Econometrics and Statistics 26 (2023) 84–98

Contents lists available at ScienceDirect

Econometrics and Statistics

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

Fuzzy sets and (fuzzy) random sets in Econometrics and Statistics

Ana Colubi[a,b,*], Ana Belén Ramos-Guajardo[c,d]

[a] Department of Mathematics, King's College London, UK
[b] Frederick University, Cyprus
[c] Department of Statistics, University of Oviedo, Spain
[d] Indurot, University of Oviedo, Spain

ARTICLE INFO

Article history:
Received 11 April 2021
Revised 1 July 2022
Accepted 1 July 2022
Available online 8 July 2022

Keywords:
Interval data
Fuzzy data
Set-valued data
Hilbert spaces
Fuzzy clustering
Fuzzy regression discontinuity designs
Random (fuzzy) sets
Fuzzy regression
Regression with fuzzy data

ABSTRACT

Fuzzy sets generalize the concept of sets by considering that elements belong to a class (or fulfil a property) with a degree of membership (or certainty) ranging between 0 and 1. Fuzzy sets have been used in diverse areas to model gradual transitions as opposite to abrupt changes. In econometrics and statistics, this has been especially relevant in clustering, regression discontinuity designs, and imprecise data modelling, to name but a few. Although the membership functions vary between 0 and 1 as the probabilities, the nature of the imprecision captured by the fuzzy sets is usually different from stochastic uncertainty. The aim is to illustrate the advantages of combining fuzziness, imprecision, or partial knowledge with randomness through various key methodological problems. Emphasis will be placed on the management of non-precise data modelled through (fuzzy) sets. Software to apply the reviewed methodology will be suggested. Some open problems that could be of future interest will be discussed.

1. Introduction

Fuzzy sets were introduced in Zadeh (1965) as mathematical objects defined via a membership function that generalizes the characteristic function of a set by associating each element of a class with a grade ranging between 0 and 1. The concept of fuzzy set has been long related to fuzzy logic and fuzzy systems, since their original use regards largely to that area (see, e.g., Mamdani and Assilian, 1975; Takagi and Sugeno, 1985 and Zadeh, 1973; Zadeh, 1975). However, the flexibility of considering gradual transitions as opposite to abrupt changes has made the idea flourish fast in a variety of problems where full membership/non-membership is sometimes too restrictive, such as clustering or decision-making (see, e.g., Bezdek, 1981; Bellman and Zadeh, 1970; Ruspini, 1969 and Zimmermann, 1976).

Fuzzy logic and fuzzy inference systems have been used to incorporate imprecise or vaguely-defined information, such as natural language concepts, in decision-making problems or knowledge-based forecasting in economics (see, e.g., Chen et al., 2006 and Cheng et al., 2013). The impact of many of these approaches in applied econometrics literature has been discrete, partially due to the lack of any objective or statistical validity measure of the conclusions, even when statistical concepts are involved (see, e.g., Maciel et al., 2016 and Wang et al., 2011). Nevertheless, the use of imprecise information has been

[*] Corresponding author.
E-mail addresses: colubi@uniovi.es (A. Colubi), ramosana@uniovi.es (A.B. Ramos-Guajardo).

https://doi.org/10.1016/j.ecosta.2022.07.001

proved to be statistically valuable in certain situations (see, e.g., Fischer et al., 2016 and Manski and Tamer, 2002 for the particular case of interval data).

Other relevant uses of the concept of fuzziness in econometrics and statistics can be found in the context of clustering, where the membership to a cluster is not sharp, or regression-discontinuity designs (see, e.g., Bezdek, 1981; Calonico et al., 2014; Hahn et al., 2001 and Yang et al., 2006), where there is some blurriness around the threshold determining who is eligible for certain treatment.

Fuzzy sets have also been frequently used to model non-precise statistical data (see, e.g., Denœux, 2000; Diamond, 1988; González-Rodríguez et al., 2012 and Viertl, 2011). This is one of the most successful uses of fuzzy sets in the area. Fuzzy data arise as a generalization of interval data, or in higher dimensions, of set-valued data, in order to formalize the blurriness of the boundaries of elements that are essentially imprecise, in the sense of not being constrained to a single point in $\mathbb{R}^n$ (see, e.g., Ferraro et al., 2010). Set-valued data have been used in statistics and econometrics through the theory of random sets to deal with convex particles, cell shapes, partial identification and multivariate risk measures (see, e.g., Carleos et al., 2014; Choirat and Seri, 2014; Molchanov and Cascos, 2016 and Molchanov and Molinari, 2018). In particular, interval data are used to represent fluctuations or ranges, grouped data, censoring, rounding, or to deal with partial identification (see, e.g., Beresteanu and Molinari, 2008; Cameron and Huppert, 1989; De Carvalho, 2007; Manski and Molinari, 2010 and Ramos-Guajardo and González-Rodríguez, 2013).

Fuzziness is also used to represent imprecision or vagueness, e.g., in the context of linguistic variables (Zadeh, 1975) or expert ratings (González-Rodríguez et al., 2012). In practice, most of the fuzzy subsets of $\mathbb{R}$ are a generalization of intervals, both conceptually and methodologically (see, e.g., Ferraro et al., 2010). Depending on the aim, fuzzy, set-valued or interval data can either be considered the (complex) statistical data of interest per-se, or the imperfect observable outcomes of some underlying precise un-observed statistical data of interest (compare, e.g., Manski and Tamer, 2002 and Ramos-Guajardo and González-Rodríguez, 2013). The latter is sometimes called epistemic approach, while the former is called ontic approach (see, e.g., Colubi and González-Rodríguez, 2015).

Among the various existing frameworks to handle fuzzy data associated with a standard random experiment, those with more impact relate to the concept of fuzzy random variable introduced by Puri and Ralescu (1986)). This concept has generated a full body of statistical literature (see, e.g. Colubi et al., 2011; Körner, 2000 and Näther, 1997). These statistical tools have been frequently used in areas such as insurance, blind testing or psychology, to name but a few (see, e.g. Coppi et al., 2006b; Ramos-Guajardo et al., 2019 and Shapiro, 2009). For instance, the space of fuzzy sets has been proposed as a rich alternative to Likert scales in order to capture more information in intrinsically imprecise data, like evaluations, medical diagnoses or quality ratings (see, e.g., González-Rodríguez et al., 2012 and Lubiano et al., 2016).

The aim is to discuss several key approaches merging fuzziness and stochastic uncertainty in econometrics and statistics and clarify their potential use for future research in this area. The selected topics cover examples where the transition between two states can be made gradual, e.g., fuzzy clustering and fuzzy regression discontinuity designs (Section 2), the management of non-precise statistical data (Section 3), and the difference between fuzzy regression and regression with (fuzzy) set-valued data (Section 4). Some remarks and a summary of relevant open problems will be put together to conclude (Section 5).

2. Fuzzy transitions

Formally, a traditional (sub-)set A of a reference class E is characterized by a function $u_A : E \to \{0, 1\}$ so that $u_A(x) = 1$ if $x \in A$ and $u_A(x) = 0$ if $x \notin A$. In contrast, a fuzzy set A of a reference class E is characterized by a function $u_A : E \to [0, 1]$ where $u_A(s)$ is interpreted as the membership degree of s to the fuzzy set A, i.e., the boundaries of the set are blurred: there is a fuzzy transition between belonging or not belonging to the set (Zadeh, 1965).

This section describes two instances of gradual transitions useful in econometrics and statistics: fuzzy clustering and fuzzy regression discontinuity designs. Note that both of them are fuzzy approaches for non-fuzzy data.

2.1. Fuzzy clustering

Traditional (hard) cluster analysis aims to determine k homogeneous and separate groups or clusters from a set of n objects according to a given dissimilarity measure based on p observed variables $\{X_i\}_{i=i}^{p}$ (see Henning et al., 2015 for an in-depth review on hard clustering). However, there are situations in which some objects have intermediate features among clusters and, as a consequence, cannot be clearly assigned. In such cases, a classical hard clustering approach leads to an unnatural classification since each data point is restricted to belonging to just one cluster, being the points similar within each group and different from those in other groups (Jain et al., 1999). Different soft-clustering approaches can be used to overcome this drawback by allowing the objects to belong to more than one cluster, as it is the case of the fuzzy clustering described below. For completeness, other soft-clustering approaches are briefly described at the end of the section.

One of the best-known contributions of the fuzzy set theory to exploratory data analysis is fuzzy clustering. Fuzzy clustering relaxes the hard clustering approach by considering that the points can belong to each group with a certain membership degree between 0 and 1. In this way, the transition between the clusters becomes fuzzy.

As for hard clustering, there are many ways to obtain fuzzy clusters depending on, e.g., the similarity criterion, the existence of constraints, the shape of the clusters, or the nature of the data set, which can be Euclidean or non-Euclidean, complex, contaminated or high-dimensional, to name but a few (see, e.g., D'Urso and Giordani, 2006; Ferraro et al., 2021; Hathaway and Bezdek, 1993; Huang et al., 2012 and Krishnapuram and Keller, 1993). Economic-related applications can be found, e.g., in production, accounting or hedge funds analysis (see, e.g., Gibson and Gyger, 2007; Qu and Zhang, 2010 and Yang et al., 2006).

The basic fuzzy $k-$means (or fuzzy $c-$means) problem can be formalized as follows (Bezdek, 1981). The aim is to find the centroids c_j that determine the clusters C_j, and the membership u_{ij} of each data point x_i to C_j, for all $i = 1, \dots, n$ and $j = 1, \dots k$, so that the dissimilarity within each cluster is minimized according to a given criterion. At first, the dissimilarity measure was based on the Euclidean distance, but it can be any d defined on the space where the data points live (originally, $\mathbb{R}$). The common fuzzy $k-$means considers that the sum of memberships of each point to all the clusters must be 1, that is, $\sum_{j=1}^{k} u_{ij} = 1$ for all $i = \{1, \dots, n\}$.

Moreover, the approach depends on a parameter $m > 1$ called fuzzifier that has a smoother effect and controls how much the clusters overlap. The usual value of the fuzzifier is $m = 2$ (see, e.g., Klawonn and Höppner, 2003). Given a set of centroids (non necessarily the optimal ones), the membership degree of each point x_i is computed in terms of the dissimilarity measure as follows:

$$u_{ij} = \frac{1}{\sum_{l=1}^{k} \left(\frac{d(x_i, c_j)^2}{d(x_i, c_l)^2} \right)^{1/(m-1)}}.$$

And given the membership degrees u_{ij} for $i = \{1, \dots, n\}$ and $j = \{1, \dots, k\}$, the centroids of the cluster C_j are computed as

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m}.$$

With this notation, the problem is to find the centroids $\{c_1^*, \dots, c_k^*\}$ such that

$$\{c_1^*, \dots, c_k^*\} = arg\,min_{\{c_1, \dots c_k\}} \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m d(x_i, c_j)^2.$$

As a generalization of the $k-$means algorithm, the fuzzy $k-$means algorithm is recursive, and updates the solution starting from an arbitrary set of centres. An extensive review of some of the most relevant fuzzy clustering methods based on the classical fuzzy $k-$means can be found in Ferraro (2021).

The R package 'fclust: Fuzzy Clustering' (Ferraro et al., 2019) implements algorithms for fuzzy clustering, cluster validity indices and plots for cluster validity and visualizing fuzzy clustering results. An illustrative example of fuzzy clustering can be found in Ferraro et al. (2019). There, the fuzzy k-means method with $m = 1.2$ has been applied to the NBA dataset available in the R package 'fclust'. The dataset contains variables on 30 NBA teams for the regular season 2017-2018, such as field goal percentage, free throw percentage and offensive rebounds. Two clusters were found according to the fuzzy silhouette index. From an interpretability point of view, the best teams are more related to the first cluster and the worst teams to the second one.

In addition to fuzzy clustering, other approaches can be considered "soft". Namely, possibilistic clustering (Krishnapuram and Keller, 1993) also assigns membership degrees to different clusters, but it aims at relaxing the unit-sum constraints of the membership degrees by adding a penalization term. The membership values can be interpreted as degrees of possibility of the points to belong to the clusters. Thus, data points that are far from all the prototypes are assigned to the different clusters with membership degrees closer to 0. Rough clustering (Pawlak, 1982; 1992) is another soft approach that is based on the concept of rough set, which is characterized by lower and upper approximations, so that each object can be assigned to a lower approximation or to two or more upper approximations or boundary regions according to the distances between each object and each prototype.

Finally, the model-based clustering can also be considered a soft method since it produces a soft partition of the units. The posterior probability of belonging to a cluster plays a role comparable to that of the membership degree. However, both approaches are conceptually different; namely, a membership degree cannot be understood as a probability measure because there is no random generation process previously assumed and, conversely, the posterior probability is neither associated with the fuzziness of the partition. The most used approach in this framework is the mixture of Gaussian densities (see, for instance, Fraley and Raftery, 2002; Kasa and Rajan, 2022 and McLachlan and Peel, 2000), although other extensions involving different parametric distributions have been proposed as, for instance, a mixture of t-distributions (see, e.g., Murray and Browne, 2017 and Peel and McLachlan, 2000) or other non-Gaussian multivariate data (Sahin and Czado, 2022). An expanded review of the latest advances on soft methods can be found in Ferraro and Giordani (2020).

2.2. Fuzzy regression discontinuity designs

Regression Discontinuity Designs (RDDs) are used to assess interventions where the eligibility to take part in an event, or receive a given treatment, or not is determined by a cutoff point $c \in \mathbb{R}$ once the individuals are ranked according to an eligibility criterion X (see, e.g., Imbens and Lemieux, 2008; Lee and Lemieux, 2010 and Thistlethwaite and Campbell, 1960).

An RDD assumes that none of the individuals i below the cutoff ($X_i < c$) receives the treatment ($W_i = 0$), while all the individuals above the cutoff ($X_i \geq c$) receive the treatment ($W_i = 1$), that is, there are no cross-overs or no-shows. Individuals just around the cutoff are analyzed to assess the potential (non-observed) causal impact of treatment, denoted $Y_i(1)$ and $Y_i(0)$, being the observed outcome $Y_i = W_iY_i(1) + (1 - W_i)Y_i(0)$. RDDs are sometimes described as local randomized experiments, in the sense that the relationship between the outcome and the eligibility criterion around the cutoff is assumed to be continuous (i.e., there are no confounders), which implies randomization.

The usual parameter of interest is the average causal effect at the cutoff, that is

$$\tau = E(Y_i(1) - Y_i(0)|X = c).$$

Since $Y_i(1) - Y_i(0)$ is not observable, under the RDD continuity conditions, τ can be equivalently expressed as

$$\tau = \lim_{x \downarrow c} E(Y_i|X_i = x) - \lim_{x \uparrow c} E(Y_i|X_i = x),$$

which can be estimated, e.g., with nonparametric techniques (Hahn et al., 2001). Further improved inferences are also available (see, e.g., Calonico et al., 2014).

The lack of cross-overs or no-shows is not always realistic, and while there may be a clear threshold determining who is eligible for the treatment, a small percentage of those below, but near the threshold, may receive the treatment, while a small percentage of those above, but near the threshold, may not receive the treatment. Since the cutoff is not sharp anymore, this RDD is called Fuzzy Regression Discontinuity Design, or FRDD for short (see, e.g., Hahn et al., 2001 and Imbens and Lemieux, 2008, which deals with both sharp and fuzzy RDD, or Bertanha and Imbens, 2020 and Dong, 2018, that are related to the FRDD approach).

In this context, however, the blurriness is actually related to the conditional probabilities $P(W = w|X = x)$, with $w = 0, 1$, which differs from most cases when fuzziness is considered in Zadeh's sense. FRDD are usually handled with a similar methodology as RDD by considering the subpopulation of compliers, i.e., those individuals who would not take the treatment if the cutoff were below X_i and would take it above X_i. In this case, the parameter of interest is the average causal effect for compliers at $x = c$, that is,

$$\tau_f = E(Y_i(1) - Y_i(0)|i \text{ complier}, X_i = c).$$

This parameter can be expressed in a more tractable way under mild conditions by scaling the jump of outcome discontinuity by the jump of the treatment discontinuity

$$\tau_f = \frac{\lim_{x \downarrow c} E(Y_i|X = x) - \lim_{x \uparrow c} E(Y_i|X_i = x)}{\lim_{x \downarrow c} E(W_i|X_i = x) - \lim_{x \uparrow c} E(W_i|X_i = x)}.$$

Thus, the same techniques used to estimate τ can be applied to estimate τ_f and no instrument from the standard fuzzy theory based on Zadeh's concept is required. This approach has been frequently applied in econometrics (see, e.g., Basten and Betz, 2013 and Cardella and Depew, 2014).

The R package 'rdd: Regression Discontinuity Estimation' (Dimmery, 2016) provides the tools to undertake estimation in both sharp and fuzzy Regression Discontinuity Designs. Other works involving further developments and applications related to the FRDD approach can be found in Arai et al. (2021); Cattaneo et al. (2016) and Choi and Lee (2018).

As an illustrative example of the treatment assignment in FRDD, the clinical study on the initiation of HIV early antiretroviral treatment in South Africa developed in Bor et al. (2014) can be considered. The study population included the patients who had a first CD4 lymphocyte count between 1 January 2007 and 11 August 2011 and were under surveillance. The previous guidelines recommended treating all people with initial CD4 counts of less than 200 cells/mm^3 and, according to this, the proportion receiving treatment declined noticeably at this threshold. Applying the FRDD approach, the probability of receiving treatment was notably higher for patients with CD4 counts below this threshold. Even so, several people below that threshold did not receive treatment, whereas several above the threshold did, likely because of other clinical symptoms. Such crossover, relative to the assignment threshold, motivates the use of fuzzy RDD. In contrast, when employing sharp RDD, the probability of receiving early antiretroviral treatment would have been exactly 1 above the threshold and 0 below it.

3. Interval and (fuzzy) set-valued data

Non-precise data are frequently modelled through (fuzzy) sets verifying some convenient conditions such as boundedness or convexity. This section will introduce first interval data and after fuzzy data following the so-called ontic approach (Colubi and González-Rodríguez, 2015). An R package containing all the methods described in Section 3 can be obtained from the author upon request.

3.1. Random intervals: Interval data

Let $\mathcal{K}_c(\mathbb{R}) = \{[a, b] : a, b \in \mathbb{R}, a < b\}$ be the space of bounded and closed intervals. Any $A \in \mathcal{K}_c(\mathbb{R})$ can be characterized by means of the mid-point and the spread, through the $t-$vector $t_A = (\text{mid}\, A, \text{spr}\, A) \in \mathbb{R}^2$ (with $\text{spr}\, A \geq 0$). This representation allows us to separate two important concepts: location (related to the mid-point) and imprecision (related to the spread).

The natural interval arithmetic on $\mathcal{K}_c(\mathbb{R})$ is defined in terms of the Minkowski addition and the product by a scalar, $A+\lambda B=\{a+\lambda b : a\in A, b\in B\}$ for all $A,B\in\mathcal{K}_c(\mathbb{R})$ and $\lambda\in\mathbb{R}$. It verifies that

$$t_{A+\lambda B}=(\operatorname{mid}A+\lambda\operatorname{mid}B, \operatorname{spr}A+|\lambda|\operatorname{spr}B).$$

Of course, this arithmetic does not agree with the natural arithmetic on $\mathbb{R}^2$, and the space $(\mathcal{K}_c(\mathbb{R}),+,\cdot)$ is not linear, but semilinear (cone structure). The natural difference is called Hukuhara difference, that is, $A-_H B=C$, with $C\in\mathcal{K}_c(\mathbb{R})$ so that $A=B+C$, and it exists if spr $B\le$ spr A.

Metrics in $\mathcal{K}_c(\mathbb{R})$ can be inherited from the general expressions of the metrics in $\mathbb{R}^2$. For instance, the well-known Hausdorff distance can be expressed as the L_1-metric, i.e.,

$$d_H(A,B)=d_1(t_A,t_B)=|\operatorname{mid}A-\operatorname{mid}B|+|\operatorname{mid}A-\operatorname{spr}B|.$$

Weights of the distance between mid-points (location) and spreads (imprecision) can be considered. The general L^2-metrics are very convenient in statistics. A general (weighted) L^2-distance can be defined as follows,

$$d_\tau^2(A,B)=d_{2,\tau}^2(t_A,t_B)=(1-\tau)(\operatorname{mid}A-\operatorname{mid}B)^2+\tau(\operatorname{spr}A-\operatorname{spr}B)^2,$$

(see, e.g., Sinova et al., 2012).

The notions of random interval and expected value can equivalently be inherited from the ones in $\mathbb{R}^2$ through the $t-$vector, from the general ones in metric spaces, or as a particular case of the theory of compact and convex random sets (see, e.g., Molchanov and Molinari, 2018). Thus, given a probability space $(\Omega,\mathcal{A},P)$ a mapping $X:\Omega\to\mathcal{K}_c(\mathbb{R})$ is random interval if $t_X=(\operatorname{mid}X,\operatorname{spr}X)$ is an $\mathbb{R}^2-$random vector (with $\operatorname{spr}X\ge_{a.s.}0$), and $E(X)$ is the interval so that $t_{E(X)}=E(t_X)$. The variance can also be defined through the Frechet approach, or equivalently, as $\sigma_X^2=(1-\tau)Var(\operatorname{mid}X)+\tau Var(\operatorname{spr}X)$. Since all these concepts depend only on the metric and the semi-linear structure, all the interpretation and usual properties hold. This is not the case for the covariance of two random intervals X and Y that can be defined only in terms of the $t-$vector in $\mathbb{R}^2$ or, equivalently, as

$$Cov(X,Y)=(1-\tau)Cov(\operatorname{mid}X,\operatorname{mid}Y)+\tau Cov(\operatorname{spr}X,\operatorname{spr}Y).$$

The covariance does have the classical meaning and properties since, for instance, X can be non-degenerated and equal to $-X$. However, it can be connected with the strength of “linear relationships” (see, e.g., Blanco-Fernández et al., 2011).

In order to handle statistical interval data and derive inference, e.g., on the expected value, asymptotic techniques are usually considered. One of the reasons for such a consideration is that no normal distribution (but a degenerated one on the mid-points) lives in $\mathbb{R}\times[0,+\infty)$, due to the boundedness restriction of the spread, and there is no parametric model for random intervals widely used in applications. Any model in $\mathbb{R}\times[0,+\infty)$ is a candidate, e.g., $(\mathcal{N}(\mu,\sigma),\chi_l^2)$ and this is usually applied in simulations, but methods not relying on parametric distributions are preferred (see, e.g., Gil et al., 2007; Ramos-Guajardo et al., 2020).

Asymptotic results are supported by general large sample probabilistic results, such as the Strong Law of Large Numbers, or the (distance-based) Central Limit Theorem particularized to $\mathcal{K}_c(\mathbb{R})$ (see, e.g. Molchanov and Molinari, 2018). Asymptotic/bootstrap confidence regions and hypothesis tests can be developed similarly to the real case.

For the confidence regions, theory developed for the (non-parametric) confidence estimation of the mean of non-negative random variables can be used in combination with the standard theory to estimate the mean of the $t-$vector in $\mathbb{R}\times[0,+\infty)$. Alternatively, the following method based on bootstrapping valid for the general case (not relying on linearity or ordering) can be applied. Note that in $\mathcal{K}(\mathbb{R})$ ‘confidence intervals’ cannot be expressed as the sample mean plus/minus a given quantity depending on the (estimated) variability and the sample size. That is why the *confidence regions* are not specified by upper and lower bounds, but expressed in terms of distances.

Confidence Regions (CR) can be defined as balls centred on the sample mean with radius determined via bootstrapping. For the case of the mean, if $(1-\beta)\in(0,1)$ be confidence level, then the $CR_\beta-$confidence ball will be

$$CR_\beta=B\big(\overline{X},\delta\big)=\Big\{A\in\mathcal{K}_c(\mathbb{R})|d_\tau(A,\overline{X})\le\delta\Big\},$$

where the radius δ verifies the coverage condition

$$P\Big(\mu\in B\big(\overline{X},\delta\big)\Big)=P\Big(d_\tau\big(\mu,\overline{X}\big)\le\delta\Big)=1-\beta.$$

Thus, δ should be the $(1-\beta)$-quantile of the distribution of $d_\tau(\mu,\overline{X})$. In practice, δ can be approximated by the corresponding bootstrap $(1-\beta)$-quantile. This approach is theoretically well-supported and provides adequate empirical results for moderate or large sample sizes, but not for small sample sizes. As usual, a way to improve the results is to consider the variability. The standardized confidence interval is slightly conservative, but it can be applied even for small sample sizes with good results (e.g., $n=10$).

Hypothesis testing approaches can also be developed (see, e.g., Blanco-Fernández and González-Rodríguez, 2016 and Blanco-Fernández and Warren-Liao, 2015). Consider, for example, the one-sample test. Let X be an integrable random interval and $A\in\mathcal{K}_c(\mathbb{R})$. The aim is to test $H_0:E(X)=A$ against $H_1:E(X)\neq A$. This can be done through the t-vector by applying techniques in $\mathbb{R}\times(0,+\infty]$. However, for the sake of the generalization, and in order to take into account the metric structure, the test can be expressed in terms of distances to avoid using (not well-defined) differences, like for the confidence regions. Thus, the null hypothesis can be written as $H_0:d_\tau(E(X),A)=0$, and the alternative hypothesis as $H_1 : d_\tau(E(X), A) > 0$. Asymptotic techniques are supported by the Central Limit Theorem (CLT). In the same way, the bootstrapped CLT supports bootstrapping. The one-sample statistics can be defined as $T = \left[d_\tau\left(\overline{X}, A\right)\right]^2/\widehat{S}^2$, where $\widehat{S}^2$ is the quasi-variance. Similarly, bootstrap two-sample and ANOVA tests can be established (Nakama et al., 2010). In practice, they show the usual empirical behaviour in the Behrens-Fisher problem, which can be lessened as for the real case.

Inference on the (real-valued) variance has been developed for both one and $k-$samples analogously to the real case. For instance, the Levene-type statistics for k random intervals $X_1 \ldots X_k$ can be defined

$$R_n^k = \frac{\sum_{i=1}^k n_i\left(\widehat{\sigma}^2_{X_i} - \widehat{\sigma}^2\right)^2}{\sum_{i=1}^k \widehat{\sigma}^2_{(d_\tau(X_i,\overline{X_i \cdot}))^2}},$$

where $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^k n_i\widehat{\sigma}^2_{X_i}$ and $\widehat{\sigma}_{X_i}$ is the sample variance of X_i (Ramos-Guajardo and Lubiano, 2012).

Hypothesis tests that regard concepts inherently related to sets are of particular interest, such as inclusion or overlapping (see, e.g., Beresteanu and Molinari, 2008 and Ramos-Guajardo et al., 2020. Among these problems, partial inclusion can be considered a first attempt to formalize statistically “fuzzy” hypothesis, and it is especially relevant for its generality (Ramos-Guajardo et al., 2014).

An example to illustrate the applicability of the hypothesis testing theory for the variance of a random interval has been provided in Ramos-Guajardo and González-Rodríguez (2013). The purpose of the study was to determine if the variability of the tidal fluctuation in Gijón (Spain) changed in 2010 with respect to 2009. Knowing that the variability of tides fluctuation in 2009 was.0593, the aim was to check if that of 2010 can be considered the same with the information of the tidal fluctuation of 50 days of 2010 by using the proposed test for the variance. Since the null hypothesis was not rejected at the usual significance levels, the variability of the tidal fluctuations in Gijón in 2010 cannot be considered different from the one in 2009.

3.2. Random sets: Set-valued data

The extension of the concept of interval to higher dimensions, even for the fuzzy case, has been frequently focused on preserving closeness and, more conveniently, convexity. Thus, the theory of (convex and bounded) closed sets of $\mathbb{R}^p$, usually based on the Minkowski addition and the Hausdorff distance, has been well-developed for statistical/econometric purposes (see, e.g., Choirat and Seri, 2014; Matheron, 2018; Molchanov and Cascos, 2016; Molchanov and Molinari, 2014 and Simó et al., 2004). The support function allows us to embed the space of convex compact sets into a linear functional space by preserving important metric and arithmetic properties, which has frequently simplified the study of set-valued random elements (see, e.g., Adusumilli and Otsu, 2017 and Beresteanu and Molinari, 2008).

Recently, a more general alternative, based on star-shaped sets, a directional Minkowski addition, and L^2-type metrics, has shown to be convenient from a formal and a practical point of view (González-Rodríguez et al., 2018). Star-shaped sets relax the condition of convexity to directional convexity from a given point (see, e.g., Klain, 1997). They are useful to represent, e.g., (almost) convex particles or cell shapes (Carleos et al., 2014; Choirat and Seri, 2014). The main advantages of this approach are:

1. It generalizes the concept of mid/location and spread/imprecision by parametrizing the sets by a crisp centre and a radial function;
2. The propagation of the uncertainty is made directionally, which avoids the general dilation produced by the Minkowski addition;
3. The cone where the space is embedded is explicitly known, in contrast to what happens for the support function, so it allows us to fully exploit the rich statistical theory in Hilbert spaces.

The aim is to handle a set A through a centre $c_A \in A$ and the polar function from ρ_A. Let $A^c = A - c_A$ be the associated set ‘centred at 0’. A closed set $A \subset \mathbb{R}^p$ with $0 \in A$ is a star-shaped set w.r.t. 0 if for any $a \in A$, $\gamma a \in A$ for all $\gamma \in [0, 1]$, that is, if the segment joining 0 and a fully belongs to A. Thus,

$$A = \left\{\gamma\,\rho_A(u) \mid \gamma \in [0, 1],\ u \in \mathbb{S}^{p-1}\right\},$$

where $\mathbb{S}^{p-1}$ is the unit sphere and $\rho_A(u) = \sup\{\gamma \in [0, \infty) \mid \gamma u \in A\}$ for all $u \in \mathbb{S}^{p-1}$ is the polar function. In order to avoid ill-definitions, and as usual when handling functional data, ρ_A will be assumed to be square-integrable, that is, $\rho_A \in L^2(\mathbb{S}^{p-1}, \vartheta_p)$, where ϑ_p is the Lebesgue measure over the unit sphere. The space of centred star-shaped sets is denoted by $\mathbb{X}_0$. The space of centred star-shaped sets with square-integrable boundary will be denoted $\mathbb{X}_0^2$.

Generally, $A \subset \mathbb{R}^p$ is a star-shaped set, i.e. $A \in \mathbb{X}$, if there exists $x \in A$ so that $A - x = \{a - x \mid a \in A\} \in \mathbb{X}_0$. Therefore, the star-shaped sets will be characterized by (c_A, ρ_{A^c}), i.e.,

$$A = \left\{c_A + \gamma\,\rho_{A^c}(u) \mid \gamma \in [0, 1],\ u \in \mathbb{S}^{p-1}\right\}.$$

By defining the interval $I_u(A^c) = [-\rho_{A^c}(-u), \rho_{A^c}(u)]$ for each $u \in \mathbb{S}^{p-1}$, a directional Minkowski addition can be defined so that $I_u(A^c + B^c) = I_u(A^c) + I_u(B^c)$ for all $A, B \in \mathbb{X}$. This is equivalent to consider

$$A^c + B^c = \left\{\gamma\,(\rho_{A^c}(u) + \rho_{B^c}(u)) \mid \gamma \in [0, 1],\ u \in \mathbb{S}^{p-1}\right\},$$

and analogously for the product by a scalar,

$$\lambda \cdot A^c = \left\{ \gamma(|\lambda| \rho_{A^c}(\mathrm{sign}(\lambda)u)) \mid \gamma \in [0,1],\ u \in \mathbb{S}^{p-1} \right\}.$$

for all $\lambda \in R$. In this way,

$$A + B = (c_A, \rho_{A^c}) + (c_B, \rho_{B^c}) = (c_A + c_B, \rho_{A^c} + \rho_{B^c}),$$

and $\lambda(c_A, \rho_{A^c}(\cdot)) = (\lambda c_A, |\lambda| \rho_{A^c}(sign(\lambda) \cdot))$. This arithmetic propagates the imprecision directionally, and not as a general dilation as happens for the usual Minkowski addition, which can be convenient in many applications.

Let $H = \mathbb{R}^p \times L^2(\mathbb{S}^{p-1}, \vartheta_p)$ and $\tau \in (0,1)$. For any $(x_1, f_1), (x_2, f_2) \in H$, inspired on the mid-spread metric for intervals, the $\tau-$inner product is defined as

$$\left\langle (x_1, f_1), (x_2, f_2) \right\rangle_\tau = (1-\tau) x_1^t \cdot x_2 + \tau [f_1, f_2]_2,$$

where $\left[f, g\right]_2 = \int_{\mathbb{S}^{p-1}} f_1(u) f_2(u) \vartheta_p(du)$. The Hilbert space $(H, \left\langle \cdot, \cdot \right\rangle_\tau)$ is separable. Let $|| \cdot ||_\tau$ denote its induced norm. The space of star-shaped sets is embedded in H as follows. Let $\Gamma : \mathbb{R}^p \times \mathbb{X}_0 \to H$ so that $\Gamma(c, A) = (c, \rho_A)$. Let $A, B \in \mathbb{X}_0^2$, $x, y \in \mathbb{R}^p$, the metric is defined as

$$d_\tau^2((x, A), (y, B)) = ||\Gamma(x, A) - \Gamma(y, B)||_\tau^2$$

$$= (1-\tau)(x-y)^t \cdot (x-y) + \tau \int_{\mathbb{S}^{p-1}} (\rho_A(u) - \rho_B(u))^2 \vartheta_p(du).$$

As usual when dealing with L^2 spaces, equivalence classes are considered. Thus, A and B are in the same equivalence class if $\rho_A = \rho_B$ a.s.$-\vartheta_p$.

The mapping $\Gamma : \mathbb{R}^p \times \mathbb{X}_0^2 \to H$ is an isometry preserving the arithmetic, and $\Gamma(\mathbb{R}^p \times \mathbb{X}_0^2) = \mathbb{R}^p \times L^2(\mathbb{S}^{p-1}, \vartheta_p)^+$ is a closed convex cone. The space $(\mathbb{R}^p \times \mathbb{X}_0^2, d_\tau)$ is complete and separable, and d_τ is invariant under rigid motions. One of the main advantages of this embedding w.r.t. the one induced by the support function is that in the first situation, the arriving cone is perfectly identified, and it is trivial to determine the set associated with any element of the cone, which is not easy in the case of the support function.

A missing point until now is the determination of the centre c_A for any $A \in \mathbb{X}^2$, where $\mathbb{X}^2$ denotes the space of star-shaped sets with square-integrable boundary. The idea is to define as centre a point of the kernel, $ker(A) = \{x \in A \mid A - x \in \mathbb{X}_0^2\}$, which is a non-empty, convex and closed set. The problem is that the $ker(A)$ may be affected by noise, and be different for two elements in the same equivalence class. Thus, a previous ‘cleaning’ is requested. Let $A \in \mathbb{X}^2$ with bounded $ker(A)$, so that it is meaningful to search for its centre, otherwise any point would serve. Thus, $ker(A) \in \mathcal{K}_c(\mathbb{R}^p)$. Consider the subspace generated by the linear span of A:

$$span(A) = \left\{ \sum_{i=1}^k \lambda_i a_i \mid k \in \mathbb{N}, a_i \in A,\ \lambda_i \in \mathbb{R} \right\},$$

and let $int(A)$ be the interior of A within its span. If $int(A) \neq \emptyset$, then $c_A =$ centre of mass of $ker(cl(int(A)))$. If $int(A) = \emptyset$, let $\lambda_{ker(A)}$ be Lebesgue measure on $ker(A)$, then $c_A =$ centre of mass w.r.t. $\lambda_{ker(A)}$ of $ker(A)$. In this way, c_A is robust against negligible noise affecting the kernel.

Fig. 1 shows an example of a star-shaped set A in $\mathbb{R}^2$ and its centred version A^c. Regarding the main probabilistic concepts and results, let (Ω, A, P) be a probability space. A mapping $X : \Omega \to \mathbb{R}^p \times \mathbb{X}_0^2$ is called Random L_2-Star-shaped Set (R2S) if $\Gamma \circ X$ is an $H-$valued random element. If X is an R2S so that $E(d_\tau(X, 0)) < \infty$, then the expected value $E(X)$ is defined as the element in $\mathbb{R}^p \times \mathbb{X}_0^2$ so that $\Gamma(E(X)) = E(\Gamma(X))$. In this case both Bochner and Pettis expectations agree.

Let X, Y be an R2Ss so that $E(d_\tau^2(X, 0)) < \infty$ and $E(d_\tau^2(Y, 0)) < \infty$, then the (scalar) variance is defined as

$$Var(X) = E\left(d_\tau^2(X, E(X))\right) = E(||\Gamma(X) - E(\Gamma(X))||_\tau^2\Big).$$

Similarly, the (scalar) covariance can be defined as

$$Cov(X, Y) = E\Big(\left\langle \Gamma(X) - E(\Gamma(X)), \Gamma(Y) - E(\Gamma(Y)) \right\rangle_\tau\Big).$$

Moreover, the covariance operator $C_{\Gamma(X)} : H \to H$ is defined as

$$C_{\Gamma(X)}(x) = E\Big(\left\langle (\Gamma(X) - E(\Gamma(X))), x \right\rangle_\tau (\Gamma(X) - E(\Gamma(X)))\Big)$$

for all $x \in H$. Thus, $Cov\Big(\left\langle \Gamma(X), x \right\rangle_\tau, \left\langle \Gamma(X), y \right\rangle_\tau\Big) = \left\langle C_{\Gamma(X)}(x), y \right\rangle_\tau$ for all $x, y \in H$.

The Central Limit Theorem (CLT), and other asymptotic results, can be derived from the well-known CLT in Hilbert Spaces (see, e.g., Araujo and Gine, 1980). Thus, let $X_1, \ldots, X_n$ be an i.i.d. sequence of R2Ss so that $E(d_\tau^2(X_1, 0)) < \infty$, then

- $n^{1/2}\Big(\Gamma\big(\frac{1}{n}\sum_{i=1}^n X_i\big) - E(\Gamma(X_1))\Big) \to Z_{\Gamma(X_1)}$ weakly in H,
- $n\, d_\tau^2\big(\frac{1}{n}\sum_{i=1}^n X_i, E(X_1)\big)_\tau^2 \to \|Z_{\Gamma(X_1)}\|_\tau^2$ weakly in $\mathbb{R}$,

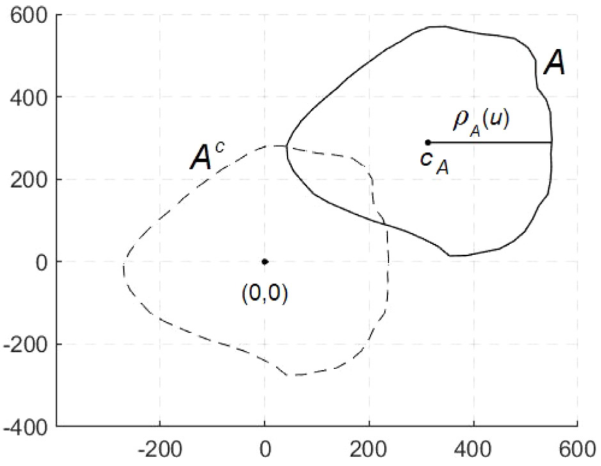

Fig. 1. Star-shaped set A in $\mathbb{R}^2$, its center c_A, its associated radial function for $u=(1,0)$, and the corresponding centred set A^c.

where $Z_{\Gamma(X_1)}$ is a centred Gaussian $H-$valued random element with $C_Z = C_{\Gamma(X_1)}$. As a result, the same inferential techniques based on distances described for random intervals, such as confidence interval for the means, ANOVA or homoscedasticity tests, can be stated for random sets in higher dimensions.

As an illustrative example involving star-shaped data, the study in Ferraro et al. (2022) can be considered. The dataset refers to 41 Cantabrian coast stones (Spain), namely, 19 river stones and 22 sea stones. The shape and the roundness were measured. The first feature was parameterised through star-shaped random sets. After applying a fuzzy k-means approach for such type of data 2 clusters were obtained. The results were consistent with what was expected, since most of the sea stones are rounder than the river ones and therefore are more related to one of the clusters.

3.3. Random fuzzy sets: Fuzzy data

Fuzzy data have been long handled as realizations of random fuzzy sets, or fuzzy random variables, whose $\alpha-$levels are convex compact random sets. That is, the theory of convex compact random sets has been extended level-wise to cover the fuzzy sets (see, e.g., González-Rodríguez et al., 2012; Körner, 2000; Näther, 1997 and Puri and Ralescu, 1986). From an ontic point of view, fuzzy data can be treated as values belonging to a metric space endowed with a semilinear structure, so it is possible to develop different statistical techniques for metric spaces. Dealing with the whole (fuzzy) sets instead of summarizing them by, for instance, their central points, allows capturing the intrinsic imprecision inherent to certain kinds of data, such as ranges or perceptions (see, e.g., Ferraro et al., 2010; Fischer et al., 2016 and González-Rodríguez et al., 2012). Specifically, in Fischer et al. (2016) it is shown that considering the daily range of process data through intervals in a financial problem involving regression produces better forecasts than considering only the mean prices.

Following the discussion in Section 3.2, the general alternative based on random star-shaped sets can be further extended level-wise. This section will introduce the basic concepts and results. For more details, see González-Rodríguez, 2020.

Fuzzy sets will be characterized by a centre, or location, in $\mathbb{R}^p$ and radial function defined level-wise for the corresponding fuzzy set centred on 0. The space of fuzzy star-shaped sets (w.r.t. 0) is denoted by

$$\mathbb{F}_0(\mathbb{R}^p) = \{A : \mathbb{R}^p \to [0,1] \,|\, A_\alpha \in \mathbb{X}_0 \ \forall \alpha \in (0,1]\},$$

where $A_\alpha = \{x \in \mathbb{R}^p \,|\, A(x) \geq \alpha\}$ for all $\alpha > 0$. The arithmetic is extended level-wise, that is, $(A+\gamma B)_\alpha = A_\alpha + \gamma B_\alpha$ for all $\alpha \in (0,1]$. The polar function is defined as $\rho_A : \mathbb{S}^{p-1} \times (0,1] \to \mathbb{R}^+$ so that $\rho_A(u,\alpha) = \rho_{A_\alpha}(u)$ for all $u \in \mathbb{S}^{p-1}, \alpha \in (0,1]$. An example of a fuzzy set A in $\mathbb{R}$ is provided in Fig. 2.

Considering square-integrable polar functions, and

$$\mathbb{F}_0^2(\mathbb{R}^p) = \left\{A \in \mathbb{F}_0(\mathbb{R}^p) \,|\, \rho_A \in L^2(\mathbb{S}^{p-1} \times (0,1], \vartheta_p \times \lambda)\right\},$$

the embedding is $\Gamma : \mathbb{R}^p \times \mathbb{F}_0^2(\mathbb{R}^p) \to H$ so that $\Gamma(c,A) = (c,\rho_A)$, where $H = \mathbb{R}^p \times L^2(\mathbb{S}^{p-1} \times (0,1], \vartheta_p \times \lambda)$. The metric on the separable Hilbert space can also be extended level-wise. Let φ be an L^2 density with support [0,1] (in practice $\varphi = 1$), and $\tau \in (0,1)$, let $(x_1,f_1),(x_2,f_2) \in H$, then

$$\left\langle (x_1,f_1),(x_2,f_2)\right\rangle_{\tau,\varphi} = (1-\tau)x_1^t \cdot x_2 + \tau \int_0^1 [f_1(\cdot,\alpha), f_2(\cdot,\alpha)]_2 \varphi(\alpha)\lambda(d\alpha),$$

where $\left[f,g\right]_2 = \int_{\mathbb{S}^{p-1}} f_1(u) f_2(u) \vartheta_p(du)$. The induced norm will be denoted by $||\cdot||_{\tau,\varphi}$. In this way, the metric on $\mathbb{R}^p \times \mathbb{F}_0^2(\mathbb{R}^p)$ can be defined as

$$d^2_{\tau,\varphi}((x,A),(y,B)) = ||\Gamma(x,A) - \Gamma(y,B)||^2_{\tau,\varphi}$$

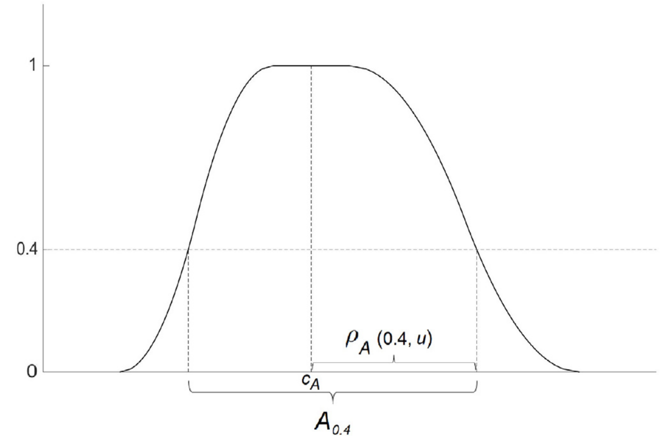

Fig. 2. Fuzzy set A in $\mathbb{R}$, its center c_A and its associated radial function for $u = (1, 0)$

$$= (1-\tau)(x-y)^t \cdot (x-y)$$

$$+\tau \int_0^1 \int_{\mathbb{S}^{p-1}} (\rho_A(u,\alpha) - \rho_B(u,\alpha))^2 \varphi(\alpha)\vartheta_p(du)\lambda(d\alpha).$$

As before, $\Gamma(\mathbb{R}^p \times \mathbb{F}_0^2) = \mathbb{R}^p \times L^2(\mathbb{S}^{p-1} \times (0,1], \vartheta_p \times \lambda)^+$ is a closed convex cone, $\Gamma : \mathbb{R}^p \times \mathbb{F}_0^2 \to H$ is an isometry preserving the arithmetic, and $(\mathbb{R}^p \times \mathbb{F}_0^2, d_\tau)$ is a complete and separable metric space.

The centre of the fuzzy sets must belong to all the α-level sets. Thus, it must belong to the 1-level set. In details, a function $A : \mathbb{R}^p \to [0,1]$ is a (non-centred at 0) fuzzy star-shaped set in $F_{\mathbb{X}}(\mathbb{R}^p)$ if there exists $c \in \mathbb{R}^p$ with $A_\alpha - c \in \mathbb{X}_0$ $\forall\alpha \in (0,1]$. Consider

$$ker(A) = \{x \in A_1 \,|\, A_\alpha - x \in \mathbb{X}_0^2 \ \forall\alpha \in (0,1]\},$$

and assume that it is bounded, then $ker(A) \in \mathcal{K}_c(\mathbb{R}^p)$. Consider the subspace generated by the linear span of A_1. If $int(A_1) \neq \emptyset$, then c_A is defined as the centre of mass of $ker^*(A)$, where

$$ker^*(A) = \{x \in A_1 \,|\, cl(int(A_\alpha - x)) \in \mathbb{X}_0^2 \ \forall\alpha \in (0,1]\} \in \mathcal{K}_c(\mathbb{R}^p).$$

However, if $int(A_1) = \emptyset$, then c_A will be defined as the centre of mass of $ker(A)$ with respect to $\lambda_{ker(A)}$, the Lebesgue measure on $ker(A)$.

In this way, all the theoretical framework provides the same structure of the embedding in a cone of a Hilbert space that was used to establish the concepts of random element, expected value, variance and covariances, as well as the asymptotic results for the case of random sets. As a result, all the inference derived from them is valid for the fuzzy case. Using essentially the same notation, bootstrap confidence sets, and hypothesis tests for the mean and the variance can be established and supported theoretically. The empirical results continue to be analogous to the real case.

An illustrative case study involving a one-way ANOVA test for fuzzy data can be found in González-Rodríguez et al. (2012). The study concerns an experiment in which people were asked for their perception of the relative length of several line segments with respect to a fixed longer segment that is used as a standard for comparison. A sample of 17 people from various countries, professions, ages and sexes have been considered. The study reflects that the test is more affected by differences in imprecision. In addition, significant differences have been detected among individuals, between men and women and among females. Still, no significant differences are detected among males, who are concluded to be more uniform in expressing their perceptions.

4. Regression: Fuzzy regression and regression with (fuzzy) sets

Fuzzy regression was not initially introduced as a statistical model. However, the difference among fuzzy regression, fitting curves with fuzzy data, and estimate regression models with fuzzy data has not always been clear, and it has been frequently argued that fuzzy regression can substitute, and even “work better”, than standard regression models under various circumstances (see, e.g., Kim et al., 1996 and Savic and Pedrycz, 1991). In contrast to the real case, the lack of linearity of the space of fuzzy or set-valued data makes the fitting problem different from the estimation problem. That is because considering or not the data generation process leads to unequal restrictions in the minimization problem (González-Rodríguez et al., 2009). In order to contribute to clarifying the particularities from a statistical point of view, these methodologies are presented in this section and their use is detailed.

4.1. Fuzzy/possibilistic regression

Fuzzy regression, also known as possibilistic regression, was introduced by (Tanaka et al., 1982) as a way to substitute the random deviations between the observed values and the predicted ones assumed in the standard regression models by fuzziness of the parameters. This is justified because those deviations are considered to be associated with "indefiniteness of the system structure" or vagueness, e.g., due to the involvement of human assessments, or a lack of precise knowledge. Thus, no data generation process is postulated. The fuzzy regression just seeks a numerical fit with fuzzy parameters belonging to certain parametrized classes by usually applying some linear programming techniques optimizing a criterion related to the fuzziness of the system.

In other words, fuzzy regression can be applied to fit a fuzzy function to relate any dependent variable with an arbitrary number of independent variables based on a set of tuples $\{y_i, x_{i1}, \ldots, x_{ik}\}_{i=1}^n$. However, this flexibility comes with a price: since the fit is performed numerically, without assuming any data generation process, there is no statistical guarantee on the estimations and no inference can be derived. This is clearly in contrast with the econometrics and statistical standards, but it can be of some help for descriptive purposes, as a fitting technique, by allowing some flexibility on the definition of the parameters through fuzziness. Applications in economics can be found, e.g., in Berry-Stölzle et al. (2010); Malyaretz et al. (2018) and Muzzioli et al., 2015. For instance, a comparison between different fuzzy regression models to estimate the implied volatility smile function in a financial framework has been carried out in Muzzioli et al. (2015).

Tanaka's fuzzy regression fits linear functions with symmetric triangular fuzzy parameters with either fuzzy or crisp output and crisp inputs. Symmetric triangular fuzzy sets are determined by a centre and a spread, so they are formally equivalent to intervals, although conceptually they are different. A fuzzy set U is called symmetric triangular fuzzy number if for a given $u \in \mathbb{R}$ and $r > 0$, its membership function is

$$U(x) = \begin{cases} 1 - \frac{|u-x|}{e} & if u - r < x < u + r \\ 0 & \text{elsewhere} \end{cases}.$$

In this case U is simply denoted by $U = (u, r)$. The arithmetic between fuzzy numbers is equivalently defined by Zadeh's extension principle (Zadeh, 1975) or by extended level-wise the interval arithmetic, i.e., for two triangular fuzzy numbers $U_1 = (u_1, r_1)$, $U_2 = (u_2, r_2)$, and $\lambda \in \mathbb{R}$,

$$U_1 + \lambda U_2 = (u_1, r_1) + \lambda(u_2, r_2) = (u_1 + \lambda u_2, r_1 + |\lambda| r_2).$$

Let $\{Y_i, x_{1i}, \ldots, x_{ki}\}_{i=1}^n$ be a set of data, where $Y_i = (y_i, e_i)$ are triangular fuzzy numbers and $x_{ij} \in \mathbb{R}$ for all $i = 1, \ldots, n$ and $j = 1 \ldots k$. Tanaka's fuzzy regression is meant to fit a fuzzy linear function

$$Y = A_0 + A_1 x_1 + \ldots + A_k x_k,$$

where A_j are symmetric triangular fuzzy numbers (α_j, c_j) for all $j = 1, \ldots k$. The aim is to find (α_j, c_j) minimizing the fuzziness of the model, defined as the sum of spreads of all the fuzzy parameters, under the constraint that the membership value of y_i to its "fuzzy estimate" Y_i^* is at least h for all $i = 1, \ldots, n$, where h is fixed a priori. The value $h \in [0, 1]$ is seen as a measure of goodness of fit or compatibility between the dataset and the fuzzy linear function. Formally, the problem can be written as follows:

$$\min \sum_{j=0}^{k} c_j,$$

subject to:

$$\sum_{j=0}^{k} \alpha_j x_{ij} + (1-h) \sum_{j=0}^{k} c_j |x_{ij}| \geq y_i, \sum_{j=0}^{k} \alpha_j x_{ij} - (1-h) \sum_{j=0}^{k} c_j |x_{ij}| \leq y_i,$$

$$c \geq 0, \alpha \in R, x_{i0} = 1.$$

The R package 'fuzzyreg: Fuzzy Linear Regression' (Skrabanek and Martinkova, 2019) implements different algorithms for fuzzy regression.

4.2. Fuzzy least-squares

A closer viewpoint of fuzzy regression to the standard regression is the so-called "fuzzy least-squares" (Diamond, 1988), where the aim is again to fit a linear function with fuzzy parameters. Instead of applying linear programming techniques to minimize the fuzziness under certain constraints, it searches for the fuzzy linear function that minimizes the overall distance between the observed fuzzy values and the predicted ones as traditional least-squares does. Some constraints must be imposed in order to cope with the lack of linearity of the space of fuzzy sets, inherited from the lack of linearity of the space of intervals or, more generally, compact and convex sets. Several distances can be defined and, depending on the type of fuzzy sets, different solutions have been established (see, e.g., Bargiela et al., 2007; Diamond and Körner, 1997 and D'Urso and Gastaldi, 2000).

The advantage of using least-squares is that the solution is delivered with an objective measure of the goodness-of-fit, based on the explained variability. However, these problems are just numerical fittings, and no data generation process is assumed, so no statistical estimation is provided. This is also the case of some regression problems for interval data (see, e.g., Diamond, 1990 and Gil et al., 2002). For example, a least squares approach for analyzing the degree of dependence between the systolic and diastolic blood pressure of 59 patients from a hospital in Asturias (Spain) from a descriptive point of view has been developed in Gil et al., 2002.

In the simpler case, given a set of fuzzy data $\{(Y_i, X_i)\}$, the problem is to find the closest affine transformation, that is, the closest linear function $Y = aX + B$, where $\alpha \in \mathbb{R}$ and B is a fuzzy set. In order to define the problem mathematically, the considered space of fuzzy sets is frequently $\mathcal{F}_c(\mathbb{R}^p)$, which is the space of fuzzy sets with convex and compact level sets (see, e.g., Puri and Ralescu, 1986). Sometimes more restrictive parametrized classes are considered, such as LR-fuzzy numbers, i.e., fuzzy sets of $\mathbb{R}$ determined by a centre, a left spread and a right spread (Dubois and Prade, 1980). The distance between the values predicted by the affine transformation and the observed outputs is measured in terms of a given metric d (see, e.g., Trutschnig et al., 2009). The minimization problem can be ill-posed if no conditions are assumed, given the conical structure associated with the usual arithmetic between fuzzy sets (see, e.g., Diamond, 1990 and Gil et al., 2002). Thus, formally the problem is to find $\alpha \in \mathbb{R}$ and B sot hat

$$\min \sum_{i=1}^{n} d(Y_i, \alpha X_i + B)$$

under certain suitable constraints making the problem and the solution well-defined in the considered space of fuzzy sets. For general fuzzy sets, the conditions are quite restrictive, while for LR-fuzzy numbers they are essentially related to the non-negativity of the spreads. The R package ‘fuzzyreg: Fuzzy Linear Regression’ (Skrabanek and Martinkova, 2019) implements the original fuzzy least-squares (Diamond, 1988).

4.3. Regression with fuzzy/interval data through separate models

Later on, some models including random errors have also been labelled as “fuzzy regression”, although their aim differs from the original Tanaka’s regression. Many of these models do not assume a data generation process of the data themselves yet. In contrast, they consider data generation processes separately for the characterizing parametrizations, e.g., minima and suprema, or centres and spreads, which does not always imply any explanatory relationship among the original variables (see, e.g., Coppi et al., 2006a and Lima Neto and de Carvalho, 2008). As an example, the dependence relationship of the daily atmospheric concentration of carbon monoxide (CO) in Rome (Italy) with a fuzzy summary of meteorological variables by considering hourly information during 21 days has been analyzed in Coppi et al. (2006a) through the least square estimation of a linear regression model with LR fuzzy response based on separate models.

In the case of interval data, let a^l, a^u, a^m and a^s denote respectively the lower bound, upper bound, mid-point and spread of an interval A. Given a set of interval-valued pairs $\{(Y_i, X_i)\}$, the technique of the separate models consists in assuming linear regression models between either lower bound and upper bounds, or between mid-points and spreads, that is:

$$y^l = \alpha_1 x^l + \beta_1 \epsilon_1 \quad \text{and} \quad y^u = \alpha_2 x^u + \beta_2 \epsilon_2 \text{ or}$$

$$y^m = \alpha_1 x^m + \beta_1 \epsilon_1 \quad \text{and} \quad y^s = \alpha_2 x^s + \beta_2 \epsilon_2,$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$ and ϵ_1 and ϵ_1 are the (real-valued) random errors. Once more, conditions are requested to guarantee that the models are well-defined and the solutions are coherent within the space of interval data. For instance, the model linking the spreads is not meaningful as defined, since both y^s and x^s are non-negative variables. During the estimation process, ill-definition in the predicted values can be avoided by using constraint least-squares, but this does not solve the problem out-of-sample. A solution developed in Ferraro and Giordani (2012) consists in using Box-Cox transformations to establish the models relating to the spreads. In Ferraro and Giordani, 2012, LR fuzzy sets are considered, which formally just adds another equation for the spreads. Cross-relationships between minima and maxima, or mid-points and spreads, could be trivially considered.

The R package ‘iRegression: Regression Methods for Interval-Valued Variables’ (Lima Neto et al., 2016) implements some separate regression methods for interval-valued variables. It should be noted that none of these models are regression models between interval-valued random elements, in the sense of modelling $E(Y|X = x)$ by considering the concept of expected value of random interval, and hence, the arithmetic between intervals.

4.4. Regression models with fuzzy/interval data

Standard regression models in statistics assume a data generation process making the conditional expected value $E(Y|X = x)$ to be a regression function in a given family, e.g., a linear, quadratic or smooth function, to name but a few. The simpler case, to compare with Section 4.2, has been developed in González-Rodríguez et al. (2009). In that study, the proposed regression model was applied to analyze the linear relationship between the quality of the trees in a reforestation area in

Asturias (Spain) and the quality of the land, where the measurement of both characteristics was given in terms of trapezoidal fuzzy numbers by expert perception.

In order to avoid ill-definitions for the non-negativity, the independent term is assumed to be embedded in the error term, that is, given two random fuzzy sets X and Y with non-negative and finite variances, the data generation process is

$$Y = \alpha X + \epsilon,$$

where $\alpha \in \mathbb{R}$, ϵ is a fuzzy-valued random error with $E(\epsilon|X) = B \in \mathcal{F}_c(\mathbb{R}^p)$. Thus, the population regression function is $E(Y|X) = \alpha X + B$. This model implies that ϵ is the Hukuhara difference between Y and αX, that is $\epsilon = Y -_H \alpha X$, and since the data are assumed to be generated from the model, the least-squares problem can be constrained to impose the existence of the residuals, that is,

$$\min_{\alpha \in A} \sum_{i=1}^{n} d(Y_i, \alpha X + B),$$

where $A = \{\alpha \in \mathbb{R} | Y_i -_H (\alpha X_i + B) \text{ exists for all } i = 1, \ldots n\}$. Under these conditions, closed expressions for the estimators of α and B can be obtained.

This model is, however, very restrictive, as it assumes the relationship among complex data to be represented in a single scalar $\alpha \in \mathbb{R}$. More flexible models exploiting functional regressions are expected in the context of fuzzy star-shaped sets (see, e.g., Ferraty et al., 2019; González-Rodríguez and Colubi, 2017 and Matsui, 2020).

Also for the interval case, more flexible models have been considered (see, e.g. Blanco-Fernández et al., 2013; Blanco-Fernández et al., 2011). By using the notation $A = [\text{mid}\,A \pm \text{spr}\,A]$, and the canonical decomposition of interval as $A = \text{mid}\,A[1 \pm 0] + sprA[0 \pm 1]$, García-Bárzana et al., 2020 proposed a general model for random intervals as:

$$Y = \alpha_1 \text{mid}\,\overrightarrow{X}[1 \pm 0] + \alpha_2 \text{spr}\,\overrightarrow{X}[0 \pm 1] + \alpha_3 \text{mid}\,\overrightarrow{X}[0 \pm 1] + \alpha_4 \text{spr}\,\overrightarrow{X}[1 \pm 0] + \epsilon,$$

where Y is the random interval-valued response, $\overrightarrow{X} = (X_1, \ldots X_k)$ are the interval-valued explanatory variables, $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}^k$, and ϵ is the random interval-valued error with $E(\epsilon|\overrightarrow{X}) = \Delta$, which is also an interval. The estimation can be done numerically by applying techniques to solve a constrained minimization problem.

All these models are standard regression problems defined within metric spaces endowed with a semilinear arithmetic rich enough to develop inferences, such as confidence regions and hypothesis testing (see, e.g., Blanco-Fernández and González-Rodríguez, 2016). R packages and algorithms in MatLab for flexible regression models for random intervals can be found at http://bellman.ciencias.uniovi.es/smire/IntervalLM.html.

5. Conclusions

The aim has been to review ways to handle blurriness and lack of precision represented through (fuzzy) sets in Econometrics and Statistics. There are two main blocks.

The first one includes problems where relaxing sharp boundaries makes the data analysis more accurate and richer. As two main examples, fuzzy clustering and FRDDs have been considered. FRDDs have a particularity with respect to the other cases considered in the present manuscript: the blurriness is associated with a frequentist distribution, and not with fuzziness in Zadeh's sense (Zadeh, 1965). However, it has been included for two reasons: 1) It is one of the topics where the word "fuzzy" more appears in Econometrics nowadays; 2) Neither the frequentist nature, nor its potential to be represented as a fuzzy set, are really exploited, and putting it in context with other problems undergoing fuzziness can serve as inspiration for future research.

The second block comprises all those studies related to the employment of (fuzzy) random sets and (fuzzy) set-valued data. This block can be subdivided into two more categories: problems related to partial identification typical in econometrics, and problems in which the available data do not reduce to a single observation of $\mathbb{R}^p$, either because they form a set, such as ranges or convex particles, or because they represent subjective perceptions, such as medical diagnosis or ratings.

Interval data appear naturally in many contexts, because it is not uncommon to have access only to grouped data or ranges. They can be easily analyzed through a two-dimensional vector, e.g., mid-spread. However, one should be aware that the natural arithmetic between intervals does not agree with the natural arithmetic in $\mathbb{R}^2$, even if the restricted cone $\mathbb{R} \times \mathbb{R}^+$ is used to account for the non-negativity of the spreads. Also, the metrics more meaningful for intervals should agree with the human perception when assessing distances between intervals. In any case, a rich set of tools for the statistical analysis of interval data is available, as shown throughout the paper.

The generalization of the interval case can lead to two situations: compact and convex sets of $\mathbb{R}^p$, or fuzzy sets of $\mathbb{R}$, which can be further extended to fuzzy sets of $\mathbb{R}^p$. The usual interval arithmetic, which is defined through the Minkowski addition, also called dilation, is not always meaningful when working with imprecision in higher dimensions. The reason is that the imprecision is expanded in all directions. An alternative has been presented through (fuzzy) star-shaped sets and an arithmetic that extends the interval arithmetic directionally. From a formal point of view, the framework reduces to work with an easy-to-handle cone within a Hilbert space. Thus, traditional tools of statistics in Hilbert space and functional data analysis can be applied. This framework has not been fully exploited yet, but standard inferences such as confidence regions and hypothesis tests for means and variances have already been implemented and are available upon request.

Special attention has been paid to the associated regression problems. The reason is that there are different points of view that are important to distinguish. First, the so-called Tanaka's fuzzy regression, which lies within the possibilistic and not the probabilistic theory. Tanaka's fuzzy regression seeks to fit a (linear) fuzzy function minimizing the fuzziness of the system under certain constraints that guarantee that the function has a certain degree of agreement with the data. It does not take into account the distance between the predictions and the data, as it does the traditional least-squares approaches. For this reason, fuzzy least-squares were considered as well. However, initially, only a numerical fitting to an affine function was considered. Probably this was done like that because in the real case the minimization problem of searching for the numerical fitting and the one to estimate the parameters of a potentially well-defined underlying regression model is the same. In contrast, these two problems are not the same in the fuzzy (or interval) case, due to the lack of linearity of the spaces and the constraints that this implies. In the same way, standard separate regression models for certain kinds of parametrized fuzzy sets (including the interval case) have also been developed. These separate models do not imply a joint model in the space of fuzzy sets, but they are useful to relate important features within the fuzzy sets, such as location and imprecision. Finally, there are some standard regression models both in the space of fuzzy sets and in the space of intervals. Although for the interval case there are quite flexible models, the space of fuzzy sets is not fully developed yet.

Acknowledgment

The research in the present manuscript has been partially supported by the Spanish Government through MINECO-18-MTM2017-89632-P Grant. Their finantial support is greatfully acknowledged.

References

Adusumilli, K., Otsu, T., 2017. Empirical likelihood for random sets. Journal of the American Statistical Association 112 (519), 1064–1075.

Arai, Y., Hsu, Y., Kitagawa, T., Mourifié, I., Wan, Y., 2021. Testing identifying assumptions in fuzzy regression discontinuity designs. Quantitative Economics 13, 1–2.

Araujo, A., Gine, E., 1980. The central limit theorem for real and Banach valued random variables. Wiley & Sons, New York.

Bargiela, A., Pedrycz, W., Nakashima, T., 2007. Multiple regression with fuzzy data. Fuzzy Sets and Systems 158 (19), 2169–2188.

Basten, C., Betz, F., 2013. Beyond work ethic: Religion, individual, and political preferences. American Economic Journal: Economic Policy 5 (3), 67–91.

Bellman, R.E., Zadeh, L.A., 1970. Decision-making in a fuzzy environment. Management Science 17 (4), 141–164.

Beresteanu, A., Molinari, F., 2008. Asymptotic properties for a class of partially identified models. Econometrica 76 (4), 763–814.

Berry-Stölzle, T., Koissi, M.-C., Shapiro, A., 2010. Detecting fuzzy relationships in regression models: The case of insurer solvency surveillance in germany. Insurance: Mathematics and Economics 46 (3), 554–567.

Bertanha, M., Imbens, G.W., 2020. External validity in fuzzy regression discontinuity designs. Journal of Business & Economic Statistics 38, 593–612.

Bezdek, J., 1981. Pattern Recognition With Fuzzy Objective Function Algorithms. Plenum Press, New York.

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. Information Sciences 247, 109–122.

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

Blanco-Fernández, A., González-Rodríguez, G., 2016. Inferential studies for a flexible linear regression model for interval-valued variables. International Journal of Computer Mathematics 93 (4), 658–675.

Blanco-Fernández, A., Warren-Liao, T., 2015. A bootstrap factorial anova for random intervals. Advances in Intelligent Systems and Computing 315, 193–201.

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., Bärnighausen, T., 2014. Regression discontinuity designs in epidemiology: causal inference without randomized trials. Epidemiology 25, 729–737.

Calonico, S., Cattaneo, M., Titiunik, R., 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica 82 (6), 2295–2326.

Cameron, T., Huppert, D., 1989. Ols versus ml estimation of non-market resource values with payment card interval data. Journal of Environmental Economics and Management 17 (3), 230–246.

Cardella, E., Depew, B., 2014. The effect of health insurance coverage on the reported health of young adults. Economics Letters 124 (3), 406–410.

Carleos, C., López-Díaz, M.C., López-Díaz, M., 2014. Ranking star-shaped valued mappings with respect to shape variability. Journal of Mathematical Imaging and Vision 48 (1), 1–12.

Cattaneo, M.D., Keele, L., Titiunik, R., Vázquez-Bare, G., 2016. Interpreting regression discontinuity designs with multiple cutoffs. Journal of Politics 78 (4), 1229–1248.

Chen, C.-T., Lin, C.-T., Huang, S.F., 2006. A fuzzy approach for supplier evaluation and selection in supply chain management. International Journal of Production Economics 102 (2), 289–301.

Cheng, C.-H., Wei, L.-Y., Liu, J.-W., Chen, T.L., 2013. Owa-based anfis model for taiex forecasting. Economic Modelling 30 (1), 442–448.

Choi, J.-Y., Lee, M.J., 2018. Relaxing conditions for local average treatment effect in fuzzy regression discontinuity. Economics Letters 173, 47–50.

Choirat, C., Seri, R., 2014. Bootstrap confidence sets for the aumann mean of a random closed set. Computational Statistics and Data Analysis 71, 803–817.

Colubi, A., González-Rodríguez, G., 2015. Fuzziness in data analysis: Towards accuracy and robustness. Fuzzy Sets and Systems 281, 260–271.

Colubi, A., González-Rodríguez, G., Gil, M.A., Trutschnig, W., 2011. Nonparametric criteria for supervised classification of fuzzy data. International Journal of Approximate Reasoning 52 (9), 1272–1282.

Coppi, R., D'Urso, P., Giordani, P., Santoro, A., 2006a. Least squares estimation of a linear regression model with lr fuzzy response. Computational Statistics and Data Analysis 51 (1), 267–286.

Coppi, R., Giordani, P., D'Urso, P., 2006b. Component models for fuzzy data. Psychometrika 71 (4), 733–761.

De Carvalho, F., 2007. Fuzzy c-means clustering methods for symbolic interval data. Pattern Recognition Letters 28 (4), 423–437.

Denœux, T., 2000. Modeling vague beliefs using fuzzy-valued belief structures. Fuzzy Sets and Systems 116 (2), 167–199.

Diamond, P., 1988. Fuzzy least squares. Information Sciences 46 (3), 141–157.

Diamond, P., 1990. Least squares fitting of compact set-valued data. Journal of Mathematical Analysis and Applications 147 (2), 351–362.

Diamond, P., Körner, R., 1997. Extended fuzzy linear models and least squares estimates. Computers and Mathematics with Applications 33 (9), 15–32.

Dimmery, D. (2016). rdd: Regression discontinuity estimation. URL https://CRAN.R-project.org/package=rdd.

Dong, Y., 2018. Alternative assumptions to identify LATE in fuzzy regression discontinuity designs. Oxford Bulletin of Economics and Statistics 80 (5), 1020–1027.

Dubois, D., Prade, H., 1980. Fuzzy Sets and Systems: Theory and Applications. Academic Press, New York.

D'Urso, P., Gastaldi, T., 2000. A least-squares approach to fuzzy linear regression analysis. Computational Statistics and Data Analysis 34 (4), 427–440.

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

Ferraro, M.B., 2021. Fuzzy k-means: History and applications. Econometrics and Statistics. In Press.

Ferraro, M.B., Coppi, R., González-Rodríguez, G., Colubi, A., 2010. A linear regression model for imprecise response. International Journal of Approximate Reasoning 51 (7), 759–770.

Ferraro, M.B., Fernández-Iglesias, E., Ramos-Guajardo, A.B., González-Rodríguez, G., 2022. On clustering of star-shaped sets with a fuzzy approach: an application to the clasts in the cantabrian coast. In: 10th international conference on soft methods in probability and statistics. Submitted

Ferraro, M.B., Giordani, P., 2012. A multiple linear regression model for imprecise information. Metrika 75 (8), 1049–1068.

Ferraro, M.B., Giordani, P., 2020. Soft clustering. Wiley Interdisciplinary Reviews: Computational Statistics 12 (1), e1480.

Ferraro, M.B., Giordani, P., Serafini, A., 2019. fclust: An r package for fuzzy clustering. The R Journal 11, 198–210. URL https://journal.r-project.org/archive/2019/RJ-2019-017/RJ-2019-017.pdf.

Ferraro, M.B., Giordani, P., Vichi, M., 2021. A class of two-mode clustering algorithms in a fuzzy setting. Econometrics and Statistics 18, 63–78.

Ferraty, F., Zullo, A., Fauvel, M., 2019. Nonparametric regression on contaminated functional predictor with application to hyperspectral data. Econometrics and Statistics 9, 95–107.

Fischer, H., Blanco-Fernández, A., Winker, P., 2016. Predicting stock return volatility: Can we benefit from regression models for return intervals? Journal of Forecasting 35 (2), 113–146.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of American Statistical Association 97, 611–631.

García-Bárzana, M., Ramos-Guajardo, A.B., Colubi, A., Kontoghiorghes, E., 2020. Multiple linear regression models for random intervals: a set arithmetic approach. Computational Statistics 35 (2), 755–773.

Gibson, R., Gyger, S., 2007. The style consistency of hedge funds. European Financial Management 13 (2), 287–308.

Gil, M.A., González-Rodríguez, G., Colubi, A., Montenegro, M., 2007. Testing linear independence in linear models with interval-valued data. Computational Statistics and Data Analysis 51 (6), 3002–3015.

Gil, M.A., Lubiano, M.A., Montenegro, M., López, M.T., 2002. Least squares fitting of an affine function and strength of association for interval-valued data. Metrika 56 (2), 97–111.

González-Rodríguez, G. (2020). Random fuzzy star-shaped sets. Submitted.

González-Rodríguez, G., Blanco-Fernández, A., Colubi, A., Lubiano, M.A., 2009. Estimation of a simple linear regression model for fuzzy random variables. Fuzzy Sets and Systems 160 (3), 357–370.

González-Rodríguez, G., Colubi, A., 2017. On the consistency of bootstrap methods in separable hilbert spaces. Econometrics and Statistics 1, 118–127.

González-Rodríguez, G., Colubi, A., Gil, M.A., 2012. Fuzzy data treated as functional data: A one-way anova test approach. Computational Statistics and Data Analysis 56 (4), 943–955.

González-Rodríguez, G., Ramos-Guajardo, A.B., Colubi, A., Blanco-Fernández, A., 2018. A new framework for the statistical analysis of set-valued random elements. International Journal of Approximate Reasoning 92, 279–294.

Hahn, J., Todd, P., Van Der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69 (1), 201–209.

Hathaway, R., Bezdek, J., 1993. Switching regression models and fuzzy clustering. IEEE Transactions on Fuzzy Systems 1 (3), 195–204.

Henning, C., Meila, M.F., Rocci, R., 2015. Handbook of cluster analysis. Chapman and Hall CRC, Boca Raton, Florida.

Huang, H.-C., Chuang, Y.-Y., Chen, C.S., 2012. Multiple kernel fuzzy clustering. IEEE Transactions on Fuzzy Systems 20 (1), 120–134.

Imbens, G., Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. Journal of Econometrics 142 (2), 615–635.

Jain, A., Murty, M., Flynn, P., 1999. Data clustering: A review. ACM Computing Surveys 31 (3), 264–323.

Kasa, S.R., Rajan, V., 2022. Improved inference of gaussian mixture copula model for clustering and reproducibility analysis using automatic differentiation. Econometrics and Statistics 22, 67–97.

Kim, K., Moskowitz, H., Koksalan, M., 1996. Fuzzy versus statistical linear regression. European Journal of Operational Research 92 (2), 417–434.

Klain, D., 1997. Invariant valuations on star-shaped sets. Advances in Mathematics 125 (1), 95–113.

Klawonn, F., Höppner, F., 2003. What is fuzzy about fuzzy clustering? understanding and improving the concept of the fuzzifier. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2810, 254–264.

Körner, R., 2000. An asymptotic α-test for the expectation of random fuzzy variables. Journal of Statistical Planning and Inference 83 (2), 331–346.

Krishnapuram, R., Keller, J., 1993. A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems 1 (2), 98–110.

Lee, D., Lemieux, T., 2010. Regression discontinuity designs in economics. Journal of Economic Literature 48 (2), 281–355.

Lima Neto, E., de Carvalho, S.F., 2008. Centre and range method for fitting a linear regression model to symbolic interval data. Computational Statistics and Data Analysis 52 (3), 1500–1515.

Lima Neto, E., de Carvalho, S. F., & Marinho, P. (2016). iregression: Regression methods for interval-valued variables. URL https://CRAN.R-project.org/package=iRegression.

Lubiano, M.A., De La Rosa De Sáa, S., Montenegro, M., Sinova, B., Gil, M.A., 2016. Descriptive analysis of responses to items in questionnaires. Why not using a fuzzy rating scale? Information Sciences 360, 131–148.

Maciel, L., Gomide, F., Ballini, R., 2016. Evolving fuzzy-garch approach for financial volatility modeling and forecasting. Computational Economics 48 (3), 379–398.

Malyaretz, L., Dorokhov, O., Dorokhova, L., 2018. Method of constructing the fuzzy regression model of bank competitiveness. Journal of Central Banking Theory and Practice 7 (2Ç), 139–164.

Mamdani, E., Assilian, S., 1975. Experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 7 (1), 1–13.

Manski, C., Molinari, F., 2010. Rounding probabilistic expectations in surveys. Journal of Business and Economic Statistics 28 (2), 219–231.

Manski, C., Tamer, E., 2002. Inference on regressions with interval data on a regressor or outcom. Econometrica 70 (2), 519–546.

Matheron, G., 2018. Random Sets and Integral Geometry. Wiley & Sons, New York.

Matsui, H., 2020. Quadratic regression for functional response models. Econometrics and Statistics 13, 125–136.

McLachlan, G.J., Peel, D., 2000. Finite mixture models. John Wiley & Sons, Hoboken, New Jersey.

Molchanov, I., Cascos, I., 2016. Multivariate risk measures: a constructive approach based on selections. Mathematical Finance 26 (4), 867–900.

Molchanov, I., Molinari, F., 2014. Applications of random set theory in econometrics. Annual Review of Economics 6, 229–251.

Molchanov, I., Molinari, F., 2018. Random Sets in Econometrics. Cambridge University Press, Cambridge.

Murray, P.M., Browne, R.P., 2017. A mixture of SDB skew-t factor analyzers. Econometrics and Statistics 3, 160–168.

Muzzioli, S., Ruggieri, A., De Baets, B., 2015. A comparison of fuzzy regression methods for the estimation of the implied volatility smile function. Fuzzy Sets and Systems 266, 131–143.

Nakama, T., Colubi, A., Lubiano, M.A., 2010. Two-way analysis of variance for interval-valued data. Advances in Intelligent and Soft Computing 77, 475–482.

Näther, W., 1997. Linear statistical inference for random fuzzy data. Statistics 29 (3), 221–240.

Pawlak, A., 1982. Rough sets. International Journal of Information and Computer Sciences 11, 145–172.

Pawlak, Z., 1992. Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers Norwell, Dordrecht, Netherlands.

Peel, D., McLachlan, G., 2000. Robust mixture modelling using the t distribution. Statistics and Computing 10, 339–348.

Puri, M., Ralescu, D., 1986. Fuzzy random variables. Journal of Mathematical Analysis and Applications 114 (2), 409–422.

Qu, X., Zhang, G., 2010. Measuring the convergence of national accounting standards with international financial reporting standards: The application of fuzzy clustering analysis. International Journal of Accounting 45 (3), 334–355.

Ramos-Guajardo, A.B., Blanco-Fernández, A., González-Rodríguez, G., 2019. Applying statistical methods with imprecise data to quality control in cheese manufacturing. Studies in Systems, Decision and Control 183, 127–147.

Ramos-Guajardo, A.B., Colubi, A., González-Rodríguez, G., 2014. Inclusion degree tests for the aumann expectation of a random interval. Information Sciences 288, 412–422.
Ramos-Guajardo, A.B., González-Rodríguez, G., 2013. Testing the variability of interval data: An application to tidal fluctuation. Studies in Fuzziness and Soft Computing 285, 65–74.
Ramos-Guajardo, A.B., González-Rodríguez, G., Colubi, A., 2020. Testing the degree of overlap for the expected value of random intervals. International Journal of Approximate Reasoning 119, 1–19.
Ramos-Guajardo, A.B., Lubiano, M.A., 2012. K-sample tests for equality of variances of random fuzzy sets. Computational Statistics and Data Analysis 56 (4), 956–966.
Ruspini, E., 1969. A new approach to clustering. Information and Control 15 (1), 22–32.
Sahin, O., Czado, C., 2022. Vine copula mixture models and clustering for non-gaussian data. Econometrics and Statistics 22, 136–158.
Savic, D., Pedrycz, W., 1991. Evaluation of fuzzy linear regression model. Fuzzy Sets and Systems 39 (1), 51–63.
Shapiro, A., 2009. Fuzzy random variables. Insurance: Mathematics and Economics 44 (2), 307–314.
Simó, A., De Ves, E., Ayala, G., 2004. Resuming shapes with applications. Journal of Mathematical Imaging and Vision 20 (3), 209–222.
Sinova, B., Colubi, A., Gil, M.A., González-Rodriguez, G., 2012. Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric. Information Sciences 199, 109–124.
Skrabanek, P., & Martinkova, N. (2019). fuzzyreg: Fuzzy linear regression. URL https://CRAN.R-project.org/package=fuzzyreg.
Takagi, T., Sugeno, M., 1985. Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man and Cybernetics SMC 15 (1), 116–132.
Tanaka, H., Uejima, S., Asai, K., 1982. Linear regression analysis with fuzzy model. IEEE Transactions on Systems, Man and Cybernetics 12 (6), 903–907.
Thistlethwaite, D., Campbell, D., 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology 51 (6), 309–317.
Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A., 2009. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences 179 (23), 3964–3972.
Viertl, R., 2011. Statistical Methods for Fuzzy Data. John Wiley & Sons, Chichester.
Wang, C.-C., Hsu, Y.-S., Liou, C.H., 2011. A comparison of arima forecasting and heuristic modelling. Applied Financial Economics 21 (15), 1095–1102.
Yang, M.-S., Hung, W.-L., Cheng, F.C., 2006. Mixed-variable fuzzy clustering approach to part family and machine cell formation for gt applications. International Journal of Production Economics 103 (1), 185–198.
Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8 (3), 338–353.
Zadeh, L.A., 1973. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man and Cybernetics SMC 3 (1), 28–44.
Zadeh, L.A., 1975. The concept of a linguistic variable and its application to approximate reasoning-i. Information Sciences 8 (3), 199–249.
Zimmermann, H.J., 1976. Description and optimization of fuzzy systems. International Journal of General Systems 2 (4), 209–215.