Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 00 (2023) 1–20
Fuzzy Sets and Systems

Significance Measures for rules in Probabilistic-Fuzzy Inference Systems based on Fuzzy transforms

Nicolás Madrid

Universidad de Málaga, Dept. Matemática Aplicada.
Blv. Louis Pasteur 35, 29071 Málaga, Spain.

Abstract

This paper defines a novel probabilistic-fuzzy inference system that considers fuzzy inputs and returns, as output, a probability distribution. In this way, it combines two different ways to represent uncertainty: the one modeled by fuzzy theory, that allows to represent reliable but vague information; and the one modeled by probability theory, that allows to represent undetermined but specific information. The novelty of this probabilistic-fuzzy inference system, with respect to the other existing in the literature, is that its inference engine combines quantile functions instead of distribution, probabilistic or density functions. Besides the formal definition of this novel kind of fuzzy inference systems, we propose: firstly, the construction of probabilistic-fuzzy rules by means of direct quantiles F-transforms; secondly, the definition of several significance measures for the obtained association rules; and finally, we present a set of experiments to validate all the assertions done throughout the paper.

Keywords: Fuzzy Transforms, Quantile regression, Fuzzy association rules, Fuzzy inference systems.

1. Introduction

Uncertainty is a feature that appears naturally in data and in different forms. Perhaps, the two most important mathematical theories to deal with uncertainty are fuzzy logic theory [17, 8] and probability theory [28, 3]. The former focuses on the uncertainty that human beings use in language for reasoning and communication. The latter focuses on the uncertainty appearing on the ignorance about the occurrence of an event. Both theories use degrees to represent the uncertainty but with two completely different meanings. One of the most typical and representative examples of this difference is the *"*5 *poisoned glasses of water with degree* 0.2*"*. From a fuzzy logic point of view, that degree means that in each glass there is a ratio of 0.8 drinkable water and 0.2 of poisoned water, whereas from a probabilistic point of view, that statement means that 1 of those 5 glasses is poisoned whereas the rest are glasses with only drinkable water. In this paper we join both theories by proposing a fuzzy inference system that takes fuzzy data as input and returns a probability distribution as output. Such a consideration in inference systems makes plenty of sense: on the one hand, since the inputs in an inference system are entirely known, in the sense that it is an event that has already happened, the probability has not place there and then, the uncertainty that may appear is only related to fuzziness; on the other hand, if the inference system aims at being applied for forecasting, the probability theory is the best theory that should be used to represent the inherent uncertainty about the unknown event that is coming.

This paper is framed in the so-called Probabilistic Fuzzy Inference Systems (PFIS), which encompasses any system that combines fuzzy set theory and probability theory to perform an inference. Each approach in PFIS can

Email address: nicolas.madrid@uma.es (Nicolás Madrid)

be classified in the way they combine fuzzy set theory and probability theories: there are approaches that consider fuzzy events (i.e., fuzzy sets with a certain probability assigned) in antecedents and consequents of rules [12, 19, 13]; approaches where a probability is assigned to fuzzy rules [31, 2]; approaches where a probabilistic uncertainty is linked to the truth values of the fuzzy sets in the consequents [30, 24]; and approaches where the antecedents of rules are fuzzy sets and the consequents are probability distributions of a (usually crisp) random variable [1, 27]. Our approach is framed in this later class. The most approaches in PFISs make use of the fuzzy probability theory [5, 29] to define rules or to perform the inference process; actually, all the previously mentioned approaches about PFIS make use of fuzzy probability theory. In contrast, our approach is not based on fuzzy probability theory but in one recent extension of Fuzzy Transforms (F-transforms) called L_1-Fuzzy transforms or Quantile Fuzzy Transforms (QF-transforms) [6, 7]. Different approaches in the literature propose the use of F-transforms for the construction of fuzzy rules in fuzzy inference systems [18, 21, 25, 4], but to the best of our knowledge, this is the first time that it is proposed for the construction of probabilistic fuzzy rules.

Apart form the use of QF-transforms to construct probabilistic fuzzy rules, the main novelty of the proposed PFIS is the inference process. Specifically, the proposed inference process is based on a weighted mean of quantiles functions instead of density functions (or probability functions), which is the common inference process in PFIS [1, 27, 31, 2]. We show that the use of quantiles functions to perform the inference reduces the uncertainty of the output with respect to the use of density functions.

In addition to the PFIS, we present some significance measures aimed at gathering information represented by each rule. The first, based on the well known Kolmogorov-Smirnov test [11, 22], pretends to determine whether there is a dependence between the dependent and independent variables. The second, based on statistical dispersion measures, pretends to measure how does the uncertainty about the dependent variable change after applying a probabilistic fuzzy rule. The third, and last measure, called translation measure, is designed to determine under which conditions, the dependent variable is greater or lesser.

The paper is organized as follows. In Section 2 we fix terminology of probability theory, fuzzy theory and recall some basic notions of F-transforms. In Section 3 we introduce the proposed PFIS, present the construction method for the probabilistic fuzzy rules based on QF-transforms and define the significance measures. Then, in Section 4, we illustrate the behaviour of the proposed inference system with some experiments. In this respect, we consider both synthetic and real data in order to show that the output is an actual approximation of the probability distribution of the dependent variable conditioned to the values of the independent variables. In Section 5 we describe the relation between our approach and others in the literature concerning PFIS and F-transforms. Finally, in Section 6, we provide some conclusions and future lines of research.

2. Preliminaries

In order to make this paper as self-contained as possible, we recall in this section some basic notions of probability theory and fuzzy transforms that are used through all the paper.

2.1. Cumulative probability distributions and quantile functions

Given a real-valued random variable $X\colon \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathscr{F}, P)$, the associated *cumulative probability distribution* of X is the mapping $F\colon \mathbb{R} \to [0,1]$ defined by $F(x) = P(X \leq x)$, that is, the mapping that assigns to each $x \in \mathbb{R}$ the probability of "X is lesser or equal than x". Every cumulative probability distribution is monotonically increasing and right continuous. The *quantile function* of a cumulative probability distribution is the mapping $Q\colon [0,1] \to \mathbb{R}$ defined by $Q(q) = \inf\{x \in \mathbb{R}\colon F(x) \geq q\}$. The value $Q(0.5)$ is called median and the values $Q(0.25)$ and $Q(0.75)$ first and third quartiles, respectively. Note that $F(Q(q)) = q$ and that if F is bijective, then $Q = F^{-1}$.

Every probability distribution of a real-valued random variable X is completely determined by its cumulative probability distribution or by its quantile function. For such a reason, and because the paper focus mainly on quantiles, hereafter the probability distribution of any real-valued random variable X will be given always in terms of its quantile function.

2.2. Fuzzy transforms (F-transforms)

In this paper we identify any fuzzy set A, on a universe $\mathscr{U}$, with its membership function $A\colon \mathscr{U} \to [0,1]$. A fuzzy partition Δ of a universe $\mathscr{U}$ is a set of fuzzy sets $\Delta_1,\dots,\Delta_n$ on $\mathscr{U}$ (called *classes*) fulfilling the covering property, i.e. for all $u \in \mathscr{U}$ there exists $k \in \{1,\dots,n\}$ such that $\Delta_k(u) > 0$. In the literature the reader can find several additional conditions that are usually imposed on fuzzy partitions; e.g., normality, continuity, convexity or Ruspini condition [20].

Since in this paper we are interested in analyzing data without a functional structure, we recall below the definition of fuzzy transforms given in [14] instead of the original one that is applied only to functions [20]. Accordingly, we consider a finite subset $\mathbf{T} = \{(x_i,y_i)\}_{i\in\mathbb{I}}$ of $\mathscr{U}\times\mathbb{R}$ which may have a non-functional structure; i.e. for the same $x \in \mathscr{U}$ it may exist either two tuples $(x,y_1),(x,y_2) \in \mathbf{T}$ satisfying $y_1 \neq y_2$ or it may exist $x_0 \in \mathscr{U}$ such that $(x_0,y) \notin \mathbf{T}$ for all $y \in \mathbb{R}$.

Definition 1. *Let $\mathbf{T} = \{(x_i,y_i)\}_{i\in\mathbb{I}} \subseteq \mathscr{U}\times\mathbb{R}$ and let $\Delta = \{\Delta_1,\dots,\Delta_n\}$ be a fuzzy partition of $\mathscr{U}$. We say that the n-tuple $\mathbf{F}_\Delta[\mathbf{T}] = [F_1,\dots,F_n] \in \mathbb{R}^n$ is the direct F-transform of $\mathbf{T}$ w.r.t. Δ if*

$$F_k = \frac{\sum_{i\in\mathbb{I}} y_i \cdot \Delta_k(x_i)}{\sum_{i\in\mathbb{I}} \Delta_k(x_i)} \tag{1}$$

for all $k \in \{1,\dots,n\}$.

It is not difficult to check that the previous definition extends the original one in the following way: given a function $f\colon \mathscr{U} \to \mathbb{R}$, Definition 1 coincides with the original definition given in [20] by identifying f with the subset $\mathbf{T}_f = \{(x,f(x)) \mid x \in \mathscr{U}\} \subseteq \mathscr{U}\times\mathbb{R}$. As in the original definition [20], the components of the direct F-transform coincide with a *least squares weighted* solution, as recalled in the following proposition.

Proposition 1 ([14])**.** *Let $\mathbf{T} = \{(x_i,y_i)\}_{i\in\mathbb{I}} \subseteq \mathscr{U}\times\mathbb{R}$ and let $\Delta = \{\Delta_1,\dots,\Delta_n\}$ be a fuzzy partition of $\mathscr{U}$. Then, the k-th component of the direct F-transform is the minimum of the following function:*

$$\phi(z) = \sum_{i\in\mathbb{I}} (y_i - z)^2 \cdot \Delta_k(x_i). \tag{2}$$

Note that from the previous result, we can interpret the k-th component of the direct F-transform as *"the mean of the values y's of those tuples $(x,y) \in \mathbf{T}$ such that x belongs to the class Δ_k."* Note that in the previous interpretation, the membership of x to Δ_k is fuzzy.

As in the original approach [20], the inverse F-transform is a function from $\mathscr{U}$ to $\mathbb{R}$ defined by means of the direct F-transform components.

Definition 2. *Let $\mathbf{T} = \{(x_i,y_i)\}_{i\in\mathbb{I}} \subseteq \mathscr{U}\times\mathbb{R}$ and let $\mathbf{F}_\Delta[\mathbf{T}] = [F_1,\dots,F_n] \in \mathbb{R}^n$ be the* direct F-transform *of $\mathbf{T}$ w.r.t. Δ. Then, the function defined, for all $x \in \mathscr{U}$, as:*

$$\mathbf{T}^F_\Delta(x) = \frac{\sum_{k=1}^n F_k \cdot \Delta_k(x)}{\sum_{k=1}^n \Delta_k(x)} \tag{3}$$

is called the inverse F-transform *of $\mathbf{T}$ w.r.t. Δ.*

Some remarks about the previous definition. Firstly, the inverse F-transform $\mathbf{T}^F_\Delta(x)$ is a function independently whether the set $\mathbf{T}$ has the structure of a function or not. Secondly, note that the domain of the inverse F-transform is $\mathscr{U}$, so $\mathbf{T}^F_\Delta(x)$ is defined even for those $x \in \mathscr{U}$ such that there is not $(x,y) \in \mathbf{T}$; i.e., it may be considered an interpolation and regression technique. Finally, the inverse F-transform is closely related to the function obtained by assigning to each $x \in \mathscr{U}$ the mean among all the y_i such that $(x,y_i) \in \mathbf{T}$ (see [14] for more details).

2.3. L_1-F-transforms and QF-transforms

Taking Proposition 1 as a reference, [6] proposes a modification of the (standard) direct F-transform as a minimizer of a residual absolute error instead of a residual square error.

Definition 3 ([6]). *Let $\mathbf{T} = \{(x_i, y_i)\}_{i\in\mathbb{I}} \subseteq \mathscr{U} \times \mathbb{R}$ and let $\Delta = \{\Delta_1, \ldots, \Delta_n\}$ be a fuzzy partition of $\mathscr{U}$. We say that the n-tuple $\mathbf{F}_{\Delta}^{L_1}[\mathbf{T}] = [F_1, \ldots, F_n] \in \mathbb{R}^n$ is the direct L_1-F-transform of $\mathbf{T}$ w.r.t. Δ, if for each $k \in \{1, \ldots, n\}$, F_k is a minimizer[1] of the following function:*

$$\phi(z) = \sum_{i\in\mathbb{I}} |y_i - z| \cdot \Delta_k(x_i) \tag{4}$$

Here it is worth recalling that the median of a dataset $Y = \{y_i\}_{i\in\mathbb{I}}$ can be characterized as the minimizer of the function $\phi(z) = \sum_{i\in\mathbb{I}} |y_i - z|$. Therefore, we can assert that the main difference between the L_1-F-transforms and standard F-transform is that the former is related to the median whereas the latter to the mean. Going one step further, and taking into account that the q-th quantile of a dataset $Y = \{y_i\}_{i\in\mathbb{I}}$ coincides with the minimizer of the objective function:

$$\Phi(z) = \sum_{i\in\mathbb{I}} w_q(y_i) \cdot |y_i - z|,$$

where $w_q(y_i)$ is the weighted function

$$w_q(y_i) = \begin{cases} 1-q & \text{if } y_i < z \\ q & \text{if } y_i \geq z, \end{cases}$$

we can define the direct q-th quantile F-transform (or the q-th QF-transform) as follows.

Definition 4. *Let $\mathbf{T} = \{(x_i, y_i)\}_{i\in\mathbb{I}} \subseteq \mathscr{U} \times \mathbb{R}$, let $\Delta = \{\Delta_1, \ldots, \Delta_n\}$ be a fuzzy partition of $\mathscr{U}$ and let $q \in [0,1]$. We say that the n-tuple $\mathbf{QF}_{\Delta}^{q}[\mathbf{T}] = [F_1, \ldots, F_n] \in \mathbb{R}^n$ is the direct q-th QF-transform of $\mathbf{T}$ w.r.t. Δ, if for each $k \in \{1, \ldots, n\}$, F_k is a minimizer of the following function:*

$$\phi(z) = \sum_{i\in\mathbb{I}} w_q(y_i) \cdot |y_i - z| \cdot \Delta_k(x_i) \tag{5}$$

where $w_k(y_i)$ is the weighted function

$$w_q(y_i) = \begin{cases} 1-q & \text{if } y_i < z \\ q & \text{if } y_i \geq z. \end{cases}$$

Note that the definition above is cyclic, but it is well-defined (see [6]). We can interpret the k-th component of the direct q-th QF-transforms as *"the q-th quantile of the set of values y's of those tuples $(x, y) \in \mathbf{T}$ such that x belongs to the class Δ_k."*

It is necessary to mention here that Definition 4 differs slightly from the original approach in the following way: in [6] the direct QF-transform is defined as a vector of fuzzy sets (or fuzzy numbers) whereas Definition 4 defines direct QF-transforms as scalar values in $\mathbb{R}$; one per each $q \in [0,1]$. Nevertheless, it is easy to prove that both definitions are equivalent in the sense that one can be obtained from the other and viceversa. The reason of considering this slight modification in the definition is because of the sake of presentation for the further use of q-th QF-transforms to define quantile functions of probability distributions.

Finally, note that each component of the direct q-th QF-transform can be obtained by means of weighted linear programming [6], therefore, the computation of direct QF-transforms can be efficiently performed in practice.

As the inverse standard F-transform, the inverse q-th QF-transform is a function from $\mathscr{U}$ to $\mathbb{R}$.

Definition 5. *Let $\mathbf{T} = \{(x_i, y_i)\}_{i\in\mathbb{I}} \subseteq \mathscr{U} \times \mathbb{R}$, let $q \in [0,1]$ and let $\mathbf{QF}_{\Delta}^{q}[\mathbf{T}] = [F_1, \ldots, F_n] \in \mathbb{R}^n$ be the direct q-th QF-transform of $\mathbf{T}$ w.r.t. Δ. Then, the function defined for all $x \in \mathscr{U}$ as*

$$\mathbf{T}_{\Delta}^{QF}(x, q) = \frac{\sum_{k=1}^{n} F_k \Delta_k(x)}{\sum_{k=1}^{n} \Delta_k(x)} \tag{6}$$

is called the q-th inverse *QF*-transform *of $\mathbf{T}$ w.r.t. Δ.*

[1] The minimizer of a function $f\colon \mathbb{R} \to \mathbb{R}$ is the value $z \in \mathbb{R}$ such that $f(z)$ is the minimum of f and, in this case, it may not be unique.

For each $x \in \mathscr{U}$ and $q \in [0,1]$, $\mathbf{T}_{\Delta}^{QF}(x,q)$ approximates the q-th quantile of the set $D_x = \{y \in \mathbb{R} \mid (x,y) \in \mathbf{T}\}$. Consequently, if we fix $x \in \mathscr{U}$ and consider $q \in [0,1]$ as a variable, then the mapping $\mathbf{T}_{\Delta}^{QF}(x,-)$ can be considered an approximation of the quantile function associated to the probability distribution of the set D_x. Moreover, note that the approximation given by $\mathbf{T}_{\Delta}^{QF}(x,q)$ takes into account no only those elements in $\mathbf{T}$ with first component equals to x, but also those elements with first component *"close"*[2] to x; as a result, the approximation is also applicable to those cases where $D_x = \emptyset$. This last property makes this approach suitable to approximate probabilities on environments affected by lack of information.

3. Constructing a probabilistic-fuzzy inference system based on QF-Transforms.

3.1. Defining a general Probabilistic Fuzzy Inference System (PFIS).

The different fuzzy inference systems in the literature differs mainly on the structure of fuzzy rules and on the procedure considered to perform the inference. However, in general, most of them has the following general pattern divided in four steps: Fuzzification, Knowledge database, Inference engine and Defuzification. Below we define a Probabilistic Fuzzy Inference System (PFIS) following those four steps, where antecedents of rules are fuzzy sets and consequents of rules are probability distributions. As a result, the output of the inference is a probability distribution.

- FUZZIFICATION. For the fuzzification we assume a given fuzzy partition Δ. For the sake of simplicity, we do not go in deep about it, and we refer to the interested reader to [10].
- KNOWLEDGE DATABASE. The *knowledge database* is formed by a set of rules If-Then. In each rule the antecedent is a fuzzy set (which may be composed by a set of fuzzy sets and fuzzy connectives) and the consequent is a probability distribution P given by its quantile function Q. As a simple example of a rule, let us consider three variables, b_1, b_2 and a; two fuzzy sets on the universe of the variables b_1 and b_2, denoted by B_1 and B_2, respectively; and a quantile function Q_a related to possible values of the variable a. Then, the rule R given by:

  $$R\colon \text{ IF } \quad B_1 \vee B_2 \quad \text{ THEN } \quad Q_a$$

  should be interpreted as *"if the input values of attributes b_1 and b_2 are in $B_1 \vee B_2$, then the variable a is distributed according to the probability distribution given by Q_a"*. Hereafter, and without loss of generalization, we omit the use of connectives in the antecedent.
- INFERENCE ENGINE. The knowledge database is useless without an *inference engine* that allows us to combine all the rules in only one inference. To describe the inference engine, let us consider a knowledge database $\mathbb{K}$ formed by the following n fuzzy rules

  $$\mathbb{K} = \left\{ R_k\colon \text{ IF } \quad B_k \quad \text{ THEN } \quad Q_k \right\}_{k \in \{1,\dots,n\}} \tag{7}$$

  where all the fuzzy sets B_k are defined on the same universe $\mathscr{U}$ that represent the input variables. Let us consider also an input value $u \in \mathscr{U}$. Then, the result of the inference engine with respect to the knowledge database $\mathbb{K}$ is the probability distribution associated to the quantile function $\mathbf{Q}$ defined for each $q \in [0,1]$ by:

  $$\mathbf{Q}(u,q) = \frac{\sum_{k=1}^{n} Q_k(q) \cdot B_k(u)}{\sum_{k=1}^{n} B_k(u)} \tag{8}$$

  Theorem 1. *Let $\mathbb{K}$ be a knowledge database as in Equation 7 and $u \in \mathscr{U}$, then the function $\mathbf{Q}(u,\text{-})$ given in Equation* (8) *defines a quantile function.*

[2] Where the relationship of closeness is given by the fuzzy partition Δ.

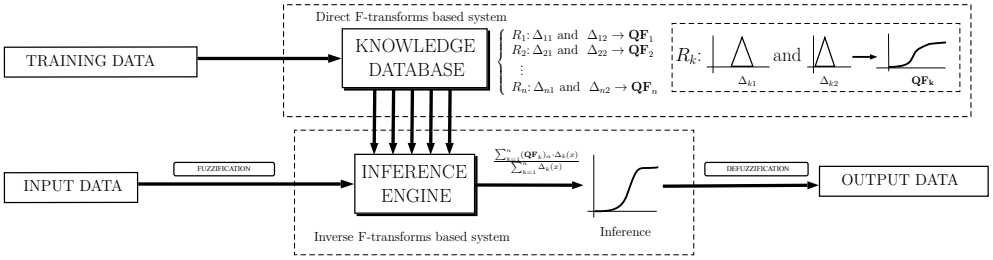

Figure 1. Diagram of the proposed probabilistic fuzzy inference system based on quantiles F-transforms.

Proof. To prove that Equation (8) defines a quantile function, we only have to show that $\mathbf{Q}(u,\text{-})$ is increasing and right continuous. Let us assume that $q_1 \leq q_2$. Since each Q_k is a quantile function, we have that $Q_k(q_1) \leq Q_k(q_2)$ for all $k \in \{1,\dots,n\}$. As a result,

$$\mathbf{Q}(u,q_1) = \frac{\sum_{k=1}^n Q_k(q_1)\cdot B_k(u)}{\sum_{k=1}^n B_k(u)} \leq \frac{\sum_{k=1}^n Q_k(q_2)\cdot B_k(u)}{\sum_{k=1}^n B_k(u)} = \mathbf{Q}(u,q_2)$$

Once we have proved that $\mathbf{Q}(u,\text{-})$ is increasing, the right continuity is proved by showing that $\mathbf{Q}(u,\inf_{i\in\mathbb{I}} q_i) = \inf_{i\in\mathbb{I}} \mathbf{Q}(u,q_i)$ for all subsets $\{q_i\}_{i\in\mathbb{I}} \subseteq [0,1]$:

$$\mathbf{Q}(u,\inf_{i\in\mathbb{I}} q_i) = \frac{\sum_{k=1}^n Q_k(\inf_{i\in\mathbb{I}} q_i)\cdot B_k(u)}{\sum_{k=1}^n B_k(u)} = \frac{\sum_{k=1}^n \inf_{i\in\mathbb{I}} Q_k(q_i)\cdot B_k(u)}{\sum_{k=1}^n B_k(u)} = \inf_{i\in\mathbb{I}}\left(\frac{\sum_{k=1}^n Q_k(q_i)\cdot B_k(u)}{\sum_{k=1}^n B_k(u)}\right) = \inf_{i\in\mathbb{I}} \mathbf{Q}(u,q_i)$$

□

Two remarks. Firstly, note that thanks to the use of quantile functions to describe probability distributions, the inference engine is defined for both, discrete and continuous sample spaces. Secondly, note that the result of the inference engine is not a fuzzy set but a probability distribution.

- DEFUZZIFICATION. In our approach this step is tricky by the nature of the PFIS. On the one hand, we may be interested in obtaining a crisp output according to the problem we are dealing with. Therefore, a step alike defuzzification to transform the probability distribution, obtained in the previous step, into a value (or into an interval) should be implemented. On the other hand, the output of the inference engine is not fuzzy, but probabilistic. Thus, the word "defuzzification" does not describe appropriately this task. In order to keep a coherent nomenclature with the existing fuzzy inference systems, we keep the name of defuzzification. Some possible defuzzifications are: the mean of the probability distribution obtained (as done in [27]); the median; an interval centered in the mean that encompass certain probability (like confidence intervals); or a quantile that determine an upper (or lower) bound of the output with certain probability. Actually, any descriptor of a probability distribution can be applied to perform this step. In this section we do not go into details in this step, since the defuzzification procedure must be chosen according to the final application of the PFIS.

3.2. Construction of a knowledge database by means of direct QF-transforms

In this section we describe a technique for the construction of probabilistic fuzzy rules in those terms described in Equation (7). The reader can easily observe that the inference engine (Equation (8)) of the PFIS is very similar to the inverse QF-transform. Hence, the use of direct QF-transforms to obtain the consequent of rules in the *knowledge database* seems to be a very convenient method to construct rules.

The full approach is summarized in Figure 1. Hereafter, we consider a training dataset $\mathbf{T} = \{(\mathbf{x}_i, y_i)\}_{i\in\mathbb{I}}$ of elements in $\mathbb{R}^m \times \mathbb{R}$; that is, we assume that $\mathscr{U} = \mathbb{R}^n$ and that for each tuple $(\mathbf{x},y) \in \mathbf{T}$ we have that $\mathbf{x} = (x_1,x_2,\dots,x_m)$. Each $((x_1,x_2,\dots,x_m),y) \in \mathbf{T}$ corresponds to one object (or instance) where each coordinate corresponds to a different

attribute (or variable), the attributes corresponding to $x_1, x_2, \ldots, x_m$ are called independent and the attribute corresponding to y, is called dependent. We fix now the set of antecedents $\Delta = \{B_k\}_{k\in 1\ldots n}$ that must form a fuzzy partition of the universe $\mathscr{U} = \mathbb{R}^m$. For each class in $B_k \in \Delta$ we define a probabilisitic fuzzy rule R_k where B_k is the antecedent of the rule and the consequent is the quantile function $\mathbf{QF}_k$ defined by the k-th component of the direct quantile F-transforms of $\mathbf{T}$ with respect to Δ (see Definition 4); i.e., $\mathbf{QF}_k(q) = \left(\mathbf{QF}^q_\Delta[\mathbf{T}]\right)_k$ for all $q \in [0,1]$. In this way, the knowledge database is formed by a set of rules with the form:

$$R_k\colon\ IF\ (x_1, x_2, \ldots, x_m) \text{ is in } B_k\ \text{ THEN }\ \mathbf{QF}_k(y). \tag{9}$$

In case the support of B_k wrt $\mathbf{T}$ is null, the respective rule is removed from the Knowledge database. Note also that, since each $\mathbf{QF}_k$ is a quantile function for each $k \in \{1, \ldots, n\}$, from Theorem 1 we can conclude that the inference engine described in Equation (8) produces a quantile function. Later on, in Section 4, we show that this technique is suitable to construct effective PFIS. That is, the result of the proposed PFIS approximates properly the probability distribution of the dependent variable conditioned to the knowledge of the independent variables.

The last step in the PFIS concerns the Defuzzification procedure. It depends strongly on the application problem and on the information we want to retrieve from the output probability distribution. Below, in Section 3.4, we analyze different Defuzzification procedures and its impact in the computation of rules.

3.3. Significance of rules

The technique for the construction of a knowledge database described in Section 3.2 creates one probabilistic fuzzy rule for each fuzzy set B_k in Δ even if there is no dependence between the antecedent (B_k) and the consequent. Such a feature may look strange at first view, but if we attend to the meaning of the probabilistic fuzzy rules we can conclude that there is nothing strange or wrong. The rule R: $IF\ B(\mathbf{x})\ THEN\ QF(y)$ in our paradigm states that if the attributes $\mathbf{x}$ of one hypothetic instance c is in B then, the attribute y distributes according to QF. Therefore, if there is no relation between the independent and dependent variables $\mathbf{x}$ and y, the quantile function QF would be very similar to the original quantile function of the variable y. Then, even in the case of independent variables, the rule is correct as long as it represents no relation. The issue treated in this section is not about whether the rules are correct or not (analyzed experimentally later in Section 4), but whether the rules provide significance or not. That is, in this section we search for an answer to the question: does a rule determine a real dependence between variables? For such a reason, in this section we present few measures that are applied on the probabilistic fuzzy rules in order to determine whether the rules obtained have certain significance or are meaningless.

Let us assume that

$$R_k\colon IF\ (x_1, x_2, \ldots, x_m) \text{ is in } \Delta_k\ \text{ THEN }\ \mathbf{QF}_k(y)$$

is one of the fuzzy rules constructed according to the dataset $\mathbf{T} = \{(x_i, y_i)\}_{i\in\mathbb{I}} \subseteq \mathbb{R}^n \times \mathbb{R}$ and the fuzzy partition $\Delta = \{\Delta_1, \ldots, \Delta_n\}$ of $\mathbb{R}^n$. Let Q be the quantile function of the probability distribution of the variable y_i in $\mathbf{T}$; i.e., the quantile function associated to the dataset $\{y_i \in \mathbb{R} \mid (x_i, y_i) \in \mathbf{T}\}$. The rule R_k states that the dependent variable y of those objects with dependent variables $x_1, x_2, \ldots, x_m$ in Δ_k distributes accordingly to the quantile function $\mathbf{QF}_k$. Then, the greater the difference between $\mathbf{QF}_k$ and Q, the more dependence between variables and the greater the significance of the rule R_k. Among the different approaches in the literature addressing the problem of defining distances between two probability distributions (in this case, given by quantile functions), we have considered the Kolmogorov-Smirnov distance estimator that leads to the well known Kolmogorov-Smirnov test [11, 23, 22]. The KS-distance is the maximum point-wise difference between two empirical probability distributions, and by following the same idea, we define the KS-significance measure of the rule R_k as the KS-distance between the cumulative probability distributions associated to $\mathbf{QF}_k$ and Q, that is:

$$\mathrm{sig}_{KS}(R_k) = \max_{x\in\mathbb{R}} |F_{\mathbf{QF}_k}(x) - F_Q(x)| \tag{10}$$

where $F_{\mathbf{QF}_k}$ and F_Q are the cumulative probability distributions associated to $\mathbf{QF}_k$ and Q, respectively.

Note that $\mathrm{sig}_{KS}(R_k) \in [0,1]$ and that the closer its value to 1, the more the difference between the distributions associated to $\mathbf{QF}_k$ and Q and therefore, the more significance of the rule R_k. The other estimator in the Kolmogorov-Smirnov test makes use of the so-called *p-value*, which is an estimator about the probability that two (different) empirical distributions comes from two (different) datasets distributed under the same probability distribution. This

estimator can be easily adapted to our approach by considering $\mathbf{QF}_k$ and Q as the empirical distributions associated to two datasets. Due to the lack of space and that the p-value is well-known by the scientific comunity, we do not go deeper in its description and the reader interested on such an estimator is referred to [9]. By the way, with respect to the this latter significance measure in our framework: the smaller the p-value, the more significance of the rule.

The two previous measures of significance ($\text{sig}_{KS}(R_i)$ and p-value) can be used to determine whether the rule R_k provides significant information or not, but they are useless if we wonder to know what kind of information does it represent. In order to solve such a limitation, we present below other measures of significance. Firstly, note that if the statistical dispersion associated to $\mathbf{QF}_k$ is lesser than the one of Q, then the (probability) uncertainty of the independent variable y decreases if the antecedents of the rule R_k hold. That information is really important if we are interested in knowing in what cases we can obtain concise and reliable outputs. We can measure the variation of probability distributions directly by quantiles. Specifically, given $q_1, q_2 \in [0,1]$ we define the uncertainty generated by R_k with respect to the quantiles q_1 and q_2 as:

$$\text{unc}(R_k; q_1, q_2) = \frac{\mathbf{QF}_k(q_1) - \mathbf{QF}_k(q_2)}{Q(q_1) - Q(q_2)}$$

Note that if $\text{unc}(R_k; q_1, q_2) < 1$, then the dispersion of $\mathbf{QF}_k$ is lesser than the dispersion of Q in the range of the quantiles q_1 and q_2. In such a case, the values of the dependent variable y in data satisfying the antecedent of R_k are more concentrated and we may conclude that R_k reduces the uncertainty of the attribute y. Conversely, if $\text{unc}(R_k; q_1, q_2) > 1$, then we may say that the rule R_k increases the uncertainty of the attribute y; i.e. the values of y are more sparse if the antecedent of R_k holds.

We could be also interested in knowing how is the variation of the attribute y (in the sense greater or smaller) according to the satisfiability of the antecedent of a rule. This variation can be estimated by computing a distance between the quantile functions $\mathbf{QF}_k$ and Q. In this paper, we propose to measure such a distance by an integral and then, we define the translational significance of R_k by

$$\text{sig}_T(R_k) = \int_0^1 \mathbf{QF}_k(q) - Q(q)\, dq \tag{11}$$

Note that if $\text{sig}_T(R_k)$ is significantly greater than 0 (resp. lesser than 0), it means that the attribute of the variable y is greater (resp. lesser) than usual when the antecedent of the rule R_k holds.

One application of the significance measures is the description of the information provided by the PFIS. Accordingly, the KS-significance measure and the p-value report which rules do not provide significant information and can be removed from the description. The measure of uncertainty reveals which rules reduce or increase the uncertainty of the dependent variable. This is very useful when we are interested in a concise and reliable output, since only rules with $\text{unc}(R_k; q_1, q_2)$ sufficiently close to 0 are suitable. Finally, the translational significance shows how much one rule increases or decreases the value of the dependent variable. This is very useful when causation is assumed in a PFIS. In this respect, the translational significance is capable to determine under which circumstances we can obtain the greatest (or lowest) value of the dependent variable. In order to illustrate the information provided by the significance measures we introduce below an example.

Example 1. *Let us assume that we have fixed a fuzzy partition $\Delta = \{\Delta_1, \Delta_2, \Delta_3\}$ and a training dataset* $\mathbf{T}$. *The knowledge database KB associated to such a fuzzy partition has three rules (one for each class in Δ):*

$$R_1 : \mathit{IF}\ \mathbf{x}\ \mathit{is\ in}\ \Delta_1\ \mathit{THEN}\ \ \mathbf{QF}_1(y)$$
$$R_1 : \mathit{IF}\ \mathbf{x}\ \mathit{is\ in}\ \Delta_2\ \mathit{THEN}\ \ \mathbf{QF}_2(y)$$
$$R_1 : \mathit{IF}\ \mathbf{x}\ \mathit{is\ in}\ \Delta_3\ \mathit{THEN}\ \ \mathbf{QF}_3(y)$$

Let us assume that the significance measures of those three rules are given in the following table:

	$\text{sig}_{KS}(R)$	$p\text{-value}(R)$	$\text{unc}(R)$	$\text{sig}_T(R)$
R_1	0.8	0.01	0.11	-12.1
R_2	0.1	0.25	0.81	-1.1
R_3	0.74	0.02	1.61	20.1

Firstly, let us focus on the KS-significance measure and p-value. In general, a p-value below 0.05 *is significant enough to ensure that the probability distribution of the variable y changes with respect to the original. In this respect, only rules* R_1 *and* R_3 *satisfy such a requirement. In other words, we cannot ensure that the independent variable* $\mathbf{x}$ *has some effect on the variable y when* $\mathbf{x}$ *is in* Δ_2 *and the only rules that report significant information are* R_1 *and* R_3*. We put our attention now to the significance measure* $\mathrm{unc}(R)$ *to know how is the information that* R_1 *and* R_3 *report. Since* $\mathrm{unc}(R_1) = 0.11$ *is considerably lesser than* 1*, we can conclude that if* $\mathbf{x}$ *is in* Δ_1*, then the values of y are more concentrated than in the original probability distribution, since the dispersion is lower. On the other hand, since* $\mathrm{unc}(R_3) = 1.61$ *is greater than* 1*, the values of the variable y are sparser when* $\mathbf{x}$ *is in* Δ_3*. In other words, if* $\mathbf{x}$ *is in* Δ_1 *we can be more precise in the output of the PFIS than if* $\mathbf{x}$ *is in* Δ_3*. Finally, we have that* $\mathrm{sig}_T(R_1) = -12.1$ *and* $\mathrm{sig}_T(R_3) = 20.1$*. Therefore, when* $\mathbf{x}$ *is in* Δ_1 *the values of y are lesser than usual and when* $\mathbf{x}$ *is in* Δ_3 *the values of y are greater than usual.* □

The information provided by the significance measures is useful for a human interpretation of rules, but it may be also used for some practical purposes. For example, it may be used for the adjustment of the PFIS to a practical problem. Specifically, one central point of a PFIS is the choice of an ad-hoc fuzzy partition, since the performance of the rest of steps depends directly on such a choice. The significance measures can be used to define an iterative procedure consisting in the following steps: choice a fuzzy partition - determine the knowledge database - compute the significance measures - choice a new fuzzy partition according to the significance measures - repeat. Accordingly, if one rule R is associate to a high p-value, that means that R does not report useful information. Since such a rule R is associated to a class in the chosen fuzzy partition then, the user may decide either to remove such a rule R by removing the corresponding class or to make the fuzzy partition thiner and split the class associated to R in different classes. The possibilities are tremendous and depend on the practical problem to solve. For such a reason, we have let this research line for future work.

3.4. A note about the computation of rules and the defuzzification procedure

It is convenient to take into account a limitation in the computation of rules in the proposed PFIS. A *"perfect"* computation of the quantile functions $\mathbf{QF}_k$ requires to compute a non-numerable number of quantiles (i.e., one for each q in the unit interval $[0,1]$), which is obviously imposible. This issue can be solved by computing simply a discrete approximation of each $\mathbf{QF}_k$ that considers a finite number of quantiles. For example, we can approximate $\mathbf{QF}_k$ by means of deciles (i.e., nine values with $q \in \{0.1, 0.2, \ldots, 0.9\}$) or percentiles (i.e., 99 values with $q \in \{0.01, 0.02, \ldots, 0.99\}$). The way this discretization is performed has a direct impact in the computational cost of the procedure (less quantiles, lighter computation) and also on the defuzzification step carried later on. Actually, the kind of defuzzification applied in the last step of the fuzzy inference system should be the central point in the election of the discretization process, since it may be used to avoid unnecessary computations. For example, let us assume that we are interested on returning a bounded interval where the 90% of data is contained. Then, we may only need to compute the 0.05 and 0.95 quantiles and to return the corresponding interval as output; so the defuzzification underlies in the implementation of the whole procedure. In Section 4.5 we deal with this issue in more detail for a practical implementation.

4. Experimental validation

The goal of this section is to validate experimentally all the assertions exposed in the previous section concerning the proposed PFIS. In [15] we have already shown that the result of the PFIS constructed by QF-transforms certainly approximates the probability distribution of the dependent variable. For such a reason, here we focus on the significance measures and the potential application of the proposed PFIS. The experiments presented in this section are the following:

- In Sections 4.1, 4.2 and 4.3 we show that the significance measures of the rules presented in Section 3.3 represent the information they pretend, that is, dependence/independence, more or less dispersion and greater/lesser values. For this task, we have considered synthetic data which suits better than real data because we can analyze the behavior of the PFIS under certain circumstances constructed ad-hoc. Accordingly, we have considered three different kinds of synthetic datasets that represent three different environments: in Sections 4.1 we consider independent attributes, Sections 4.2 functional dependent attributes and in Sections 4.3 partially dependent attributes.

	X uniform Y uniform		X uniform Y normal		X normal Y uniform		X normal Y normal		X normal Y exponential		X exponential Y uniform		X exponential Y exponential	
	sig_{KS}	p-val.	sig_{KS}	p-val.	sig_{KS}	p-val.	sig_{KS}	p-val.	sig_{KS}	p-val.	sig_{KS}	p-val.	sig_{KS}	p-val.
R_1	0.180	0.454	0.181	0.457	0.163	0.555	0.156	0.591	0.156	0.596	0.166	0.544	0.169	0.519
R_2	0.188	0.424	0.193	0.390	0.215	0.294	0.222	0.295	0.232	0.242	0.144	0.671	0.141	0.696
R_3	0.196	0.382	0.193	0.390	0.209	0.323	0.213	0.324	0.223	0.278	0.162	0.561	0.157	0.591
R_4	0.195	0.387	0.186	0.431	0.207	0.331	0.208	0.320	0.210	0.334	0.178	0.469	0.179	0.462
R_5	0.195	0.390	0.193	0.419	0.199	0.365	0.210	0.328	0.211	0.316	0.203	0.355	0.203	0.347
R_6	0.187	0.430	0.195	0.373	0.210	0.337	0.211	0.322	0.201	0.370	0.237	0.253	0.227	0.253
R_7	0.192	0.398	0.194	0.395	0.212	0.313	0.208	0.328	0.215	0.302	0.265	0.182	0.285	0.138
R_8	0.190	0.405	0.184	0.445	0.226	0.267	0.226	0.266	0.216	0.294	0.305	0.116	0.316	0.095
R_9	0.179	0.460	0.184	0.440	0.155	0.602	0.160	0.568	0.159	0.579	0.255	0.230	0.267	0.209
μ	0.189	0.414	0.189	0.416	0.200	0.376	0.202	0.371	0.203	0.368	0.213	0.376	0.216	0.368

Table 1. The KS-significance measures and the p-values concerning 7 experiments with independent variables; one value for each of the nine rules for experiment. Last row the mean of the KS-significance measure and the p-value for each family of dataset.

- Finally, in Sections 4.4 and 4.5 we use real data to show that the PFIS models the probability distribution as good as the naive approach but, additionally, PFIS provides much more useful information about the dependence of variables. Moreover, in Section 4.5 we present a simple application of the proposed PFIS for forecasting the electrical energy output of a power plant.

4.1. Significance measures for independent data

In order to show the performance under two independent variables, X and Y, we have constructed 9 different families of datasets. Each family contains 100 datasets of 200 entries constructed by considering in each variable one of the following continuous probability distributions: uniform distribution (with range between 0 and 10), normal distribution (mean 5 and random standard deviation between 1 and 10) or exponential distributions (with random mean between 1 and 10). In other words, in two datasets of the same family, the values of the variables X and Y follow the same probability distribution. In each dataset we have considered a uniform fuzzy partitions of 9 classes and then, for each dataset we obtain 9 rules (one for each class). For each rule, we have computed its respective measure of KS-significance sig_{KS} (Equation (10)) and its p-value. Table 1 shows the mean of those measures obtained for each of the nine classes in seven of those families of datasets. Note that in all cases but two, the KS-significance sig_{KS} is lesser than 0.3, which is a low value in terms of statistical significance; the two exceptions are 0.305 and 0.316 which are neither high. Moreover, the mean of the all p-values is 0.384 which is a high value (note that usually, in statistical hypothesis testing, the p-value should be lesser than 0.05 to reject the null hypothesis). Note that in almost all cases, the p-value is greater than 0.2; the four exceptions are reached in the case of considering in the variable X an exponential distribution and for rules R_7 and R_8. The reason of those low values may be the low cardinality of the classes Δ_7 and Δ_8 in the uniform partition considered for the construction of rules R_7 and R_8; i.e., only few data belonged to Δ_7 and Δ_8. Any case, the associated p-values are still greater than 0.05. Although omitted in Table 1 we have also computed the standard deviation of the obtaining measures. The standard deviations have been, in general, low in these experiments: below 0.05 for the sig_{KS} and below 0.2 for the p-value. These low values for the standard deviations mean that the obtained results are fairly representative.

As a conclusion of this first group of experiments, we may conclude that for independent data, the rules constructed with our approach do not establish any relationship between variables, as expected.

4.2. Significance measures for functional data

In the next group of experiments we consider functional data, that is, we consider a variable X and construct the variable Y by means of one function $f\colon \mathbb{R} \to \mathbb{R}$ and the equality $Y = f(X)$. In all cases, the variable X is uniformly distributed with range $[0,10]$. In the first experiment, we have constructed different datasets of 200 elements by considering six types of functions: straight lines ($y = a + bx$ with $a,b \in (0,10]$), parabolas with vertex 0 ($y = ax^2 + b$ with $a,b \in (0,10]$), parabolas with vertex in $x = 5$ ($y = a(x-5)^2 + b$ with $a,b \in (0,10]$), natural logarithms ($y = \log(ax+1)$ with $a \in (0,10]$), exponentials ($y = a^x$ with $a \in (0,1)$) and the trigonometric function sin ($y = a \cdot \sin(x)$ with $a \in (0,10]$). We have constructed 100 datasets of each type and we have computed the means of the KS-significances, p-values and the (interquartile) uncertainty measure $\text{unc}(R, 0.25, 0.75)$ for each rule constructed by considering a uniform fuzzy partition of 9 classes (i.e., 9 rules). The results are given in Table 2. It is worth mentioning that,

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
$ax+b$	p-value	$2.9\cdot10^{-10}$	$7.5\cdot10^{-8}$	$7.9\cdot10^{-6}$	$1.0\cdot10^{-4}$	$2.2\cdot10^{-4}$	$1.2\cdot10^{-4}$	$2.4\cdot10^{-5}$	$2.0\cdot10^{-7}$	$8.8\cdot10^{-11}$
	sig_{KS}	0.8168	0.7116	0.6194	0.5254	0.4732	0.5128	0.6054	0.7094	0.8197
	$\text{unc}(R)$	0.1496	0.1127	0.1149	0.1163	0.1136	0.1205	0.1161	0.1189	0.1459
ax^2	p-value	$6.8\cdot10^{-11}$	$6.8\cdot10^{-7}$	$1.3\cdot10^{-5}$	$1.1\cdot10^{-4}$	$2.7\cdot10^{-4}$	$1.7\cdot10^{-4}$	$1.2\cdot10^{-5}$	$1.5\cdot10^{-7}$	$3.3\cdot10^{-11}$
	sig_{KS}	0.822	0.7108	0.616	0.5232	0.4707	0.5153	0.6143	0.7156	0.822
	$\text{unc}(R)$	0.0224	0.047	0.0719	0.0956	0.1236	0.1398	0.1618	0.192	0.2865
$a(x-5)^2$	p-value	$7.3\cdot10^{-7}$	$1.1\cdot10^{-3}$	$1.1\cdot10^{-3}$	$2\cdot10^{-6}$	$7.4\cdot10^{-11}$	$3\cdot10^{-6}$	$1\cdot10^{-3}$	$1.1\cdot10^{-03}$	$7.6\cdot10^{-7}$
	sig_{KS}	0.6774	0.4602	0.4548	0.6503	0.8245	0.6537	0.4588	0.4603	0.6705
	$\text{unc}(R)$	0.5162	0.2692	0.1835	0.09	0.0177	0.0929	0.1823	0.279	0.4907
$\ln(ax+1)$	p-value	$1.3\cdot10^{-10}$	$1.2\cdot10^{-6}$	$1.1\cdot10^{-5}$	$9.8\cdot10^{-5}$	$2.1\cdot10^{-4}$	$1.1\cdot10^{-4}$	$8.8\cdot10^{-6}$	$1.2\cdot10^{-7}$	$1.7\cdot10^{-10}$
	sig_{KS}	0.8142	0.7079	0.6075	0.5154	0.472	0.5172	0.6128	0.7143	0.8192
	$\text{unc}(R)$	0.9975	0.2664	0.1757	0.1335	0.1037	0.0836	0.0756	0.0642	0.0704
$a^x(a>1)$	p-value	$6\cdot10^{-11}$	$1\cdot10^{-7}$	$1.6\cdot10^{-5}$	$1.2\cdot10^{-4}$	$2.3\cdot10^{-4}$	$1.2\cdot10^{-4}$	$2\cdot10^{-5}$	$4.9\cdot10^{-7}$	$4.5\cdot10^{-10}$
	sig_{KS}	0.8204	0.7142	0.6128	0.5215	0.4737	0.5186	0.6094	0.7123	0.8195
	$\text{unc}(R)$	0.01559	0.0197	0.0281	0.0432	0.0808	0.1497	0.3299	0.7214	2.5071
$a^x(a<1)$	p-value	$1\cdot10^{-10}$	$1.7\cdot10^{-7}$	$2.5\cdot10^{-5}$	$1\cdot10^{-4}$	$2.5\cdot10^{-4}$	$1.2\cdot10^{-4}$	$1\cdot10^{-5}$	$1\cdot10^{-7}$	$5.3\cdot10^{-11}$
	sig_{KS}	0.8189	0.7039	0.6064	0.5168	0.4712	0.5257	0.6267	0.7372	0.8452
	$\text{unc}(R)$	141.18	2.3782	0.3677	0.117	0.0608	0.0401	0.0255	0.0209	0.0191
$a\cdot\sin(x)$	p-value	$1.1\cdot10^{-2}$	$9.7\cdot10^{-7}$	$8.2\cdot10^{-2}$	$2.3\cdot10^{-6}$	$1.2\cdot10^{-9}$	$5.5\cdot10^{-3}$	$1.5\cdot10^{-4}$	$4.3\cdot10^{-4}$	$4.4\cdot10^{-3}$
	sig_{KS}	0.4186	0.5815	0.3275	0.6825	0.7908	0.4949	0.4099	0.6768	0.2844
	$\text{unc}(R)$	0.4424	0.1896	0.4767	0.3056	0.1318	0.4479	0.3517	0.1051	0.5871

Table 2. The KS-significance measures, the p-values and uncertainty measures concerning 7 experiments with independent variables; one value for each of the nine rules for experiment.

although omitted in the table, the standard deviations of the p-values have been considerably low in all cases (in general below 10^{-4} but two cases which was below 10^{-3}), so the parameters of the corresponding functions used to construct the datasets do not have a significant impact. The reader can check directly in Table 2 that we have obtained low p-values and high sig_{KS} in all cases, which means that all the rules obtained provide significant information. In other words, the fuzzy inference rules are able to capture the dependence between functional variables, as expected. The reader can also observe that, in the most cases, the p-values and sig_{KS} values are greater in rules related to extremes classes than in rules related to central classes (i.e., greater for the rules R_1 and R_9). Such a feature is due to the monotonicity of the mappings considered to create the datasets. Exceptions of such a behavior are the datasets constructed by means of parabolas with vertex in $x=5$ and sin, where the p-value oscillates according to the function.

Once the dependence and the significance of the constructed rules has been confirmed, we can analyze the type of information reported by the rules. For such a reason, we have computed in all the previous experiments also the uncertainty measure $\text{unc}(R,0.25,0.75)$. The reader can see in Table 2 that the uncertainty measure $\text{unc}(R,0.25,0.75)$ depends on the function used to define each family of datasets. Actually, there is an observable relationship between the value of the uncertainty measure $\text{unc}(R,0.25,0.75)$ and the first derivative of the respective function. This measure is easily interpretable in the datasets $Y=a^X$ for $a\in(0,1)$. In such a case, elements with X closer to 0 (i.e., in R_1 where the antecedent is x in Δ_1) have a dispersion 141.18 times greater than usual and elements with X closer to 10 (i.e., in R_9 where the antecedent is x in Δ_9) have a dispersion 0.0191 times greater than usual (note that a value lesser than 1 implies a reduction of the dispersion). Some brief remarks: in almost all cases the dispersion is reduced ($\text{unc}(R)<1$); in straight lines the uncertainty measure is almost constant for all rules; and in parabolas the minimum uncertainty measure is always associated to the vertex (in our cases, either 0 or 5).

The second experiment with functional data aims at showing the information provided by the translational significance sig_T. That measure has been avoided in the previous experiment because it is highly dependent on parameters used in the functions (e.g., straight lines with considerably different slopes provide considerably different translational significance sig_T). In this experiment we have considered 50 datasets where X is uniformly distributed with range $[0,10]$ and $Y=2X+4$. We have considered again a uniform fuzzy partition of 9 classes and we have obtained the

following translational significance measures:

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
$\text{sig}_T(R)$	-8.69	-6.45	-4.48	-2.5	-0.62	1.21	3.13	5.03	7.3

From the previous values, we can infer that the greater the X the greater the variable Y, since they are increasing. In more detail, the value associated to R_1 says that when X is in Δ_1 (i.e., x is close to 1) then the variable Y has approximately a value 8.69 times lesser than the mean[3]. On the other extreme, the value associated to the rule R_9 states that when X is in Δ_9 (i.e., x is close to 9) then the variable Y has approximately a value 7.3 times greater than the mean.

The two final experiments with functional data aim at showing the importance of the number of entries in the training dataset and the number of classes (rules) considered to construct PFISs. We have considered three families of randomly generated datasets where X is uniformly distributed with range $[0,10]$ and $Y = \sin(X)$. We have chosen the function sin because the results are more illustrative, although the consideration of different functions reports similar conclusions. In the first family we have considered 30 datasets with 50 entries, in the second 30 datasets with 100 entries and in the third, 30 datasets with 1500 entries. In the procedure of rule construction, we have varied the number of classes from 3 to 22 and then, for each case, we have computed the means of the p-values, KS-significance and uncertainty measure. Figure 2 shows the evolution of those values according to the number of classes considered. Concerning the number of classes in the fuzzy partition, we can say that the more classes in the fuzzy partition, the greater the KS-significance. However, although the increment of classes in the fuzzy partition roughly reduces the p-values and the uncertainty measures, we can find some oscillations. That oscillations appear independently on the function used, but it looks that the more local maxima and local minima in the function, the more number of oscillations. Any case, note that the mean of p-values and the uncertainty measure tends to 0 when we increase the number of classes. Concerning the number of entries, note that the results for datasets with 50 entries and 1500 are very similar, so we can conclude that, for retrieving the information contained in functional data by means of fuzzy inference rules, we only need very few amount of data.

4.3. Significance measures in non-functional data

Now we consider non-functional data i.e., it may exist for the same $x \in X$ two values in $y_1, y_2 \in Y$ such that the tuples (x, y_1) and (x, y_2) belongs to the dataset. We present the results of 11 experiments that illustrate the effect of noise in functional data and the behaviour of the approach when it is applied to partially dependent data and/or data constructed from the union of datasets (as if it would come from different sources). In all experiments we have created 50 datasets where X is uniformly distributed with range $[0,10]$ and contains 200 entries. In order to show the effect of noise in functional data, we have considered the normal distribution $N(0,\sigma)$ for different standard deviations and considered 6 datasets where either $Y = sin(X) + N(0,\sigma)$ with $\sigma \in \{0.1, 0.5, 1\}$ or $Y = X + N(0,\sigma)$ $\sigma \in \{1,3,5\}$. The results are displayed in Table 3. As expected, with a (relative) low value for σ, the results of these experiments are similar to the results obtained for functional data (see Table 2). In contrast, since the greater the value of σ, the more similar the data to independent data, then the results for high values of σ are more similar to those obtained for independent data (see Table 1). Nevertheless, note that that even for considerably high standard deviations (note that $\sigma = 1$ and $\sigma = 5$ are the half of the range of $\sin(x)$ and x, respectively), the p-values for some rules (i.e., for some classes) are significantly low; that indicates a dependence between variables.

In order to analyze the partial dependence between variables, we have created also a set of datasets where the noise depends also on X. In Table 3 we show the results for $Y = X + X^2 N(0,10)$ and $Y = X + (X-5)^2 N(0,10)$. In the former dataset, the noise is incorporated into $Y = X$ proportionally to the value of X^2, therefore, the noise is lesser for values closer to $X = 0$. In this experiment, the p-values for the rules R_1 and R_9 have a significant low value. In the case of R_1, the reason is because the associated entries (with X close to 0) are infected with a low amount of noise. In the case of R_9 the reason is the opposite, the entries have been infected with a big amount of noise and then, the distribution is very different from the rest of data. Taking a look to the measures of uncertainty for R_1 (0.1324) and R_9 (5.0108), we have that the data with X close to 0 have very concentrate values for the variable Y whereas the variable Y for X close to 9 have very dispersed values; that indicates how the noise has infected the data.

[3]The mean of the variable Y in the original dataset, i.e., the mean of $\{y \mid (\mathbf{x}, y) \in \mathbf{T}\}$.

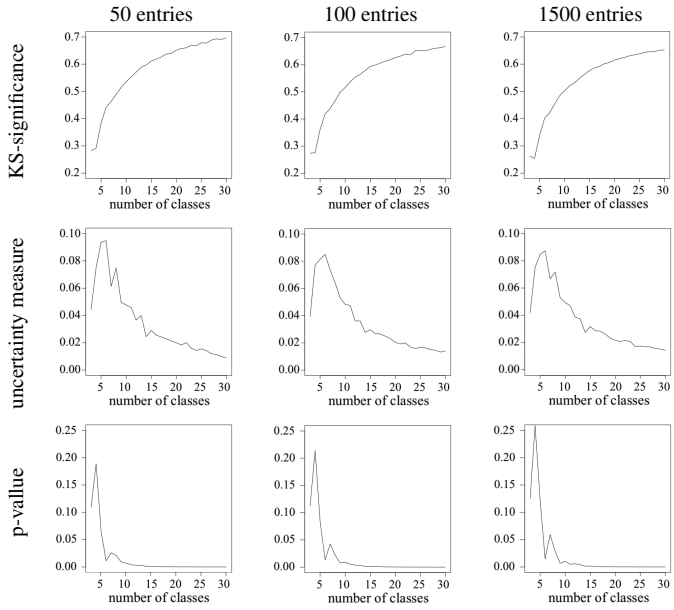

Figure 2. Evolution of the KS-significance, uncertainty measure and p-value for functional data with respect of the number of classes considered in the independent variable. There are three kinds of datasets according to the number of entries: 50, 1000 and 1500.

In the other family of datasets, the one with $Y = X + (X-5)^2 N(0,10)$, the noise is incorporated proportionally to $(X-5)^2$ and then, the closer X to 5, the smaller the noise in Y. As in the previous experiment, the lowest values for p-values are reached for the classes where the noise is minimal (i.e., R_5 where the antecedent is "X close to 5") and those where the noise is maximal (i.e., R_1 and R_9). The explantation of this behavior is exactly the same than in the previous family of datasets. Moreover, taking a look to the measures of uncertainty for R_1, R_5 and R_9, it is clear that such a measure is capable to determine how is the local dispersion of data: sparser when X is in Δ_1 or Δ_9 but more concentrate when X is in Δ_5.

In the following three experiments we analyze the behaviour of the fuzzy inference system when it is applied to non-functional dependent data. We have considered three different structures with three different shapes. The first dataset is constructed by means of $Y = Y_1 \cup Y_2$, where $Y_1 = X$ and $Y_2 = X + 4$, which create two parallel straight lines. The second dataset is constructed by means of $Y = Y_1 \cup Y_2$, where $Y_1 = 10X$ and $Y_2 = -10X + 10$, which create a crux. The third dataset considers pairs (Y_1, Y_2) where $Y_1 = \sin(X)$ and $Y_2 = \cos(X)$, which create a circle (in this last case $X \in [0, 2\pi]$ instead of $X \in [0, 10]$). In all the cases, we have considered 50 datasets and we have displayed the means of the results. Note that the p-values are low enough to consider that there is a dependence between the variables. Note that the uncertainty measure for the parallel lines is almost constant in all rules, which implies that the dispersion of data is similar along the axis X. On the other hand, the measures of uncertainty clearly show where dispersion is greater for the "crux" and "circle" datasets: in the extremes for the crux and in the middle for the circle.

4.4. PFIS constructed from real data

In this and in the subsequent section we make use of real data in order to show the potential capabilities of the proposed construction of PFIS. We use the uci-combined-cycle-power-plant dataset[4] which was analyzed in [26] under different regression models. The dataset contains 5 hourly average variables, namely: Ambient Temperature (AT), Ambient Pressure (AP), Relative Humidity (RH), Exhaust Vacuum (V) and electrical Energy outPut (EP). The dataset contains 9568 instances collected from a Combined Cycle Power Plant over 6 years (2006-2011).

[4] available in https://archive.ics.uci.edu.

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
$\sin+N_{0,1}$	p-value	$6.8\cdot10^{-3}$	$1.4\cdot10^{-5}$	$3.2\cdot10^{-2}$	$1.9\cdot10^{-7}$	$2.1\cdot10^{-10}$	$8.4\cdot10^{-4}$	$4.5\cdot10^{-3}$	$2.6\cdot10^{-7}$	$8\cdot10^{-2}$
	sig_{KS}	0.3998	0.5812	0.3237	0.6753	0.7814	0.4709	0.4158	0.6624	0.2753
	$\text{unc}(R)$	0.4431	0.2075	0.4785	0.3232	0.1613	0.4780	0.3731	0.1551	0.5824
$\sin+N_{0,5}$	p-value	$2.8\cdot10^{-2}$	$1.2\cdot10^{-3}$	$2\cdot10^{-1}$	$3.2\cdot10^{-5}$	$1.5\cdot10^{-6}$	$9.9\cdot10^{-3}$	$1.6\cdot10^{-2}$	$4.4\cdot10^{-4}$	$2.5\cdot10^{-1}$
	sig_{KS}	0.3458	0.4593	0.233	0.5716	0.6524	0.3746	0.3558	0.4964	0.2205
	$\text{unc}(R)$	0.6736	0.5899	0.6895	0.5657	0.3874	0.7115	0.6376	0.561	0.719
$\sin+N_1$	p-value	0.1655	0.0568	0.3483	0.0036	0.001	0.1171	0.1306	0.0135	0.4111
	sig_{KS}	0.2679	0.3367	0.2013	0.4272	0.4699	0.2829	0.273	0.3785	0.1869
	$\text{unc}(R)$	0.8832	0.879	0.8834	0.8322	0.8636	0.8535	0.841	0.8186	0.9464
$X+N_1$	p-value	$1.4\cdot10^{-8}$	$4\cdot10^{-6}$	$3.9\cdot10^{-4}$	$2.9\cdot10^{-3}$	$6.9\cdot10^{-3}$	$3\cdot10^{-3}$	$3.8\cdot10^{-4}$	$8\cdot10^{-6}$	$4.2\cdot10^{-9}$
	sig_{KS}	0.7435	0.626	0.5324	0.4323	0.3742	0.4221	0.5178	0.6146	0.7345
	$\text{unc}(R)$	0.2938	0.2868	0.2927	0.2913	0.2929	0.2939	0.2789	0.2927	0.3023
$X+N_3$	p-value	$1.7\cdot10^{-4}$	$2.6\cdot10^{-3}$	$2.2\cdot10^{-2}$	$1.4\cdot10^{-1}$	$2.6\cdot10^{-1}$	$1.4\cdot10^{-1}$	$2\cdot10^{-2}$	$4.1\cdot10^{-3}$	$8.9\cdot10^{-5}$
	sig_{KS}	0.5251	0.4383	0.3596	0.2652	0.2164	0.2612	0.3691	0.4372	0.539
	$\text{unc}(R)$	0.693	0.7264	0.6972	0.6746	0.6945	0.7071	0.6972	0.7345	0.7323
$X+N_5$	p-value	0.0061	0.0319	0.1581	0.2656	0.4142	0.3195	0.1166	0.0287	0.0042
	sig_{KS}	0.4117	0.3505	0.2739	0.2205	0.1879	0.209	0.279	0.3474	0.417
	$\text{unc}(R)$	0.87	0.8587	0.8737	0.8634	0.8643	0.8636	0.8627	0.858	0.9177
$X+X^2N_{10}$	p-value	0.0003	0.0096	0.07664	0.2244	0.2639	0.1399	0.0707	0.0328	0.01187
	sig_{KS}	0.4574	0.3624	0.2789	0.2252	0.2229	0.2585	0.2892	0.3279	0.3788
	$\text{unc}(R)$	0.1324	0.4048	0.8203	1.0239	1.3994	2.0125	2.9625	4.0109	5.0108
$X+(X-5)^5N_{10}$	p-value	0.024	0.0987	0.2709	0.0495	0.001	0.0362	0.2696	0.0832	0.022
	sig_{KS}	0.3425	0.2766	0.2177	0.2996	0.426	0.3091	0.2229	0.2923	0.3526
	$\text{unc}(R)$	2.5227	1.0342	0.265	0.0845	0.0259	0.2565	1.0954	2.6166	4.5636
Parallels	p-value	$3.9\cdot10^{-7}$	$3.2\cdot10^{-5}$	$8.2\cdot10^{-4}$	$1.6\cdot10^{-2}$	$7.5\cdot10^{-2}$	$2.9\cdot10^{-2}$	$2.5\cdot10^{-3}$	$9.3\cdot10^{-5}$	$1\cdot10^{-6}$
	sig_{KS}	0.6509	0.5451	0.4549	0.352	0.2769	0.3287	0.431	0.5276	0.6311
	$\text{unc}(R)$	0.8039	0.8039	0.8039	0.8039	0.8039	0.8039	0.8039	0.8039	0.8039
Crux	p-value	0.0036	0.0744	0.0841	0.0068	0.0002	0.0065	0.0837	0.0857	0.0042
	sig_{KS}	0.3892	0.2775	0.2722	0.3746	0.4616	0.3751	0.2714	0.2714	0.3862
	$\text{unc}(R)$	1.6989	1.1989	0.7797	0.4117	0.1135	0.3862	0.7978	1.1785	1.6682
Circle	p-value	0.0093	0.0669	0.0078	0.0005	0.0001	0.0005	0.0103	0.1011	0.0181
	sig_{KS}	0.3816	0.2978	0.4128	0.5131	0.5329	0.4976	0.3918	0.271	0.3528
	$\text{unc}(R)$	0.4952	1.1952	1.3301	1.4453	1.4609	1.4329	1.3287	1.2021	0.5209

Table 3. The KS-significance measures and the p-values concerning 7 experiments with independent variables; one value for each of the nine rules for experiment. Last row the mean of the KS-significance measure and the p-value for each family of dataset.

sig	Δ_1^V	Δ_2^V	Δ_3^V	Δ_4^V	Δ_5^V
Δ_1^{AT}	28.75	19.32	11.28	4.45	\
Δ_2^{AT}	27.69	17.66	7.43	-5.75	-14.79
Δ_3^{AT}	27.48	10.82	-1.38	-8.90	-14.85
Δ_4^{AT}	\	0.26	-6.68	-14.02	-19.05
Δ_5^{AT}	\	-4.82	-13.38	-18.63	-20.89

unc$(R_1, 0.1, 0.9)$	Δ_1^V	Δ_2^V	Δ_3^V	Δ_4^V	Δ_5^V
Δ_1^{AT}	0.2845	0.2829	0.1964	0.1502	\
Δ_2^{AT}	0.2568	0.3071	0.3037	0.3345	0.0151
Δ_3^{AT}	0.2916	0.2612	0.3084	0.2778	0.1314
Δ_4^{AT}	\	0.1855	0.2704	0.2597	0.2449
Δ_5^{AT}	\	0.3643	0.2313	0.1851	0.1714

Table 4. Translational and uncertainty measures for the 25 rules of the PFIS obtained considering V and AT as independent variables.

In the experiment presented in this section, we compute PFIS associated to the deciles and compare them with the results obtained with the naive approach. We call naive approach to the one that estimates directly the quantile distribution of the dependent variable without taking into account the independent variables. Obviously, the naive approach describes an approximation of a probability distribution of the electrical energy output (EP). The question we answer in this section is: how much better are the results of the PFIS compared to the naive approach to define a quantile function. To do such a comparison we have used 1000 random instances of the dataset for the training step and the rest for testing, that is 7568 instances for testing. We have chosen the variables AT and V as independent variables and a uniform fuzzy partition of 25 classes (5 clases for each independent variable). In order to illustrate how good both approaches model a probability distribution, we have computed the number of entries with EP below each quantile and the respective "quantile" obtained by the PFIS; the closer the value to the respective quantile the better the approximation. The results can be seen in Table 5 and the reader can observe that the naive approach provides a better performance. However, taking into a look that the differences between the training and testing datasets in the results corresponding to the PFIS, we may consider a readjust of the PFIS and compute the percentiles that in the training phase, provide the most accurate result to the deciles. For example, the results of the PFIS associated to the percentile 0.17 provides an accuracy in the training dataset of 0.107%. Therefore, it seems more appropriate to consider the percentile 0.17 to compute the first decile in the PFIS. In Table 5, we have included also the results corresponding to this variation and the results are more accurate than with the standard PFIS and very similar to those obtained by the naive approach.

Once we have seen that the PFIS performs an approximation of a quantile function as good as the naive approach, it is natural to wonder what is the advantage of using PFIS instead of the naive approach. The answer is given by the significance measures. The PFIS provides information about when the PE is greater/lower (by means of the translational significance measure) and when the uncertainty about the value of PE is reduced/increased (by means of the significance measure unc). Table 4 shows the values of those measure for each rule; we have chosen the deciles 1 and 9 for the uncertainty unc and the bar denotes rules without support. Observing the values of sig, the reader can quickly note that the smaller the values of AT and V, the greater the EP. On the other hand, in the right table, the reader can observe that the uncertainty is always reduced by the application of the PFIS; so the forecasted values of EP by the PFIS are more precise than those forecasted by the naive approach.

4.5. Brief application to forecast electrical energy output

In the last section, we make use of the uci-combined-cycle-power-plant dataset considered in the previous section and apply the PIFS to forecast a range of values of the electrical energy output (EP) with an accuracy of the 95%; i.e., the 95% of data will belong to the inferred interval of values. Hence, we have considered EP as the dependent variable and the rest as independent; so the universe is $\mathbb{R}^5$. For each independent variable we have considered 5 uniform fuzzy partitions (in the range of each variable) with 5 clases each and we have combined all them by means of the cartesian product to obtain a fuzzy partition in $\mathbb{R}^5$ of 625 classes. In the defuzzification procedure, we have considered the interval obtained by the quantiles 0.025 and 0.975, which is expected to contain the 95% of data.

For the proper implementation of the experiment, we have chosen 3000 random instances of the dataset for the training step and the rest for testing, that is 6568 instances for testing. Since in the defuzzification step we will consider only the quantiles 0.025 and 0.975, we have only computed the QF-transforms associated to them and we have obtained a PFIS with 625 rules of which only 439 have support different from 0 (i.e., the associated classes are not empty). Once the rules had been computed, we have done a double testing, one with the training dataset and

		Naive	PFIS	PFIS readjusted
D1	Training dataset	10%	6.3%	10.7%
	Testing dataset	9.61%	5.64%	10.8%
D2	Training dataset	20%	14.4%	20.4%
	Testing dataset	17.95%	13.67%	19.85%
D3	Training dataset	30%	24.5%	29.9%
	Testing dataset	28.66%	24.58%	29.63%
D4	Training dataset	40%	36.8%	40.2%
	Testing dataset	38.44%	36.94%	40.26%
D5	Training dataset	50%	50.1%	50.1%
	Testing dataset	49.05%	54.84%	49.78%
D6	Training dataset	60%	64.5%	60.5%
	Testing dataset	59.99%	67.41%	59.08%
D7	Training dataset	70%	75.1%	70%
	Testing dataset	69.58%	73.57%	69.41%
D8	Training dataset	80%	87%	80.3%
	Testing dataset	78.69%	85.63%	78.57%
D9	Training dataset	90%	95.5%	90.3%
	Testing dataset	88.79%	94.72%	88.94%

Table 5. Comparison between the accuracy of the PFIS and naive approach with respect to deciles.

		Accuracy		Error
	quantiles	Training Dataset	Testing Dataset	
QF-Transforms	0.025 and 0.975	98.13 %	97.97 %	2.97%
	0.0563 and 0.9437	95.03%	94.97	0.06%
Naive approach	0.025 and 0.975	95 %	95.64 %	0.64%

Table 6. Accuracy of the PFIS and naive approach obtained for forecasting the energy output on both, the testing datase and on the training dataset.

another with the testing dataset. In both cases, we have introduced the independent variables of the testing dataset in the PFIS and we have checked whether the output is contained in the interval determined by the QF-transforms associated to the quantiles 0.025 and 0.975 or not. The results are in Table 6. As the reader can observe, the accuracy obtained is 98.13% in the training dataset, more than 3% greater than expected. It is good enough, but if we really pretend to get the 95% of data, we should adjust the system choosing different quantiles. If we compute the QF-transforms associated to the quantiles 0.0563 and 0.9437 (instead of those for the quantiles 0.025 and 0.975), the obtained accuracy is 95.03% for the training dataset, which is closer to the goal of 95%. When we analyze the results in the testing dataset, we observe that the results are very similar to the one obtained in the training dataset. The error is very small, the difference between the obtained percentages is 0.16% and 0.06% when we consider the pairs of quantiles $(0.025, 0.975)$ and $(0.0563, 0.9437)$, respectively. As a conclusion, the output of the PIFS, with the pair of quantiles $(0.0563, 0.9437)$, is an interval that contains the 95% of real electrical energy output values with a very good accuracy, as we expected to obtain. For the sake of a better interpretation of the results, we have compared the results with the naive approach. In Table 6 the reader can see that the accuracy obtained for PIFS (for the quantiles 0.0563 and 0.9437) is even better than the one obtained by the naive approach. As a result, we can say that the proposed PIFS satisfactorily determines the probability distribution of the dependent variable electrical Energy outPut (EP) conditioned to the information of the independent variables Ambient Temperature (AT), Ambient Pressure (AP), Relative Humidity (RH), Exhaust Vacuum (V).

entry number	AT	V	AP	RH	PFIS interval for EP	Naive interval for EP	Real EP
#1748	14.48	46.18	1015.83	92.15	$\big(453.97\ ,\ 476.19\big)$	(430.64, 482.83)	462.30
#2548	29.72	54.20	1012.86	38.48	(436.35, 452.11)	(430.64, 482.83)	438.84
#4050	24.73	69.05	1003.25	66.44	(429.38, 445.62)	(430.64, 482.83)	437.93
#5789	9.30	43.14	1010.99	88.12	(464.21, 485.26)	(430.64, 482.83)	477.02
#6371	27.01	43.21	1011.71	77.84	(435.35, 457.41)	(430.64, 482.83)	434.09
#7045	5.73	40.35	1012.24	91.84	(471.06, 490.72)	(430.64, 482.83)	490.50

Table 7. Some results of the PFIS and the naive approach for different entries of the testing dataset.

As in the previous section, it is natural to wonder what is the difference between PFIS and the naive approach. To answer this doubt, we take a look into the information provided by the corresponding uncertainty and translation measures. These two measures compare the results of the PFIS with the naive approach. Let us recall that the uncertainty measure compares the dispersion of the data distribution in the naive approach with the one of each rule in the PFIS. The mean of the uncertainty measures is 0.2242 (with respect to the quantiles 0.0563 and 0.9437), which means that the PFIS is capable to reduce the uncertainty around 2.4 times the one resulted by the naive approach. The rule with the maximum uncertainty measure is 0.5, which is the worst case and even so, it is capable to reduce the uncertainty to the half. There are 33 rules with an uncertainty measure below 0.1, which reduce 10 times the uncertainty of the original data. In order to illustrate better this reduction of uncertainty, in Table 7 we have included the result of the PFIS for different entries in the testing dataset. The reader can see how the range of the forecasted energy output is reduced in all those cases; that is because all the uncertainty measures are lesser than 1.

Concerning the translation measures, there are 198 rules that estimate a reduction of electrical energy output (EP) (a translation measure lesser than 0) and 241 rules that estimate an increment (a translation measure greater than 0). Such a piece of information is useful in the case we pretend to determine under which circumstances the energy output increases or decreases.

In summary, the proposed PFIS is capable to forecast the electrical output from a set of data with an accuracy given. In our example, we have prefixed an accuracy of 95% and in the testing dataset we have obtained a final accuracy of 94.97% (only 0.03% of error); i.e., in a very similar way we would approximate the distribution with the naive approach. The advantage of using the PFIS instead of the naive approach is that we can reduce the uncertainty and forecast greater or lesser energy output according to the information we have from the independent variables.

5. Related Works

This article is related to two different families of approaches, those based on F-transforms and those dealing with Probabilistic Fuzzy Inference Systems (PFIS). Concerning the former, the reader can find in the literature articles that propose the use of standard F-transforms to construct Fuzzy Inference Systems [21]; but none of them involve probability in the inference. To incorporate the uncertainty related to probability in our approach, we have used QF-transforms (or L_1-F-transforms) introduced originally in [6] which have been recently applied to forecasting and Bitcoin Analysis in [7]. There are two main differences between those two last mentioned articles and our approach. Firstly, in [6] the QF-transforms are joined forming fuzzy numbers but in our approach they are joined forming probability distributions. This modification is simple but also crucial to define a PFIS from QF-transforms. Secondly, the approaches of [6, 7] present a whole process that does not focus on the values obtained in the direct QF-transforms but on the results obtained by composing direct and inverse QF-transforms. In our approach we put a lot of focus on analyzing those values obtained in the direct QF-transforms. Specifically, the significance measures presented in our approach consider only the results of direct QF-transforms, since they are the kernel of the constructed probabilistic fuzzy rules. Thank to those rules based on direct QF-transforms and the respective significance measures, we can analyze data from a different perspective than in [6]. For example, in Section 4.5, we have shown that the uncertainty and translation measures of each rule allow us to determine under which circumstances the forecasting concerning the energy power is more precise (resp. vague) and higher (resp. lower).

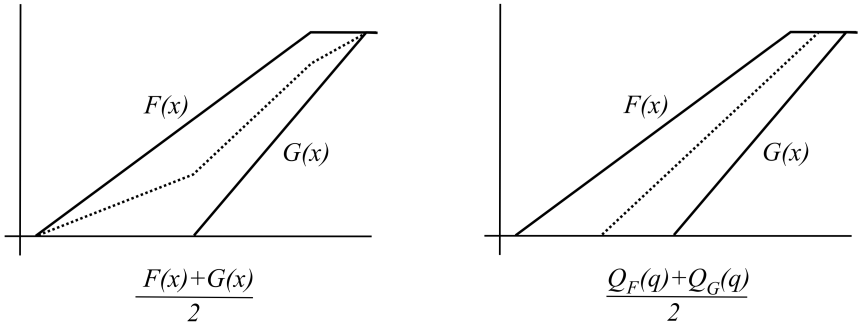

Figure 3. Left, the cumulative probability distribution obtained by computing the weighted mean of the cumulative probability distributions of F and G; an identical result is obtained by computing the weighted mean of the density functions of F and G. Right, the cumulative probability distribution obtained by computing the weighted mean of the quantile functions associated to the probability distributions of F and G

On the other hand, PFIS is any system that combines fuzzy set theory and probability theory to perform an inference. Based on these general considerations, many different approaches are related to ours. To be more precise in the comparison, we can classify PFIS as follows: approaches that consider fuzzy events (i.e., fuzzy sets with a certain probability assigned) in antecedents and consequents of rules [12, 19, 13]; approaches where a probability is assigned to each fuzzy rule [31, 2]; approaches where a probabilistic uncertainty is linked to the truth values (or shape) of the fuzzy sets in the consequents [30, 24]; and approaches where the antecedents of rules are fuzzy sets and the consequents are probability distributions of a (usually crisp) random variable [1, 27]. Our approach is framed in the later class, since the combination between fuzzy set theory and probability theory is given separately, namely: the antecedent is a set of fuzzy sets (no probability) and the consequent is a probability distribution of a random variable (no fuzziness). It is worth mentioning that [31, 2] can be also reinterpreted in such a way as well, since the consequent of fuzzy rules are crisp sets of classes $\mathscr{C}$ and the probability assigned to each rule can be identified directly with a discrete probability distribution in the respective class of $\mathscr{C}$. In any case, our approach differs from those approaches in two features. Firstly, the most obvious, in our approach we use the technique of F-transforms to construct the rules, whereas in [1, 27, 31, 2] use a fuzzy probabilistic version of Bayes theorem. The use of Bayes theorem in those approaches requires the discretization of the dependent variables, in [31, 2] the discretization is directly given by a discrete set of classes, whereas in [1, 27] the discretization is given by the choice of a fuzzy partition in the dependent variable. The use of F-transforms allows to define a probability function in the dependent variable for continuous cases without the choice of fuzzy partitions in the dependent variable.

The second difference is that the inference in [1, 27, 31, 2] is based on obtaining the mean of density probability functions or discrete probabilities, whereas our inference does the mean of quantile functions. Although quantile functions, cumulative probability distributions, density probability functions (resp. discrete probabilities) are equivalent in probability theory, the inference performed with them provides different results. In Figure 3 we show with a simple example that the probability distribution obtained by computing the mean of two density functions (equivalent to compute the mean of cumulative probability functions) is clearly different from the probability distribution obtained by computing the mean of the respective quantile functions. In other words, our proposed inference is different from the one proposed in [1, 27, 31, 2].

6. Conclusions and future works

In this paper, we have presented a method for constructing probabilistic fuzzy rules by means of direct quantile F-transforms. The obtained rules consider fuzzy sets in the antecedent and probability distributions in the consequent. As a result, we obtain a Probabilistic Fuzzy Inference System (PFIS) in which output is a probability distribution. The inference engine is based on the inverse quantile F-transforms and the procedure reminds the Takagi-Sugeno inference. We have also presented four significance measures that pretend to represent the information provided by each rule: two measures to determine whether the rule determines a dependence between variables or not; one to measure how the dispersion change according to one rule; and the last, to measure how much increase or decrease the

dependent variable according to one rule. Additionally, we have shown in a sequence of experiments, with synthetic and real data, that the proposed construction of PFIS captures the dependencies between different variables and that the significance measures report valuable information about that dependence.

There are different lines of future works. In the procedure presented in this paper we have obtained one rule for each class in the considered fuzzy partition. Therefore, on the one hand, it would be interesting to put effort in the construction of ad-hoc fuzzy partitions in order to avoid empty classes in the dataset (as done in [16]). On the other hand, it is interesting to develop a procedure to reduce the number of rules in the knowledge database and still, to be capable to retrieve right information. Another future lines will focus on the development of different defuzzification procedures and on the construction of fuzzy probabilistic rules where the kind of distribution is known (e.g., Normal, Exponential, Poisson, etc).

We also aim at the application of the proposed probabilistic-fuzzy inference system in practical problems. One interesting target is in extracting knowledge in the context of digital forensics, since the author participates in a European COST action related to such a topic (CA17124 - Digital forensics: evidence analysis via intelligent systems and practices).

Acknowledgment

Partially supported by the Spanish Ministry of Sciences project PGC2018-095869-B-I00, by the Junta de Andalucía project UMA2018-FEDERJA-001 (European Regional Development Funds) and by the European Cooperation in Science & Technology (COST) Action CA17124.

References

[1] R. J. Almeida, N. Baştürk, U. Kaymak, and J. M. C. Sousa. Analysis of probabilistic fuzzy systems' parameters in conditional density estimation. In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 2136–2143, 2016.

[2] M. Amiri, A. Ardeshir, and M. H. Fazel Zarandi. Fuzzy probabilistic expert system for occupational hazard assessment in construction. *Safety Science*, 93:16–28, 2017.

[3] P. Billingsley. *Probability and Measure*. Wisley-Interscience, 1995.

[4] F. Di Martino and S. Sessa. Detection of fuzzy association rules by fuzzy transforms. *Advances in Fuzzy Systems*, pages 1–12, 2012.

[5] S. Gudder. Fuzzy probability theory. *Demonstratio Mathematica*, 31(1):235–254, 1998.

[6] M. L. Guerra, L. Sorini, and L. Stefanini. Quantile and expectile smoothing based on l1-norm and l2-norm fuzzy transforms. *International Journal of Approximate Reasoning*, 107:17 – 43, 2019.

[7] M. L. Guerra, L. Sorini, and L. Stefanini. Bitcoin analysis and forecasting through fuzzy transform. *Axioms*, 9(4), 2020.

[8] P. Hájek. *Metamathematics of Fuzzy Logic*. Kluwer Academic Publishers, 1998.

[9] R. V. Hogg, J. W. McKean, and A. T. Craig. *Introduction to Mathematical Statistics*. Pearson, (8th Edition) 2019.

[10] G. Klir and B. Yuan. *Fuzzy sets and fuzzy logic - theory and applications*. Pearson, 1995.

[11] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, (4):83–91, 1933.

[12] C. Li, L. Ledo, M. Delgado, M. Cerrada, F. Pacheco, D. Cabrera, R.-V. Sánchez, and J. Valente de Oliveira. A bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification. *Knowledge-Based Systems*, 129:39–60, 2017.

[13] Z. Liu and H.-X. Li. A probabilistic fuzzy logic system for modeling and control. *IEEE Transactions on Fuzzy Systems*, 13(6):848–859, 2005.

[14] N. Madrid. An extension of f-transforms to more general data: potential applications. *Soft Computing*, 21(13):3551–3565, Jul 2017.

[15] N. Madrid. Toward the use of quantile fuzzy transforms for the construction of fuzzy association rules. In *29th IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2020.*, pages 1–8, 2020.

[16] N. Madrid and S. Díaz-Gómez. F-transforms for the definition of contextual fuzzy partitions. In L. T. Kóczy, J. Medina-Moreno, E. Ramírez-Poussa, and A. Šostak, editors, *Computational Intelligence and Mathematics for Tackling Complex Problems*, pages 167–173. Springer International Publishing, Cham, 2020.

[17] V. Novák, I. Perfilieva, and J. Močkoř. *Mathematical Principles of Fuzzy Logic*. Kluwer Academic Publishers, 1999.

[18] K. Nozaki, H. Ishibuchi, and H. Tanaka. A simple but powerful heuristic method for generating fuzzy rules from numerical data. *Fuzzy Sets and Systems*, 86(3):251 – 270, 1997.

[19] Y. PAN and B. YUAN. Bayesian inference of fuzzy probabilities. *International Journal of General Systems*, 26(1-2):73–90, 1997.

[20] I. Perfilieva. Fuzzy transforms: Theory and applications. *Fuzzy Sets and Systems*, 157(8):993 – 1023, 2006.

[21] I. Perfilieva, V. Novák, and A. Dvořák. Fuzzy transform in the analysis of data. *International Journal of Approximate Reasoning*, 48(1):36 – 46, 2008. Special Section: Perception Based Data Mining and Decision Support Systems.

[22] R. Simard and P. L'Ecuyer. Computing the two-sided kolmogorov–smirnov distribution. *Journal of Statistical Software*, 39(11):1–18, 2011.

[23] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics.*, 19(2):279–281, 1948.

[24] N. Sozhamadevi and S. Sathiyamoorthy. Modeling and control of an unstable system using probabilistic fuzzy inference system. *Archives of Control Sciences*, (No 3), 2015.

[25] M. Štěpnička, A. Dvořák, V. Pavliska, and L. Vavříčková. A linguistic approach to time series modeling with the help of f-transform. *Fuzzy Sets and Systems*, 180(1):164 – 184, 2011. Fuzzy Transform as a New Paradigm in Fuzzy Modeling.

[26] P. Türecki. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126 – 140, 2014.

[27] J. van den Berg, U. Kaymak, and R. J. Almeida. Conditional density estimation using probabilistic fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 21(5):869–882, 2013.

[28] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[29] Z. Xia. Fuzzy probability system: fuzzy-probability space (1). *Fuzzy Sets and Systems*, 120(3):469–486, 2001.

[30] G. Zhang, H.-X. Li, and M. Gan. Design a wind speed prediction model using probabilistic fuzzy system. *IEEE Transactions on Industrial Informatics*, 8(4):819–827, 2012.

[31] J. Zheng and Y. Tang. Fuzzy inference system with probability factor and its application in data mining. In Y. Zhang, K. Tanaka, J. X. Yu, S. Wang, and M. Li, editors, *Web Technologies Research and Development - APWeb 2005*, pages 944–949, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.