as in Step 5.1–4 and GO TO Step 6;

ELSE Stop.

The next theorem shows that the clusters obtained by applying this approach are statistically pairwise different, and the FRVs within each cluster are statistically equal in mean.

Theorem 4.1. The above Clustering algorithm generates groups $C_1 \ldots C_m$ such that

(1) for all $l_1 \neq l_2 \in \{1, \ldots, m\}$ there exists $i_1 \in C_{l_1}$ and $i_2 \in C_{l_2}$ so that $\mu_{i_1} \neq \mu_{i_2}$ at the fixed significance level α;
(2) all the FRVs $\mathscr{X}_i$ in the cluster C_l have the same (fuzzy) expected value μ_l at the fixed significance level α.

Proof. Concerning (1), let C_{l_1} and C_{l_2} be two clusters of the final solution. Since the algorithm terminates when the null hypothesis of equality of expected values of the variables belonging to the joint set is rejected (at the significance level α) for every pair of clusters, then there exists at least $i_1 \in C_{l_1}$ and $i_2 \in C_{l_2}$ so that $\mu_{i_1} \neq \mu_{i_2}$. Otherwise the null hypothesis would not have been rejected and the algorithm would continue.

On the other hand, let C_l be a cluster of the final solution. If C_l is a singleton, (2) is obvious. Otherwise, C_l was obtained at the end of an iteration in which the hypothesis of equality of the fuzzy expected values of the variables belonging to this cluster was not rejected, otherwise the linking criterion would not have been fulfilled. Then, the expected value of all FRVs $\mathscr{X}_i$ in C_l is the same. □

Theorem 4.1 shows the main advantages of the clustering procedure proposed in this paper. In contrast to classical methods, as hierarchical clustering or k-means clustering approaches, we assure that the groups are statistically equal and different from each other. In addition it is not necessary to fix the number of clusters or to apply further procedures to find the optimal configuration, because it is automatically (and objectively) determined by the probability of type I error that we are willing to tolerate.

It should be noted that the iterative process prevents us from finding a global significance level of the approach. On the other hand, the clustering procedure in this section may be also interpreted as a "post hoc" statistical procedure to determine equal sub-groups when the hypothesis of equality of all the population means is rejected by the bootstrap multi-sample test.

5. Simulation studies

Simulations in this section are considered to illustrate the statistical validity of the inferential conclusions contained in Theorem 4.1. Weighted measures W and φ have been chosen to be the Lebesgue measure on $[0,1]$. Each simulation corresponds to 10,000 iterations, the number of bootstrap replications was 1000, and the significance level α was chosen to be equal to 0.05.

In order to simulate general FRVs the methodology introduced in [18] has been employed. A finite number of $n_0 = 101$ equally spaced alpha cuts ($\alpha_i = (i-1)/(n_0-1), i = 1, \ldots, n_0$) has been fixed. Three FRVs $\mathscr{X}_1, \mathscr{X}_2$ and $\mathscr{X}_3$ with different expectations were simulated. In Table 1 the parameters defining these FRVs according to the methodology in [18] are established. Fuzzy set V in Table 1 stands for the expectation of the corresponding FRV. Concretely, it is recursively defined level-wise by means of

$$V_{1-\alpha_1} = V^c + [-c_1^l, c_1^r] \quad \text{with } c_1^k = V^k F^k(\alpha_1), \quad k = l, r,$$
$$V_{1-\alpha_i} = V_{1-\alpha_{i-1}} + [-c_i^l, c_i^r] \quad \text{with } c_i^k = V^k F^k(\alpha_i) - c_{i-1}^k, \quad k = l, r;\ i = 2, \ldots, n_0.$$

On the other hand, C is connected with the way of perturbing the above-defined coefficients. To be precise, a general FRV $\mathscr{X}$ is level-wise defined as follows:

$$\mathscr{X}_{1-\alpha_1} = C^0 + [-c_1^l T_i^l, c_1^r T_i^r], \quad \mathscr{X}_{1-\alpha_i} = \mathscr{X}_{1-\alpha_{i-1}} + [-c_i^l T_i^l, c_i^r T_i^r], \quad i = 2, \ldots, n_0,$$

where $(C^0, T_1^l, \ldots, T_{n_0}^l, T_1^r, \ldots, T_{n_0}^r)$ is an $\mathbb{R}^{2n_0+1}$-valued random vector determined by the $(2n_0+1)$-dimensional copula Cp and with marginal distributions $C^0 \sim D^0$, $T_i^l \sim D^l$ and $T_i^r \sim D^r$, $i = 1, \ldots, n_0$.

Simulations concerning the probabilities of type I error have been firstly developed. In Table 2, k populations with the same mean (all of them distributed as $\mathscr{X}_1$) have been simulated for different values of k, and the frequency distribution of the obtained number of clusters has been recorded. Since the bootstrap techniques produce suitable results for moderate

Table 1
Parameters defining the FRVs for the simulation approach in [18].

	$V = [V^c, V^l, V^r, F^l, F^r]$	$C = [D^0, D^l, D^r, C_p]$
$\mathscr{X}_1$	[3,1,1,Beta(1,1),Beta(1,1)]	$[\chi_3, \chi_2/2, \chi_1/\sqrt{2}, \Pi]$
$\mathscr{X}_2$	[4,2,2,Beta(0.5,1),Beta(3,2)]	$[U(0,8), \chi_1/\sqrt{2}, \chi_2/2, \Pi]$
$\mathscr{X}_3$	[5,0.5,0.5,Beta(1,2),Beta(2,1)]	$[\mathscr{N}(5,3), \chi_1/\sqrt{2}, \chi_3/\sqrt{6}, \mathit{\Pi}]$

Table 2
Simulations for clustering k variables with equal means. Percentage of the obtained number of clusters ($\alpha = 0.05$).

	$k = 5$	$k = 10$	$k = 20$
1 Cluster	95.36	95.43	95.28
2 Clusters	4.64	4.59	4.72

sample sizes, we have fixed $n = 30$. As to be expected, in this case the procedure has a success rate similar to the empirical size of the bootstrap method.

We have also considered three main populations with different mean values (distributed as $\mathscr{X}_1$, $\mathscr{X}_2$ and $\mathscr{X}_3$, respectively). The results of the bootstrap procedure depend on both the sample size and the distance between the population means. We have fixed the distance between the different means and we have considered several sample sizes in order to observe the improvement. We have simulated three FRVs from the first population, four from the second and two from the third. In Table 3 we have recorded the frequency distribution of the obtained number of clusters. We see that the results are better for large sample sizes. However, it should be noted that the applied method implies that the larger the sample size is, the higher the fragmentation.

6. Experimental results

From a survey about the personal experience as graduate students of different people, we have collected the answers to 15 questions which seems to be a representative sample of the perception of such people, in the sense of being independent (which seems to be supported by the sample data) and covering different aspects. Graduate students, graduates, post-graduate students and post-graduates from different degrees and universities have been considered. A total of 58 people were asked to evaluate the following aspects:

Q.1 Academic standards.
Q.2 Variety of subjects.
Q.3 Teaching methods.
Q.4 Evaluation system.
Q.5 Practice teaching system.
Q.6 Opportunities for specialization.
Q.7 Academic advising.
Q.8 Technical equipment.
Q.9 Stress on practice.
Q.10 Stress on methodology.
Q.11 Internship programs.
Q.12 Relationship with labor market demands.
Q.13 Opportunity of taking part in I+D projects.
Q.14 Contact with the teachers.
Q.15 Contact with other students.

Although, in these kinds of surveys people are often asked to answer in a discrete scale from 1 to 5, (1 meaning "total disagreement", or "the worst option", and 5 meaning "total agreement", or "the best option"), people here were allowed to express the degree of precision of their opinions by using trapezoidal fuzzy sets. In this way, people fixed the 0-level as the set of all values that they consider compatible with their opinion to a greater or lesser extent, and the 1-level consists of all the values that they consider completely compatible with their opinion. In Fig. 2 some of the obtained sample values are shown.

The trapezoidal sample means as well as the dispersion $\left(\text{i.e.,}\widehat{S}_i = \sqrt{\widehat{S}_i^2}\right)$ of the sample data are gathered in Table 4. At the first glance, it is difficult to obtain a classification, since the fuzzy values are not easy to rank, and in the procedure both the fuzzy mean value and the stochastic variability are involved. In Fig. 3 we have represented the scatter plot of the questions

Table 3
Simulations for clustering 9 FRVs with three different means. Percentage of the obtained number of clusters ($\alpha = 0.05$).

	$n = 30$	$n = 50$	$n = 100$
1 Cluster	4.38	0.09	0
2 Clusters	71.97	37.93	3.12
3 Clusters	**23.53**	**60.58**	**88.82**
4 Clusters	0.12	1.40	7.94
5 Clusters	0	0	0.12

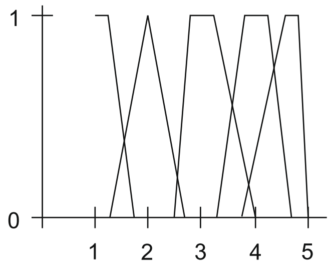

Fig. 2. Triangular and trapezoidal fuzzy description of some sample values.

Table 4
Sample means and dispersions.

	$\inf(\overline{Q_i})_0$	$\inf(\overline{Q_i})_1$	$\sup(\overline{Q_i})_0$	$\sup(\overline{Q_i})_1$	Dispersion
Q.1	2.98	3.75	4.21	4.47	0.56
Q.2	3.14	3.96	4.31	4.54	0.66
Q.3	2.76	3.59	3.93	4.29	0.77
Q.4	2.19	2.96	3.38	3.92	0.87
Q.5	2.00	2.71	2.95	3.58	1.11
Q.6	2.54	3.22	3.49	3.93	1.30
Q.7	1.75	2.42	2.76	3.44	1.03
Q.8	2.22	2.86	3.31	3.83	1.06
Q.9	1.97	2.62	2.87	3.51	1.20
Q.10	1.81	2.50	2.75	3.42	1.08
Q.11	1.66	2.16	2.31	3.02	1.23
Q.12	1.81	2.47	2.72	3.36	1.13
Q.13	1.86	2.45	2.70	3.34	1.22
Q.14	2.85	3.66	3.83	4.19	1.23
Q.15	3.14	3.99	4.25	4.47	0.88

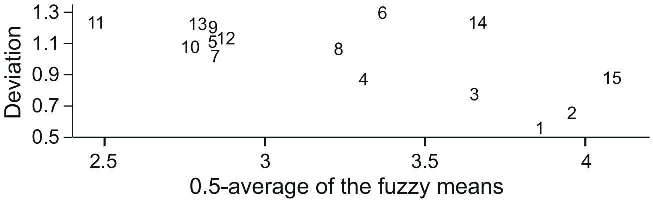

Fig. 3. Scatter plot. Sample means vs. dispersion.

by considering a defuzzification of the sample means (according to the 0.5-average criterion [8]) in the x-axis, and the dispersion in the y-axis. Since in our approach the means are more important than the variability, we have weighted the x-scale to better observe the possible clusters. We observe a clear cluster grouping questions 5, 7, 9, 10, 12 and 13. Question 11 has a lower 0.5-average, although similar dispersion. Questions 4, 6 and 8 are similar in 0.5-average, but more variable in dispersion. Finally, questions 1, 2, 3, 14 and 15 have a greater 0.5-average and are less concentrated both in mean and dispersion.

For comparative purposes, we have applied four clustering procedures, namely,

C1. A classical "Euclidean-type" distance-based hierarchical clustering (i.e., the similarity between the variables is computed in terms of the sum of the square D_W^φ-distance between the answer of the different interviewee).

C2. A distance-based hierarchical clustering focused on the fuzzy means (i.e., the similarity between the variables is computed in terms of D_W^φ-distance between the mean answer of the different respondents).

C3. The hierarchical clustering based on the matrix of similarities provided by the p-value of the two-sample test [9] explained in Section 4.

C4. The procedure suggested in this paper.

Different linkage criteria were considered for the classical hierarchical clusterings. When different results were obtained, one close to the intuitive solution in Fig. 3 was chosen. To be precise, we have chosen the Wald linkage for **C1** and **C2** and single linkage for **C3**.

The results are displayed in Figs. 4–7. **C2** and **C3** provide us with the same solution than **C4** in one of the steps, although for nonnatural stopping points. In contrast, the results for **C1** are completely different (as one may expect, since it is the unique one which does not deal with the means).

The number of cluster is prefixed in none of the procedures, however, while an objective stopping criterion leads automatically to three groups for the **C4**, further analysis are required to find a suitable cut for the other approaches.

In Table 5 we present the questions in each of the groups determined by **C4** (see also Fig. 3). The first group includes the most highly valued aspect, the second one includes aspect which produces the most indifferent/disperse perception, and the

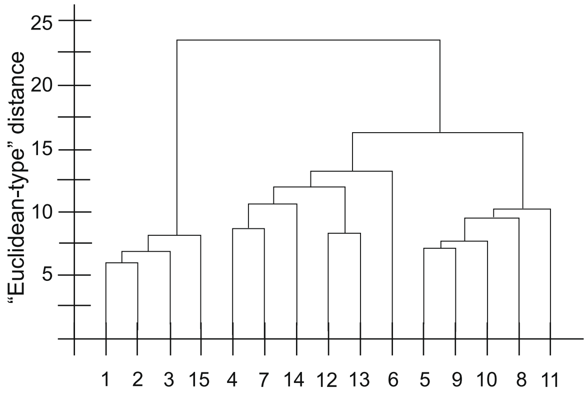

Fig. 4. Dendrogram of **C1**.

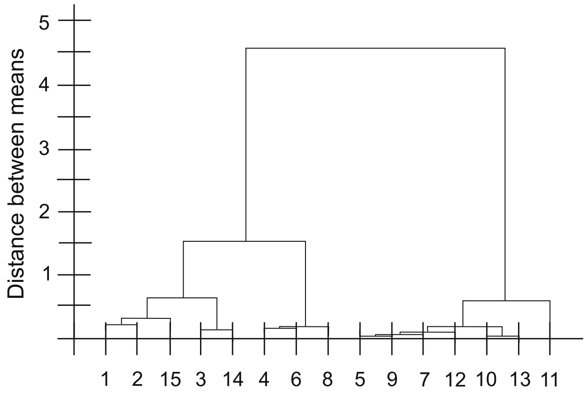

Fig. 5. Dendrogram of **C2**.

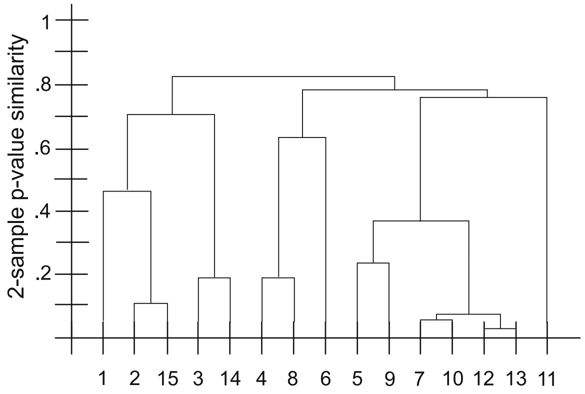

Fig. 6. Dendrogram of **C3**.

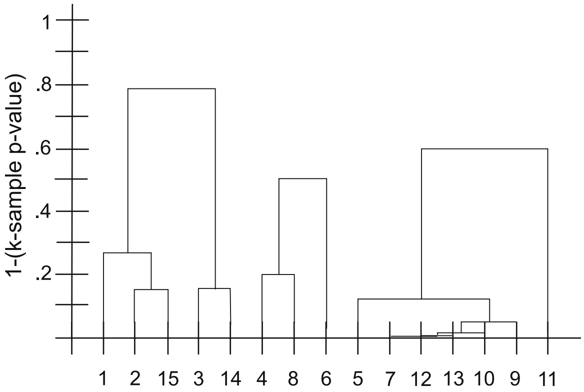

Fig. 7. Dendrogram of **C4** for $\alpha = 0.05$.

Table 5
Groups of questions with similar mean responses.

Q.1	Academic standards
Q.2	Variety of subjects
Q.3	Teaching methods
Q.14	Contact with the teachers
Q.15	Contact with other students
Q.4	Evaluation system
Q.6	Opportunities for specialization
Q.8	Technical equipment
Q.5	Practice teaching system
Q.7	Academic advising
Q.9	Stress on the practice
Q.10	Stress on methodology
Q.11	Internship programs
Q.12	Relationship with labor market demands
Q.13	Opportunity of taking part in ID projects

questions with worst response are included in the third group. The conclusions from the applied method can be interpreted by saying that people consider that the *background* acquired at the university (standards, subjects, teaching quality, and so on) is a good one, but the *implementation* of this background *to practice* (practice and training systems, useful advising,...) does not fit their expectations.

7. Conclusions

We have suggested a partition algorithm for classifying FRVs based on a statistical test procedure. We explicitly considered in the clustering process a double source of uncertainty: randomness and fuzziness. In this sense, the main features of the suggested clustering procedures are:

- the possibility of applying the clustering algorithms in various observational settings, including some of the usual categorical data such as subjective judgments, imprecise measurements and so on;
- the use of fuzzy sets to represent the imprecision or vagueness of data obtained in a random experiment;
- the use of powerful inferential tools concerning FRVs;
- the consideration of suitable proximity distances between FRVs and fuzzy data taking into account randomness and fuzziness.

In contract to classical methods, the approach in this paper leads to statistically equal groups different from each other with an objective and natural stopping criterion.

Since the clustering method is based on a multi-sample tests for independent FRVs, the approach could be applied also in the case of samples with different size, and may be also interpreted as a "post hoc" statistical procedure to determine equal sub-groups when the hypothesis of equality of all the population means is rejected by the bootstrap multi-sample test.

The performance of the clustering algorithms for FRVs was checked by means of a simulation study and a real-life application.

Some future directions related to the research in this paper concern the classification of dependent FRVs, the partition of FRVs based on fuzzy (nonhierarchical) clustering procedures, the clustering of FRVs based on mixture models, the cluster analysis of more complex data (e.g., fuzzy random three-ways arrays, fuzzy random time arrays, and so on), along with comparative studies.

Acknowledgements

Authors are very grateful to the anonymous referees and the Editor. Their valuable comments and suggestions have substantially improved the manuscript. The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grants MTM2005-00045 and MTM2006-07501 and the Italian Ministry of Education, University and Research Grant PRIN 2005. Their financial support is gratefully acknowledged.

References

[1] M.C. Alonso, T. Brezmes, M.A. Lubiano, C. Bertoluzza, A generalized real-valued measure of the inequality associated with a fuzzy random variable, Int. J. Approx. Reason. 26 (2001) 47–66.

[2] B.F. Arnold, O. Gerke, Testing fuzzy linear hypotheses in linear regression models, Metrika 57 (2003) 81–95.

[3] D.J. Bartholomew, F. Steele, I. Moustaki, J.I. Galbraith, The Analysis and Interpretation of Multivariate Data for Social Scientists, Chapman & Hall, 2002.

[4] C. Baudrit, I. Couso, D. Dubois, Joint propagation of probability and possibility in risk analysis: towards a formal framework, Int. J. Approx. Reason. 45 (2007) 82–105.

[5] C. Baudrit, D. Guyonnet, D. Dubois, Joint propagation of variability and imprecision in assessing the risk of groundwater contamination, J. Contam. Hydrol. 93 (2007) 72–84.

[6] C. Bertoluzza, N. Corral, A. Salas, On a new class of distances between fuzzy numbers, Mathware Soft Comput. 2 (1995) 71–84.

[7] A. Colubi, J.S. Domínguez-Menchero, M. López-Díaz, D.A. Ralescu, On the formalization of fuzzy random variables, Inform. Sci. 133 (2001) 3–6.

[8] L. Campos, A. Gonzalez, Further contributions to the study of the average value for ranking fuzzy numbers, Int. J. Approx. Reason. 10 (1994) 135–153.

[9] A. Colubi, G. González-Rodríguez, M. Montenegro, P. D'Urso, A cluster procedure for independent fuzzy random variables, in: Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty, Paris, 2006, pp. 965–969.

[10] A. Colubi, Statistical inference about the means of fuzzy random variables: applications to the analysis of fuzzy- and real-valued data, Fuzzy Sets Syst. 160 (2009) 344–356.

[11] R. Coppi, Management of uncertainty in Statistical Reasoning: the case of Regression Analysis, Int. J. Approx. Reason. 47 (2008) 284–305.

[12] S. Csörgõ, W.B. Wu, On the clustering of independent uniform random variables, Random Struct. Algor. 25 (2004) 396–420.

[13] T. Denoeux, M. Masson, P.-A. Hébert, Nonparametric rank-based statistics and significance tests for fuzzy data, Fuzzy Sets Syst. 153 (2005) 1–28.

[14] P. D'Urso, Fuzzy clustering of fuzzy data, in: J.V. de Oliveira, W. Pedrycz (Eds.), Advances in Fuzzy Clustering and its Applications, John Wiley and Sons, 2007, pp. 155–192.

[15] A.H. El-Shaarawi, W.W. Piegorsch, Encyclopedia of Environmetrics, Wiley, 2002.

[16] M.A. Gil, M. Montenegro, G. González-Rodríguez, A. Colubi, M.R. Casals, Bootstrap approach to the Multi-Sample Test of Means with Imprecise Data, Comput. Stat. Data Anal. 51 (2006) 148–162.

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

[18] G. González-Rodríguez, A. Colubi, W. Trutschnig, Simulation of fuzzy random variables, Inform. Sci. (2008), doi:10.1016/j.ins.2008.10.018.

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

[20] R. Körner, W. Näther, On the variance of random fuzzy variables, in: C. Bertoluzza, M.A. Gil, D.A. Ralescu (Eds.), Statistical Modeling Analysis and Management of Fuzzy Data, Physica-Verlag, Heidelberg, 2002, pp. 22–39.

[21] H. Kwakernaak, Fuzzy random variables, definition and theorems, Inform. Sci. 15 (1989) 1–29.

[22] M.A. Lubiano, M.A. Gil, M. López-Díaz, M. López-García, The $\overrightarrow{\lambda}$-mean squared dispersion associated with a fuzzy random variable, Fuzzy Sets Syst. 111 (2000) 307–317.

[23] E.A. Maharaj, A significance test for classifying ARMA models, J. Stat. Comput. Simul. 54 (1996) 305–331.

[24] S. Mitra, Y. Hayashi, Bioinformatics with soft computing, IEEE Trans. Syst. Man Cyber. – Part C 36 (2006) 616–635.

[25] M. Montenegro, M.R. Casals, M.A. Lubiano, M.A. Gil, Two-sample hypothesis tests of means of a fuzzy random variable, Inform. Sci. 133 (2001) 89–100.

[26] M. Montenegro, A. Colubi, M.R. Casals, M.A. Gil, Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable, Metrika 59 (2004) 31–49.

[27] W. Näther, Regression with fuzzy random data, Comput. Stat. Data Anal. 51 (2006) 235–252.

[28] A. Torres, J.J. Nieto, Fuzzy logic in medicine and bioinformatics, J. Biomed. Biotechnol. (2006) 1–7.

[29] W. Trutschnig, G. González-Rodríguez, A. Colubi, M.A. Gil, A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread, submitted for publication.

[30] S. Petit-Renaud, T. Denœux, Nonparametric regression analysis of uncertain and imprecise data using belief functions, Int. J. Approx. Reason. 35 (2004) 1–28.

[31] M.L. Puri, D.A. Ralescu, Fuzzy random variables, J. Math. Anal. Appl. 114 (1986) 409–422.

[32] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, Part 1, Inform. Sci. 8 (1975) 199–249; Part 2, Inform. Sci. 8 (1975) 301–353; Part 3, Inform. Sci. 9 (1975) 43–80.