

Interpretability of Fuzzy Systems: Current Research Trends and Prospects

Jose M. Alonso, Ciro Castiello, and Corrado Mencar

Abstract Fuzzy systems are universally acknowledged as valuable tools to model complex phenomena while preserving a readable form of knowledge representation. The resort to natural language for expressing the terms involved in fuzzy rules, in fact, is a key-factor to conjugate mathematical formalism and logical inference with human-centered interpretability. That makes fuzzy systems specifically suitable in every real-world context where people are in charge of crucial decisions. That is because the self-explanatory nature of fuzzy rules profitably supports expert assessments. Additionally, as far as interpretability is investigated, it appears that: a) the simple adoption of fuzzy sets in modeling is not enough to ensure interpretability; b) fuzzy knowledge representation must confront the problem of preserving the overall system accuracy, thus yielding a trade-off which is frequently debated. Such issues have attracted a growing interest in the research community and became to assume a central role in the current literature panorama of Computational Intelligence. This chapter gives an overview of the topics related to fuzzy system interpretability, facing the ambitious goal of proposing some answers to a number of open challenging questions: What is interpretability? Why interpretability is worth considering? How to ensure interpretability, and how to assess (quantify) it? Finally, how to design interpretable fuzzy models?

Jose M. Alonso
European Centre for Soft Computing, e-mail: jose.alonso@softcomputing.es

Ciro Castiello
Department of Informatics, University of Bari, e-mail: ciro.castiello@uniba.it

Corrado Mencar
Department of Informatics, University of Bari, e-mail: corrado.mencar@uniba.it

1 Introduction

The key factor for the success of fuzzy logic stands in the ability of modeling and processing *perceptions* instead of measurements [79]. In most cases, such perceptions are expressed in natural language. Thus, fuzzy logic acts as a mathematical underpinning for modeling and processing perceptions described in natural language.

Historically, it has been acknowledged that fuzzy systems are endowed with the capability to conjugate a complex behavior and a simple description in terms of linguistic rules. In many cases, the compilation of fuzzy systems has been accomplished *manually*; with human knowledge purposely injected in fuzzy rules in order to model the desired behavior (the rules could be eventually tuned to improve the system accuracy). In addition, the great success of fuzzy logic led to the development of many algorithms aimed at acquiring knowledge from data (expressing it in terms of fuzzy rules). This made feasible the automatic design of fuzzy systems (through data-driven design techniques). Moreover, theoretical studies proved the universal approximation capabilities of such systems [75].

The adoption of data-driven design techniques is a common practice nowadays. Nevertheless, while fuzzy sets can be generally used to model perceptions, some of them do not lead to a straight interpretation in natural language. In consequence, the adoption of accuracy-driven algorithms for acquiring knowledge from data often results in unintelligible models. In those cases, the fundamental plus of fuzzy logic is lost and the derived models are comparable to other measurement-based models (like neural networks) in terms of knowledge interpretability.

In a nutshell, interpretability is not granted by the adoption of fuzzy logic which represents a necessary yet not a sufficient requirement for modeling and processing perceptions. However, interpretability is a quality that is not easy to define and quantify. Several open and challenging questions arise while considering interpretability in fuzzy modeling: *What* is interpretability? *Why* interpretability is worth considering? How to *ensure* interpretability? How to *assess* (quantify) interpretability? How to *design* interpretable fuzzy models? And so on.

The objective of this chapter is to provide some answers for the questions posed above. Section 2 deals with the challenging task of setting a proper definition of interpretability. Section 3 introduces the main constraints and criteria that can be adopted to ensure interpretability when designing interpretable fuzzy systems. Section 4 gives a brief overview of the soundest indexes for assessing interpretability. Section 5 presents the most popular approaches for designing fuzzy systems endowed with a good interpretability-accuracy trade-off. Section 6 enumerates some application fields where interpretability is a main concern. Section 7 sketches a number of challenging tasks which should be addressed in the near future. Finally, some conclusions are drawn in Section 8.

2 The quest for interpretability

Answering the question “*What is interpretability?*” is not straightforward. Defining interpretability is a challenging task since it deals with the analysis of the relation occurring between two heterogeneous entities: a model of the system to be designed (usually formalized through a mathematical definition) and a human user (meant not as a passive beneficiary of a system’s outcome, but as an active reader and interpreter of the model’s working engine). In this sense, interpretability is a quality which is inherent in the model and yet it refers to an act performed by the user who is willing to grasp and explain the meaning of the model.

To pave the way for the definition of such a relation, a common ground must be settled. This could be represented by a number of fundamental properties to be incorporated into a model, so that its formal description becomes compatible with the user’s knowledge representation. In this way, the human user may interface the mathematical model resting on concepts that appear to be suitable to deal with it. The quest for interpretability, therefore, calls for the identification of several features. Among them, resorting to an appropriate framework for knowledge representation is a crucial element and the adoption of a fuzzy inference engine based on fuzzy rules is straightforward to approach the linguistic-based formulation of concepts which is typical of the human abstract thought.

A distinguishing feature of a fuzzy rule-based model is the double level of knowledge representation. The lower level of representation is constituted by the formal definition of the fuzzy sets in terms of their membership functions, as well as the aggregation functions used for inference. This level of representation defines the *semantics* of a fuzzy rule-based model as it determines the behavior of the model, i.e. the input/output mapping for which it is responsible.

On the higher level of representation, knowledge is represented in form of rules. They define a formal structure where linguistic variables are involved and reciprocally connected by some formal operators, such as “AND”, “THEN”, and so on. Linguistic variables correspond to the inputs and outputs of the model. The (symbolic) values they assume are related to linguistic terms which, in turn, are mapped to the fuzzy sets defined in the lower level of representation. The formal operators are likewise mapped to the aggregation functions. This mapping provides the interpretative transition that is quite common in the mathematical context: a formal structure is assigned semantics by mapping symbols (linguistic terms and operators) to objects (fuzzy sets and aggregation functions).

In principle, the mapping of linguistic terms to fuzzy sets is arbitrary. It just suffices that identical linguistic terms are mapped to identical fuzzy sets. Of course, this is not completely true for formal operators (e.g., t-norms, implications, etc.). The corresponding aggregation functions should satisfy a number of constraints; however some flexibility is possible. Nevertheless, the mere use of symbols in the high level of knowledge representation implies the establishment of a number of semiotic relations that are fundamental for the quest of interpretability of a fuzzy model. In particular, linguistic terms — as usually picked from natural language — must be fully meaningful for the expected reader since they denote concepts, i.e.

mental representations that allow people to draw appropriate inferences about the entities they encounter.

Concepts and fuzzy sets, therefore, are both denoted by linguistic terms. Additionally, concepts and fuzzy sets play a similar role: the former (being part of the human knowledge) contribute to determine the behavior of a person; the latter (being the basic elements of a fuzzy rule base) contribute to determine the behavior of a system to be modeled. As a consequence, concepts and fuzzy sets are implicitly connected by means of the common linguistic terms they are related to, which refer to object classes in the real world. The key essence of interpretability is therefore the property of *cointension* [80] between fuzzy sets and concepts, consisting in the possibility of referring to similar classes of objects: such a possibility is assured by the use of common linguistic terms.

Semantic cointension is a key-issue when dealing with interpretability of fuzzy systems. It has been introduced and centered on the role of fuzzy sets, but it can be easily extended to refer to some more complex structures, such as fuzzy rules or the whole fuzzy models. In this regard, a crisp assertion about the importance of cointension pronounced at the level of the whole model is given by the Michalski's "Comprehensibility Postulate" [58]:

The results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single "chunks" of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion.

It should be observed that the above postulate has been formulated in the general area of Machine Learning. Nevertheless, the assertion made by Michalski has important consequences in the specific area of fuzzy modeling (FM) too. According to the Comprehensibility Postulate, results of computer induction should be described symbolically. Symbols are necessary to communicate information and knowledge, hence pure numerical methods, such as neural networks, are not suited for meeting interpretability unless an interpretability-oriented post-processing of the resulting knowledge is performed.

The key-point of the Michalski's postulate is the human centrality of the results of a computer induction process. The importance of the human component implicitly suggests a novel aspect to be taken into account in the quest for interpretability. Actually, the semantic cointension is related to one facet of the interpretability process, which can be referred to as *comprehensibility* of the content and behavior of a fuzzy model. In other words, cointension concerns the semantic interpretation performed by a user determined to comprehend such model. On the other hand, when we turn to consider the cognitive capabilities of human brains and their intrinsic limitations, then a different facet of the interpretability process can be defined in terms of *readability* of the bulk of information conveyed by a fuzzy model. In that case, simplicity is required to perform the interpretation process because of the limited ability to store information in the human brain's short term memory [59]. Therefore, structural measures concerning the complexity of a rule base affect the cognitive efforts of a user determined to read and interpret a fuzzy model.

Comprehensibility and readability represent two facets of a common issue and both of them are to be considered while assessing the interpretability process. In particular, this distinction should be acknowledged when criteria are specifically designed to provide a quantitative definition of interpretability.

2.1 Why is interpretability so important?

A great number of inductive modeling techniques are currently available to acquire knowledge from data. Many of these techniques provide predictive models that are very accurate and flexible enough to be applied in a wide range of applications. Nevertheless, the resulting models are usually considered as black-boxes, i.e. models whose behavior cannot be easily explained in terms of the model structure. On the other hand, the use of fuzzy rule-based models is a matter of design choice: whenever interpretability is a key factor, fuzzy rule-based models should be naturally preferred. It is worth noting that interpretability is a distinguishing feature of fuzzy rule-based models. Several reasons justify a choice inclined towards interpretability. They include but are not limited to:

Integration. In an interpretable fuzzy rule-based model the acquired knowledge can be easily verified and related to the domain knowledge of a human expert. In particular, it is easy to verify if the acquired knowledge expresses new and interesting relations about the data; also, the acquired knowledge can be refined and integrated with expert knowledge.

Interaction. The use of natural language as a mean for knowledge communication enables the possibility of interaction between the user and the model. Interactivity is meant to explore the acquired knowledge. In practice, it can be done at symbolical level (by adding new rules or modifying existing ones) and/or at numerical level (by modifying the fuzzy sets denoted by linguistic terms; or by adding new linguistic terms denoting new fuzzy sets).

Validation. The acquired knowledge can be easily validated against common-sense knowledge and domain-specific knowledge. This capability enables the detection of semantic inconsistencies that may have different causes (misleading data involved in the inductive process, local minimum where the inductive process may have been trapped, data overfitting, etc.). This kind of anomaly detection is important to drive the inductive process towards a qualitative improvement of the acquired knowledge.

Trust. The most important reason to adopt interpretable fuzzy models is their inherent ability to convince end-users about the reliability of a model (especially those users not concerned with knowledge acquisition techniques). An interpretable fuzzy rule-based model is endowed with the capability of explaining its inference process so that users may be confident on how it produces its outcomes. This is particularly important in such domains as medical diagnosis, where a human expert is the ultimate responsible for a decision.

2.2 A historical review

It has been long time since Zadeh's seminal work on fuzzy sets [76] and nowadays there are lots of fruitful research lines related to fuzzy logic [6]. Hence, we can state that fuzzy sets and systems have become the subjects of a mature research field counting several works both theoretical and applied in their scope. Fig. 1 shows the distribution of publications per year regarding interpretability issues. Three main phases can be identified taking into account the historical evolution of FM.

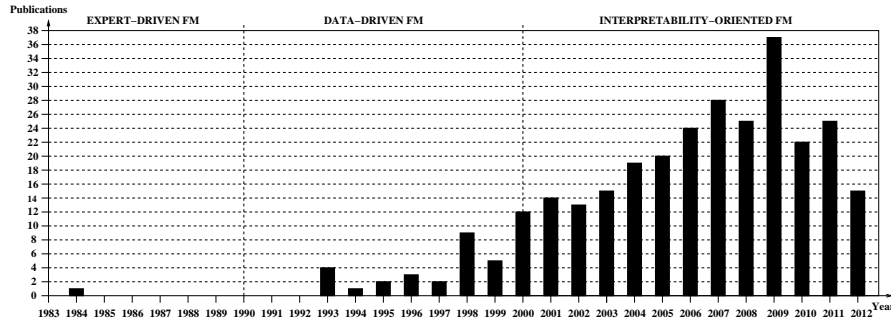


Fig. 1 Publications per year related to interpretability issues.

From 1965 to 1990. During this initial period interpretability emerged naturally as the main advantage of fuzzy systems. Researchers concentrated on building fuzzy models mainly working with expert knowledge and a few simple linguistic variables [78] and linguistic rules usually referred as Mamdani rules [52]. As a result, those designed fuzzy models were characterized by their high interpretability. Moreover, interpretability is assumed as an intrinsic property of fuzzy systems. Therefore, there are only a few publications regarding interpretability issues. Notice that, the first proposal of a Fuzzy Rule Based System (FRBS) was presented by Mamdani who was able to augment Zadeh's initial formulation allowing the application of fuzzy systems to a control problem. These kinds of fuzzy systems are also referred to as *fuzzy logic controllers*, as proposed by the author in his pioneering paper. In addition, Mamdani-type FRBSs became soon the main tool to develop linguistic models. Of course, many other rule formats were arising and gaining importance. In addition to Mamdani FRBSs, probably the most famous FRBSs are those proposed by Takagi and Sugeno [70], the popular TSK fuzzy systems, where the conclusion is a function of the input values. Due to their current popularity, in the following we will use the term "fuzzy system" to denote Mamdani-type FRBSs and their subsequent extensions.

From 1990 to 2000. In the second period the focus was set on accuracy. Researchers realized that expert knowledge was not enough to deal with complex systems. Thus, they explored the use of fuzzy machine learning techniques to automatically extract knowledge from data [44, 45]. Accordingly, those designed

fuzzy models became composed of extremely complicated fuzzy rules with high accuracy but at the cost of disregarding interpretability as a side effect. Obviously, automatically generated rules were rarely as readable as desired. Along this period some researchers started claiming that fuzzy models are not interpretable *per se*. Interpretability is a matter of careful design. Thus, interpretability issues must be deeply analyzed and seriously discussed. Although the amount of publications related to interpretability issues is still small in this period, please pay attention to the fact that publications begin to grow exponentially at the end of this second phase.

From 2000 to 2012. After the two previous periods, researchers realized that both expert-driven (from 1965 to 1990) and data-driven (from 1990 to 2000) design approaches have their own advantages and drawbacks, but they are somehow complementary. For instance, expert knowledge is general and easy to interpret but hard to formalize. On the contrary, knowledge derived from data can be extracted automatically but it becomes quite specific and its interpretation is usually hard [39]. Moreover, researchers were aware of the need of taking into account simultaneously interpretability and accuracy during the design of fuzzy models. As a result, during this third phase the main challenge was how to combine expert knowledge and knowledge extracted from data, with the aim of designing compact and robust systems with a good interpretability-accuracy trade-off. When considering both interpretability and accuracy in FM, two main strategies turn up naturally [1]: *Linguistic Fuzzy Modeling* (LFM) and *Precise Fuzzy Modeling* (PFM). On the one hand, in LFM system designers first focus on the interpretability of the model, and then they try to improve its accuracy [22]. On the other hand, in PFM designers first build a fuzzy model maximizing its accuracy, and then they try to improve its interpretability [23]. As an alternative, since accuracy and interpretability represent conflicting goals by nature, multi-objective fuzzy modeling strategies (considering accuracy and interpretability as objectives) have become very popular [26, 42].

At the same time, there has been a great effort for formalizing interpretability issues. As a result, the number of publications has grown a lot. Researchers have actively looked for the right definition of interpretability. In addition, several interpretability constraints have been identified. Moreover, interpretability assessment has become a hot research topic. In fact, several interpretability indexes (able to guide the FM design process) have been defined. Nevertheless, a universal index widely admitted is still missing. Hence, further research on interpretability issues is demanded.

Unfortunately, although the number of publications was growing exponentially until 2009, later it started decreasing. In 2012 the number of publications dropped down dramatically reaching the same levels of 2003. We would like to emphasize the impact of the two pioneer books [22, 23] edited in 2003. They contributed to make the fuzzy community aware of the need to take into account again interpretability as a main research concern. It is worth noting that the first formal definition of interpretability (in the fuzzy literature) was included in [23]. It was

given by Bodenhofer and Bauer [17] who established an axiomatic treatment of interpretability at the level of linguistic variables.

We encourage the fuzzy community to keep paying attention to interpretability issues because there is still a lot of research to be done. Interpretability must be the central point on system modeling. In fact, some of the hottest and most recent research topics like Precisiated Natural Language, Computing With Words, and Human Centric Computing strongly rely on the interpretability of the designed models. The challenge is to better exploit fuzzy logic techniques for improving the human-centric character of many intelligent systems. Therefore, interpretability deserves consideration as a main research concern and the number of publications should grow again in the next years.

3 Interpretability constraints and criteria

Interpretability is a quality of fuzzy systems that is not immediate to quantify. Nevertheless, a quantitative definition is required both for assessing the interpretability of a fuzzy system and for designing new fuzzy systems. This requirement is especially stringent when fuzzy systems are automatically designed from data, through some knowledge extraction procedure.

A common approach for defining interpretability is based on the adoption of a number of constraints and criteria that, taken as a whole, provide for a definition of interpretability. This approach is inherent to the subjective nature of interpretability, because the validity of some conditions/criteria is not universally acknowledged and may depend on the application context.

In literature, a large number of interpretability constraints and criteria can be found. Some of them are widely accepted, while others are controversial. The nature of these constraints and criteria is also diverse. Some are neatly defined as a mathematical condition, others have a fuzzy character and their satisfaction is a matter of degree. This Section is addressed to give a brief yet homogeneous outline of the best known interpretability constraints and criteria. The reader is referred to the specialized literature for deeper insights on this topic [57, 73].

Several ways are available to categorize interpretability constraints and criteria. It could be possible to refer to their specific nature (e.g., crisp vs. fuzzy), to the components of the fuzzy system where they are applied, or to the description level of the fuzzy system itself. Here, as depicted in Fig. 2, we choose a hierarchical organization that starts from the most basic components of a fuzzy system, namely the involved fuzzy sets, and goes on toward more complex levels, such as fuzzy partitions, fuzzy rules, up to considering the model as a whole.



High-Level  Abstraction Levels  Low-Level	Fuzzy Rule Bases	-> Compactness -> Average firing rules -> Logical view -> Completeness -> Locality
	Fuzzy Rules	-> Description length -> Granular output
	Fuzzy Partitions	-> Justifiable number of elements -> Distinguishability -> Coverage -> Relation preservation -> Prototypes on special elements
	Fuzzy Sets	-> Normality -> Continuity -> Convexity

Fig. 2 Interpretability constraints and criteria in different abstraction levels.

3.1 Constraints and criteria for fuzzy sets

Fuzzy sets are the basic elements of fuzzy systems and their role is to express elementary yet imprecise concepts that can be denoted by linguistic labels. Here we assume that fuzzy sets are defined on a universe of discourse represented by a closed interval of the real line (this is the case of most fuzzy systems, especially those acquired from data). Thus, fuzzy sets are the building blocks to translate a numerical domain in a linguistically quantified domain that can be used to communicate knowledge.

Generally speaking, single fuzzy sets are employed to express elementary concepts and, through the use of connectives, are combined to represent more complex concepts. However, not all fuzzy sets can be related to elementary concepts, since the membership function of a fuzzy set may be very awkward but still legitimate from a mathematical viewpoint. Actually, a sub-class of fuzzy sets should be considered, so that its members can be easily associated to elementary concepts and tagged by the corresponding linguistic labels. Fuzzy sets of this sub-class must verify a number of basic interpretability constraints, including:

Normality. At least one element of the universe of discourse is a prototype for the fuzzy set, i.e. it is characterized by a full membership degree. A normal fuzzy set represents a concept that fully qualifies at least one element of the universe of discourse, i.e. the concept has at least one example that fulfills it. On the other

hand, a sub-normal fuzzy set is usually a consequence of a partial contradiction (it is easy to show that the degree of inclusion of a sub-normal fuzzy set in the empty set is non-zero).

Continuity. The membership function is continuous on the universe of discourse. As a matter of fact, most concepts that can be naturally represented through fuzzy sets derive from a perceptual act, which comes from external stimuli that usually vary in continuity. Therefore, continuous fuzzy sets are better in accordance with the perceptive nature of the represented concepts.

Convexity. In a convex fuzzy set, given three elements linearly placed on the axis related to the universe of discourse, the degree of membership of the middle element is always greater than or equal to the minimum membership degree of the side elements [63]. This constraint encodes the rule that if a property is satisfied by two elements, then it is also satisfied by an element settled between them.

3.2 Constraints and criteria for fuzzy partitions

The key success factor of fuzzy logic in modeling is the ability of expressing knowledge *linguistically*. Technically this is realized by linguistic variables, i.e. variables that assume symbolic values called linguistic terms. The peculiarity of linguistic variables with respect to classical symbolic approaches is the interpretation of linguistic terms as fuzzy sets. The collection of fuzzy sets used as interpretation of the linguistic terms of a linguistic variable forms a fuzzy partition of the universe of discourse.

To understand the role of a fuzzy partition, we should consider that it is meant to define a relation among fuzzy sets. Such a relation must be co-intensive with the one connecting the elementary concepts represented by the fuzzy sets involved in the fuzzy partition. That is the reason why the design of fuzzy partitions is so crucial for the overall interpretability of a fuzzy system. The most critical interpretability constraints for fuzzy partitions are:

Justifiable number of elements. The number of fuzzy sets included in a linguistic variable must be small enough so that they can be easily remembered and recalled by users. Psychological studies suggest at most nine fuzzy sets or even less [59, 68]. Usually, three to five fuzzy sets are convenient choices to set the partition cardinality.

Distinguishability. Since fuzzy sets are denoted by distinct linguistic terms, they should refer to well distinguished concepts. Therefore, fuzzy sets in a partition should be well separated, although some overlapping is admissible because usually perception-based concepts are not completely disjoint. Several alternatives are available to quantify distinguishability, including similarity and possibility [54].

Coverage. Distinguishable fuzzy sets are necessary, but if they are too much separated they risk to under-represent some subset of the universe of discourse. The

coverage constraint requires that each element of the universe of discourse must belong to at least one fuzzy set of the partition with a membership degree not less than a threshold [57]. This requirement involves that each element of the universe of discourse has some quality that is well represented in the fuzzy partition. On the other hand, the lack of coverage is a signal of incompleteness of the fuzzy partition that may hamper the overall comprehensibility of the system's knowledge. Coverage and distinguishability are somewhat conflicting requirements that are usually balanced by fuzzy partitions that enforce the intersection of adjacent fuzzy sets to elements whose maximum membership degree is equal to a threshold (usually the value of this threshold is set to 0.5).

Relation preservation. The concepts that are represented by the fuzzy sets in a fuzzy partition are usually cross-related. The most immediate relation which can be conceived among concepts is related to the order (e.g., LOW preceding MEDIUM, preceding HIGH, and so on). Relations of this type must be preserved by the corresponding fuzzy sets in the fuzzy partition [18].

Prototypes on special elements. In many problems some elements of the universe of discourse have some special meaning. A common case is the meaning of the bounds of the universe of discourse, which usually represent some extreme qualities (e.g., VERY LARGE or VERY SMALL). Other examples are possible, which could be aside from the bounds of the universe of discourse being, instead, more problem-specific (e.g., prototypes could be conceived for the icing point of water, the typical human body temperature, etc.). In all these cases, the prototypes of some fuzzy sets of the partition must coincide with such special elements.

3.3 Constraints and criteria for fuzzy rules

In most cases a fuzzy system is defined over a multi-dimensional universe of discourse that can be split into many one-dimensional universes of discourse, each of them associated to a linguistic variable. A subset of these linguistic variables is used to represent the input of a system, while the remaining variables (usually only one variable) are used to represent the output. The input/output behavior is expressed in terms of rules. Each rule prescribes a linguistic output value when the input matches the rule condition (also called rule premise), usually expressed as a logical combination of soft constraints. A soft constraint is a linguistic proposition (specification) that ties a linguistic variable to a linguistic term (e.g., TEMPERATURE IS HIGH). Furthermore the soft constraints combined in a rule condition may involve different linguistic variables (e.g., TEMPERATURE IS HIGH AND PRESSURE IS LOW).

A fuzzy rule is a unit of knowledge that has the twofold role of determining the system behavior and communicating this behavior in a linguistic form. The latter feature urges to adopt a number of interpretability constraints which are to be added up to the constraints required for fuzzy sets and fuzzy partitions. Some of the most general interpretability constraints and criteria for fuzzy rules are the following:

Description length. The description length of a fuzzy rule is the sum of the number of soft constraints occurring in the condition and in the consequent of the rule (it is usually known as *total rule length*). In most cases, only one linguistic variable is represented in a rule consequent, therefore the description length of a fuzzy rule is directly related to the complexity of the condition. A small number of soft constraints in a rule implies both high readability and semantic generality, hence short rules should be preferred in fuzzy systems.

Granular outputs. The main strength of fuzzy systems is their ability to represent and process imprecision in both data and knowledge. Imprecision is part of fuzzy inference, therefore the inferred output of a fuzzy system should carry information about the imprecision of its knowledge. This can be accomplished by using fuzzy sets as outputs. Defuzzification collapses fuzzy sets into single scalars; it should be therefore used only when strictly necessary and in those situations where outputs are not the object of user interpretation.

3.4 Constraints and criteria for fuzzy rule bases

As previously stated, the interpretability of a rule base taken as a whole has two facets: (1) a structural facet (*readability*), which is mainly related to the easiness of reading the rules; and (2) a semantic facet (*comprehensibility*), which is related to the information conveyed to the users who are willing to understand the system behavior. The following interpretability constraints and criteria are commonly defined to ensure the structural and semantic interpretability of fuzzy rule bases:

Compactness. A compact rule base is defined by a small number of rules. This is a typical structural constraint that advocates for simple representation of knowledge in order to allow easy reading and understanding. Nevertheless, a small number of rules usually involves low accuracy; it is therefore very common to balance compactness and accuracy in a trade-off that mainly depends on user needs.

Average firing rules. When an input is applied to a fuzzy system, the rules whose conditions are verified to a degree greater than zero are “firing”, i.e. they contribute to the inference of the output. On the average, the number of firing rules should be as small as possible, so that users are able to understand the contributions of the rules in determining the output.

Logical view. Fuzzy rules resemble logical propositions when their linguistic description is considered. Since linguistic description is the main mean for communicating knowledge, it is necessary that logical laws are applicable to fuzzy rules; otherwise, the system behavior may result counter-intuitive. Therefore the validity of some basic laws of the propositional logic (like *Modus Ponens*) and the truth-preserving operations (e.g., application of distributivity, De Morgan laws, etc.) should be verified also for fuzzy rules.

Completeness. The behavior of a fuzzy system is well defined for all inputs in the universe of discourse; however when the maximum firing strength determined by

an input is too small, it is not easy to justify the behavior of the system in terms of the activated rules. It is therefore required that for each possible input at least one rule is activated with a firing strength greater than a threshold value (usually set to 0.5) [57].

Locality. Each rule should define a local model, i.e. a fuzzy region in the universe of discourse where the behavior of the system is mainly due to the rule and only marginally by other rules that are simultaneously activated [65]. This requirement is necessary to avoid that the final output of the system is a consequence of an interpolative behavior of different rules that are simultaneously activated with high firing strengths. On the other hand, a moderate overlapping of local models is admissible in order to enable a smooth transition from a local model to another when the input values gradually shift from one fuzzy region to another.

On summary, a number of interpretable constraints and criteria apply to all levels of a fuzzy system. This Section highlighted only the constraints that are general enough to be applied independently on the modeling problem; however, several problem-specific constraints are also reported in literature (e.g., attribute correlation). Sometimes interpretability constraints are conflicting (as exemplified by the dichotomy distinguishability vs. coverage) and, in many cases, they conflict with the overall accuracy of the system. A balance is therefore required, asking in its turn for a way to assess interpretability in a qualitative but also quantitative way. This is the main subject of the next Section.

4 Interpretability assessment

The interpretability constraints and criteria presented in previous section belong to two main classes: (1) structural constraints and criteria referring to the static description of a fuzzy model in terms of the elements that compose it; and (2) semantic constraints and criteria quantifying interpretability by looking at the behavior of the fuzzy system. Whilst structural constraints address the *readability* of a fuzzy model, semantic constraints focus on its *comprehensibility*.

Of course, interpretability assessment must regard both global (description readability) and local (inference comprehensibility) points of view. It must also take into account both structural and semantic issues when considering all components (fuzzy sets, fuzzy partitions, linguistic partitions, linguistic propositions, fuzzy rules, fuzzy operators, etc.) of the fuzzy system under study.

Thus, assessing interpretability represents a challenging task mainly because the analysis of interpretability is extremely subjective. In fact, it clearly depends on the feeling and background (knowledge, experience, etc.) of the person who is in charge of making the evaluation. Even though having subjective indexes would be really appreciated for personalization purposes, looking for a universal metric widely admitted makes mandatory also the definition of objective indexes. Hence, it is necessary to consider both objective and subjective indexes. On the one hand, objective indexes are aimed at making feasible fair comparisons among different fuzzy models

designed for solving the same problem. On the other hand, subjective indexes are thought for guiding the design of customized fuzzy models, thus making easier to take into account users' preferences and expectations during the design process.

The rest of this section gives an overview on the most popular interpretability indexes which turn out from the specialized literature. Firstly, Zhou and Gan [81] established a two-level taxonomy regarding interpretability issues. They distinguished between low-level (also called fuzzy set level) and high-level (or fuzzy rule level). This taxonomy was extended by Alonso et al. [7] who introduced a conceptual framework for characterizing interpretability. They considered both fuzzy partitions and fuzzy rules at several abstraction levels. Moreover, in [55] the authors remarked the need to distinguish between readability (related to structural issues) and comprehensibility (related to semantic issues). Later, Gacto et al. [36] proposed a double axis taxonomy regarding semantic and structural properties of fuzzy systems, at both partition and rule base levels. Accordingly, they pointed out four groups of indexes. Below, we briefly introduce the two most sounded indexes inside each group (they are summarized in Fig. 3).

	Fuzzy Partition Level	Fuzzy Rule Base Level
Structural-based Interpretability	G1 Number of features Number of membership functions	G2 Number of rules Number of conditions
	G3 Context-adaptation based index GM3M index	G4 Semantic-cointension based index Co-firing based comprehensibility index

Fig. 3 Interpretability indexes considered in this work.

G1. Structural-based interpretability at fuzzy partition level:

- *Number of features.*
- *Number of membership functions.*

G2. Structural-based interpretability at fuzzy rule base level:

- *Number of rules.* This index is the most widely used [7].
- *Number of conditions.* This index corresponds to the previously mentioned *total rule length* which was coined by Ishibuchi et al. [47].

G3. Semantic-based interpretability at fuzzy partition level:

- *Context-adaptation based index* [19]. This index was introduced by Botta et al. with the aim of guiding the so-called context adaptation approach for multi-objective evolutionary design of fuzzy rule-based systems. It is actually an interpretability index based on fuzzy ordering relations.
- *GM3M index* [35]. Gacto et al. proposed an index defined as the geometric mean of three single metrics. The first metric computes the displacement of

the tuned membership functions with respect to the initial ones. The second metric evaluates the changes in the shapes of membership functions in terms of lateral amplitude rate. The third metric measures the area similarity. This index was used to preserve the semantic interpretability of fuzzy partitions along multi-objective evolutionary rule selection and tuning processes aimed at designing fuzzy models with a good interpretability-accuracy trade-off.

G4. Semantic-based interpretability at fuzzy rule base level:

- *Semantic-cointension based index* [56]. This index exploits the cointension concept coined by Zadeh [80]. In short, two different concepts referring almost to the same entities are taken as cointensive. Thus, a fuzzy system is deemed as comprehensible only when the explicit semantics (defined by fuzzy sets attached to linguistic terms as well as fuzzy operators) embedded in the fuzzy model is cointensive with the implicit semantics inferred by the user while reading the linguistic representation of the rules. In the case of classification problems, semantic cointension can be evaluated through a logical view approach, which evaluates the degree of fulfillment of a number of logical laws exhibited by a given fuzzy rule base [55]. The idea mainly relies on the assumption that linguistic propositions resemble logical propositions, for which a number of basic logical laws are expected to hold.
- *Co-firing based comprehensibility index* [10]. It measures the complexity of understanding the fuzzy inference process in terms of information related to co-firing rules, i.e. rules firing simultaneously with a given input vector. This index emerges in relation with a novel approach for fuzzy system comprehensibility analysis, based on visual representations of the fuzzy rule-based inference process. Such representations are called fuzzy inference-grams (fin-grams) [61, 62]. Given a fuzzy rule base, a fingram plots it graphically as a social network made of nodes representing fuzzy rules and edges connecting nodes in terms of rule interaction at inference level. Edge weights are computed by paying attention to the number of co-firing rules. Thus, looking carefully at all the information provided by a fingram it becomes easy and intuitive understanding the structure and behavior of the fuzzy rule base it represents.

Notice that, most published interpretability indexes only deal with structural issues, so they correspond to groups G1 and G2. Indexes belonging to these groups are mainly quantitative. They essentially analyze the structural complexity of a fuzzy model by counting the number of elements (membership functions, rules, etc.) it contains. As a result, these indexes can be deemed as objective ones. Although these indexes are usually quite simple (that is the reason why we have just listed them above), they are by far the most popular ones. On the contrary, only a few interpretability indexes are able to assess the comprehensibility of a fuzzy model dealing with semantic issues (they belong to groups G3 and G4). This is mainly due to the fact that these indexes must take into account not only quantitative but also qualitative aspects of the modeled fuzzy system. They are inherently subjective and therefore not easy to formalize (that is the reason why we have provided more de-

tails above). Anyway, the interested reader is referred to the cited papers for further information. Moreover, a much more exhaustive list of indexes can be found in [36].

Even though there has been a great effort in the last years to propose new interpretability indexes, a universal index is still missing. Hence, defining such an index remains a challenging task. Anyway, we would like to highlight the need to address another encouraging challenge that is the careful design of interpretable fuzzy systems guided by one or more of the already existing interpretability indexes.

5 Designing interpretable fuzzy systems

Linguistic (Mamdani-type) fuzzy systems are widely known as a powerful tool to develop linguistic models [52]. They are made up of two main components:

- the *inference engine*, that is the component of the fuzzy system in charge of the fuzzy processing tasks;
- the *knowledge base* (KB), that is the component of the fuzzy system that stores the knowledge about the problem being solved. It is composed of:
 - the *fuzzy partitions*, describing the linguistic terms along with the corresponding membership functions defining their semantics, and
 - the *fuzzy rule base*, constituted by a collection of linguistic rules with the following structure:

IF X_1 is A_1 and ... and X_n is A_n **THEN** Y_1 is B_1 and ... and Y_m is B_m

with X_i and Y_j being input and output linguistic variables respectively, and A_i and B_j being linguistic terms defined by the corresponding fuzzy partitions. This structure provides a natural framework to include expert knowledge in the form of linguistic fuzzy rules. In addition to expert knowledge, induced knowledge automatically extracted from experimental data (describing the relation between system input and output) can also be easily formalized in the same rule base. Expert and induced knowledge are complementary. Furthermore, they are represented in a highly interpretable structure. The fuzzy rules are composed of input and output linguistic variables which take values from their term sets having a meaning associated to each linguistic label. As a result, each rule is a description of a condition-action statement that offers a clear interpretation to a human.

The accuracy of a fuzzy system directly depends on two aspects, the composition of the KB (fuzzy partitions and fuzzy rules) and the way in which it implements the fuzzy inference process. Therefore, the design process of a fuzzy system includes two main tasks which are going to be further explained in the following subsections, regarding both interpretability and accuracy:

- *Generation of the KB* in order to formulate and describe the knowledge that is specific to the problem domain.
- *Conception of the inference engine*, that is the choice of the different fuzzy operators that are employed by the inference process.

Mamdani-type fuzzy systems favor interpretability. Therefore they are usually considered when looking for interpretable fuzzy systems. However, it is important to remark that they are not interpretable *per se*. Notice that designing interpretable fuzzy systems is a matter of careful design.

5.1 Design strategies for the generation of a KB regarding the interpretability-accuracy trade-off

The two main objectives to be addressed in the FM field are *interpretability* and *accuracy*. Of course, the ideal aim would be to satisfy both objectives to a high degree but, since they represent conflicting goals, it is generally not possible. Regardless of the approach, a common scheme is found in the existing literature:

- Firstly, the main objective (interpretability or accuracy) is tackled defining a specific model structure to be used, thus setting the FM approach.
- Then, the modeling components (model structure and/or modeling process) are improved by means of different mechanisms to achieve the desired ratio between interpretability and accuracy.

This procedure resulted in four different possibilities: (1) LFM with improved interpretability, (2) LFM with improved accuracy, (3) PFM with improved interpretability, and (4) PFM with improved accuracy.

Option (1) gives priority to interpretability. Although a fuzzy system designed by LFM uses a model structure with high descriptive power, it has some problems (curse of dimensionality, excessive number of input variables or fuzzy rules, garbled fuzzy sets, etc.) that make it not as interpretable as desired. In consequence, there is a need of interpretability improvements to restore the pursued balance.

On the contrary, option (4) considers accuracy as the main concern. However, obtaining more accuracy in PFM does not pay attention to the interpretability of the model. Thus, this approach goes away from the aim of this book chapter. It acts close to black box techniques. So it does not follow the original objective of FM (not taking profit from the advantages that distinguish it from other modeling techniques).

Finally, the two remaining options, (2) and (3), propose improvement mechanisms to compensate for the initial imbalance in the quest for the best trade-off between interpretability and accuracy. In summary, three main approaches exist depending on how the two objectives are optimized (sequentially or at once):

- First Interpretability Then Accuracy (*LFM with improved accuracy*).
- First Accuracy Then Interpretability (*PFM with improved interpretability*).

- **Multi-Objective Design.** Both objectives are optimized at the same time.

The rest of this section provides additional details related to each of these approaches.

First Interpretability Then Accuracy. LFM has some inflexibility due to the use of linguistic variables with global semantics that establishes a general meaning of the used fuzzy sets [16]:

1. There is a lack of flexibility in the fuzzy system because of the rigid partitioning of the input and output spaces.
2. When the system input variables are dependent, it is very hard to find out right fuzzy partitions of the input spaces.
3. The usual homogeneous partitioning of the input and output spaces does not scale to high-dimensional spaces. It yields to the well-known curse of dimensionality problem that is characteristic of fuzzy systems.
4. The size of the KB directly depends on the number of variables and linguistic terms in the model. The derivation of an accurate linguistic fuzzy system usually requires a big number of linguistic terms. Unfortunately, this fact causes the number of rules to rise significantly, which may cause the system to lose the capability of being readable by human beings. Of course, in most cases it would be possible to obtain an equivalent fuzzy system with a much smaller number of rules by renouncing to that kind of rigidly partitioned input space.

However, it is possible to make some considerations to face the disadvantages enumerated above. Basically, two ways of improving the accuracy in LFM can be considered by performing the improvement in:

- the *model structure*, slightly changing the rule structure to make it more flexible, or in
- the *modeling process*, extending the model design to other components beyond the rule base, such as the fuzzy partitions, or even considering more sophisticated derivations of it.

Notice that, the so-called strong fuzzy partitions are widely used because they satisfy most of the interpretability constraints introduced in Section 3.2. The design of fuzzy partitions may be integrated within the whole derivation process of a fuzzy system with different schemata:

- *Preliminary design.* It involves extracting fuzzy partitions automatically by induction (usually performed by non-supervised clustering techniques) from the available dataset.
- *Embedded design.* Following a meta-learning process, this approach first derives different fuzzy partitions and then samples its efficacy running an embedded basic learning method of the entire KB [28].
- *Simultaneous design.* The process of designing fuzzy partitions is developed together with the derivation of other components such as the fuzzy rule base [43].

- *A posteriori design*. This approach involves tuning of the previously defined fuzzy partitions once the remaining components have been obtained. Usually, the tuning process changes the membership function shapes with the aim of improving the accuracy of the linguistic model [51]. Nevertheless, sometimes it also takes care of getting better interpretability (e.g., merging membership functions [31]).

It is also possible to opt for using more sophisticated rule base learning methods while the fuzzy partitions and the model structure are kept unaltered. Usually, all these improvements have the final goal of enhancing the *interpolative reasoning* the fuzzy system develops. For instance, the COR (cooperative rules) method follows the primary objective of inducing a better cooperation among linguistic rules [21].

As an alternative, other authors advocate the extension of the usual linguistic model structure to make it more flexible. As Zadeh highlighted in [77], a way to do so without losing the description ability to a high degree is to use linguistic hedges (also called *linguistic modifiers* in a wider sense). In addition, the rule structure can be extended through the definition of double-consequent rules, weighted rules, rules with exceptions, hierarchical rule bases, etc.

First Accuracy Then Interpretability. The birth of more flexible fuzzy systems such as TSK or approximate ones (allowing the FM to achieve higher accuracy) entailed the eruption of PFM. Nevertheless, the modeling tasks with these kinds of fuzzy systems increasingly resembled black box processes. Consequently, nowadays several researchers share the idea of rescuing the seminal intent of FM, i.e. to preserve the good interpretability advantages offered by fuzzy systems. This fact is usually attained by reducing the complexity of the model [67]. Furthermore, there are approaches aimed at improving the local description of TSK-type fuzzy rules:

1. *Merging/removing fuzzy sets in precise fuzzy systems*. The interpretability of TSK-type fuzzy systems may be improved by removing those fuzzy sets that, after an automatic adaptation and/or acquisition, do not contribute significantly to the model behavior. Two aspects must be considered:
 - *Redundancy*. It refers to the coexistence of similar fuzzy sets representing compatible concepts. In consequence, models become more complex and difficult to understand (the distinguishability constraint is not satisfied).
 - *Irrelevancy*. It arises when fuzzy sets with a constant membership degree equal to one, or close to it, are used. These kinds of fuzzy sets do not furnish relevant information.

The use of similarity measures between fuzzy sets has been proposed to automatically detect these undesired fuzzy sets [69]. Through first merging/removing fuzzy sets and then merging fuzzy rules, the precise fuzzy model goes through an interpretability improvement process that makes it less complex (more compact) and more easily interpretable (more transparent).

2. *Ordering/selecting TSK-type fuzzy rules*. An efficient way to improve the interpretability in FM is to select a subset of significant fuzzy rules that represent

in a more compact way the system to be modeled. Moreover, as a side effect this selection of important rules reduces the possible redundancy existing in the fuzzy rule base, thus improving the generalization capability of the system, i.e., its accuracy. For instance, resorting to orthogonal transformations [53] is one of the most successful approaches in this sense.

3. *Exploiting the local description of TSK-type fuzzy rules.* TSK-type fuzzy systems are usually considered as the combination of simple models (the rules) that describe local behaviors of the system to be modeled. Hence, insofar as each fuzzy rule is either forced to have a smoother consequent polynomial function or to develop an isolated action, the interpretability will be improved:
 - *Smoothing the consequent polynomial function* [34]. Through imposing several constraints to the weights involved in the polynomial function of each rule consequent then a convex combination of the input variables is performed. This contributes to a better understanding of the model.
 - *Isolating the fuzzy rule actions* [67]. The description of each fuzzy rule is improved when the overlapping between adjacent input fuzzy sets is reduced. Notice that the performance region of a rule is more clearly defined by avoiding that other rules have high firing degree in the same area.

Multi-objective Design. Since interpretability and accuracy are widely recognized as conflicting goals, the use of multi-objective evolutionary (MOE) strategies is becoming more and more popular in the quest for the best interpretability-accuracy trade-off [26, 33]. Ducange and Marcelloni [29] proposed the following taxonomy of multi-objective evolutionary fuzzy systems:

1. *MOE Tuning.* Given an already defined fuzzy system, its main parameters (typically membership function parameters but also fuzzy inference parameters) are refined through MOE strategies [4, 32].
2. *MOE Learning.* The components of a fuzzy system KB, both fuzzy partitions forming the data-base (DB) and fuzzy rules forming the rule-base (RB), are automatically generated from experimental data.
 - *MOE DB Learning.* The most relevant variables are identified and the optimum membership function parameters are defined from scratch. It usually wraps an RB heuristic-based learning process [2].
 - *MOE RB Selection.* Starting from an initial RB, a set of non-dominated RBs is generated by selecting subsets of rules exhibiting different trade-offs between interpretability and accuracy [46]. In some works [3, 35], MOE RB selection and MOE tuning are carried out together.
 - *MOE RB Learning.* The entire set of fuzzy rules is fully defined from scratch. In this approach uniformly distributed fuzzy partitions are usually considered [24].
 - *MOE KB Learning.* Simultaneous evolutionary learning of all KB components (DB and RB). Concurrent learning of fuzzy partitions and fuzzy rules proved to be a powerful tool in the quest for a good balance between interpretability and accuracy [12].

It is worthy to note that for the sake of clarity we have only cited some of the most relevant papers in the field of MOE fuzzy systems. For further details, the interested reader is referred to [29, 33] where a much more exhaustive review of related works is carried out.

5.2 Design decisions at fuzzy processing level

Although there are studies analyzing the behavior of the existing fuzzy operators for different purposes, unfortunately this question has not been considered yet as a whole from the interpretability point of view. Keeping in mind the interpretability requirement, the implementation of the inference engine must address the following careful design choices:

Select the right conjunctive operator T to be used in the antecedent of the rule.

Different operators (belonging to the t-norm family) are available to make this choice [41].

Select the operator I to be used in the fuzzy implication of “IF-THEN” rules.

Mamdani proposed to use the minimum operator as the t-norm for implication. Since then, various other t-norms have been suggested as implication operator [41], for instance the algebraic product. Other important family of implication operators are the fuzzy implication functions [71], one of the most usual being the Lukasiewicz’s one. Less common implication operators such as force-implications [30], t-conorms and operators not belonging to any of the most known implication operator families [49] have been considered too.

Choose the right inference mechanism. Two main strategies are available:

- *FATI (First Aggregation Then Inference)*. All antecedents of the rules are aggregated to form a multidimensional fuzzy relation. Via the composition principle the output fuzzy set is derived. This strategy is preferred when dealing with implicative rules [48].
- *FITA (First Inference Then Aggregation)*. The output of each rule is first inferred, and then all individual fuzzy outputs are aggregated. This is the common approach when working with the usual conjunctive rules. This strategy has become by far the most popular, especially in case of real-time applications. The choice for an output aggregation method (in some cases this is called the *also* operator) is closely related to the considered implication operator since it has to be related to the interpretation of the rules (which is connected to the kind of implication).

Choose the most suitable defuzzification interface operation mode. There are different options being the most widely used the Center of Area (COA), also called Center of Gravity (COG), and the Mean of Maxima (MOM). Even though most methods are based on geometrical or statistical interpretations, there are also parametric methods, adaptive methods including human knowledge, and even evolutionary adaptive methods [27].

6 Interpretable fuzzy systems in the real world

Interpretable fuzzy systems have an immediate impact on real-world applications. In particular, their usefulness is appreciable in all application areas that put humans at the center of computing. Interpretable fuzzy systems, in fact, conjugate knowledge acquisition capabilities with the ability of communicating knowledge in a human-understandable way.

Several application areas can take advantage from the use of interpretable fuzzy systems. In the following, some of them are briefly outlined, along with a few notes on specific applications and potentialities.

Environment. Environmental issues are often challenging because of the complex dynamics, the high number of variables and the consequent uncertainty characterizing the behavior of subjects under study. Computational Intelligence techniques come into play when tolerance for imprecision can be exploited to design convenient models that are suitable to understand phenomena and take decisions. Interpretable fuzzy systems show a clear advantage over black-box systems in providing knowledge that is capable of explaining complex and non-linear relationships by using linguistic models. Real-world environmental applications of interpretable fuzzy systems include: harmful bioaerosol detection [64]; modeling habitat suitability in river management [74]; modeling pesticide loss caused by meteorological factors in agriculture [40]; and so on.

Finance. This is a sector where human-computer cooperation is very tight. Cooperation is carried out in different ways, including the use of computers to provide business intelligence for decision support in financial operations. In many cases financial decisions are ultimately made by experts, who can benefit from automated analyses of big masses of data flowing daily in markets. To this pursuit, Computational Intelligence approaches are spreading among the tools used by financial experts in their decisions, including interpretable fuzzy systems for stock return predictions [50], exchange rate forecasting [25], portfolio risk monitoring [38], etc.

Industry. Industrial applications could take advantage from interpretable fuzzy systems when there is the need of explaining the behavior of complex systems and phenomena, like in fault detection [11]. Also, control plans for systems and processes can be designed with the aid of fuzzy systems. In such cases, a common practice is to start with an initial expert knowledge (used to design rules which are usually highly interpretable) that is then tuned to increase the accuracy of the controller. However, any unconstrained tuning could destroy the original interpretability of the knowledge base, whilst, by taking into account interpretability, the possibility of revising and modifying the controller (or the process manager) can be enhanced [66].

Medicine and Health-care. As a matter of fact, in almost all medical contexts intelligent systems can be invaluable decision support tools, but people are the ultimate actors in any decision process. As a consequence, people need to rely on intelligent systems, whose reliability can be enhanced if their outcomes may be

explained in terms that are comprehensible by human users. Interpretable fuzzy systems could play a key role in this area because of the possibility of acquiring knowledge from data and communicating it to users. In literature several approaches have been proposed to apply interpretable fuzzy systems in different medical problems, like assisted diagnosis [37], prognosis prediction [5], patient subgroup discovery [20], etc.

Robotics. The complexity of robot behavior modeling can be tackled by an integrated approach where a first modeling stage is carried out by combining human expert and empirical knowledge acquired from experimental trials. This integrated approach requires that the final knowledge base is provided to experts for further maintenance: this task could be done effectively only if the acquired knowledge is interpretable by the user. Some concrete applications of this approach can be found in robot localization systems [9] and motion analysis [8, 60].

Society. The focus of intelligent systems for social issues has noticeably increased in recent years. For reasons that are common to all the previous application areas, interpretable fuzzy systems have been applied in a wide variety of scopes, including Quality of Service improvement [15], data mining with privacy preservation [72], social network analysis [10], and so on.

7 Future research trends on interpretable fuzzy systems

Research on interpretable fuzzy systems is open in several directions. Future trends involve both theoretical and methodological aspects of interpretability. In the following, some trends are outlined amongst the possible lines of research development [6].

Interpretability definition. The blurred nature of interpretability requires continuous investigations on possible definitions that enable a computable treatment of this quality in fuzzy systems. This requirement casts the research on interpretable fuzzy systems towards cross-disciplinary investigations. For instance, this research line includes investigations on computable definitions of some conceptual qualities, like *vagueness* (which has to be distinguished from imprecision and fuzziness). Also, the problem of interpretability of fuzzy systems can be intended as a particular instance of the more general problem of communication between granular worlds [13], where many aspects of interpretability could be treated in a more abstract way.

Interpretability assessment. A prominent objective is the adoption of a common framework for characterizing and assessing interpretability with the aim of avoiding misleading notations. Within such a framework, novel metrics could be devised, especially for assessing subjective aspects of interpretability, and integrated with objective interpretability measures to define more significant interpretability indexes.

Design of interpretable fuzzy models. A current research trend in designing interpretable fuzzy models makes use of multi-objective genetic algorithms in or-

der to deal with the conflicting design objectives of accuracy and interpretability. The effectiveness and usefulness of these approaches, especially those concerning advanced schemes, has to be verified against a number of indexes, including indexes that integrate subjective measures. This verification process is particularly required when tackling high-dimensional problems. In this case the combination of linguistic and graphical approaches could be a promising approach for descriptive and exploratory analysis of interpretable fuzzy systems.

Representation of fuzzy systems. For very complex problems the use of novel forms of representation (different from the classical rule-based) may help in representing complex relationship in comprehensible ways thus yielding a valid aid in designing interpretable fuzzy systems. For instance, a multi-level representation could enhance the interpretability of fuzzy systems by providing different granularity levels for knowledge representation. On the one hand, the highest granulation levels give a coarse (yet immediately comprehensible) description of knowledge, while lower levels provide for more detailed knowledge.

As a final remark, it is worth observing that interpretability is one aspect of the multi-faceted problem of *human-centered* design of fuzzy systems [14]. Other facets include acceptability (e.g., according to ethical rules), interestingness of fuzzy rules, applicability (e.g., with respect to law), etc. Many of them are not yet in the research mainstream but they clearly represent promising future trends.

8 Conclusions

Interpretability is an indispensable requirement for designing fuzzy systems, yet it cannot be assumed to hold by the simple fact of using fuzzy sets for modeling. Interpretability must be encoded in some computational methods in order to drive the design of fuzzy systems, as well as to assess the interpretability of existing models. The study of interpretability issues started about two decades ago and led to a number of theoretical and methodological results of paramount value in fuzzy modeling. Nevertheless, research is still open both in depth — through new ways of encoding and assessing interpretability — and in breadth, by integrating interpretability in the more general realm of Human Centered Computing.

References

1. R. Alcalá, J. Alcalá-Fdez, J. Casillas, O. Cordón, and F. Herrera. Hybrid learning models to get the interpretability-accuracy trade-off in fuzzy modeling. *Soft Computing*, 10(9):717–734, 2006.
2. R. Alcalá, M. J. Gacto, and F. Herrera. A fast and scalable multi-objective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems. *IEEE Transactions on Fuzzy Systems*, 19(4):666–681, 2011.

3. R. Alcalá, Y. Nojima, F. Herrera, and H. Ishibuchi. Multiobjective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions. *Soft Computing*, 15(12):2303–2318, 2011.
4. J. Alcalá-Fdez, F. Herrera, F. Márquez, and A. Peregrín. Increasing fuzzy rules cooperation based on evolutionary adaptive inference systems. *International Journal of Intelligent Systems*, 22(4):1035–1064, 2007.
5. J. M. Alonso, C. Castiello, M. Lucarelli, and C. Mencar. Modelling interpretable fuzzy rule-based classifiers for medical decision support. In R. Magdalena, E. Soria, J. Guerrero, J. Gómez-Sanchis, and A.J. Serrano, editors, *Medical Applications of Intelligent Data Analysis: Research advancements*, pages 254–271. IGI Global, 2012.
6. J. M. Alonso and L. Magdalena. Editorial: Special issue on interpretable fuzzy systems. *Information Sciences*, 181(20):4331–4339, 2011.
7. J. M. Alonso, L. Magdalena, and G. González-Rodríguez. Looking for a good fuzzy system interpretability index: An experimental approach. *International Journal of Approximate Reasoning*, 51(1):115–134, 2009.
8. J. M. Alonso, L. Magdalena, S. Guillaume, M. A. Sotelo, L. M. Bergasa, M. Ocaña, and R. Flores. Knowledge-based intelligent diagnosis of ground robot collision with non detectable obstacles. *Journal of Intelligent and Robotic Systems*, 48(4):539–566, 2007.
9. J. M. Alonso, M. Ocaña, N. Hernandez, F. Herranz, A. Llamazares, M. A. Sotelo, L. M. Bergasa, and L. Magdalena. Enhanced WiFi localization system based on Soft Computing techniques to deal with small-scale variations in wireless sensors. *Applied Soft Computing*, 11(8):4677–4691, 2011.
10. J. M. Alonso, D. P. Pancho, O. Córdón, A. Quirin, and L. Magdalena. Social network analysis of co-fired fuzzy rules. In R. R. Yager, A. M. Abbasov, M. Reformat, and S. N. Shahbazova, editors, *Soft Computing: State of the Art Theory and Novel Applications*, pages 113–128. Springer-Verlag Berlin Heidelberg, 2013.
11. S. Altug, M.-Y. Chow, and H. J. Trussell. Heuristic constraints enforcement for training of and rule extraction from a fuzzy/neural architecture. Part II: Implementation and application. *IEEE Transactions on Fuzzy Systems*, 7(2):151–159, 1999.
12. M. Antonelli, P. Ducange, B. Lazzerini, and F. Marcelloni. Learning concurrently data and rule bases of Mamdani fuzzy rule-based systems by exploiting a novel interpretability index. *Soft Computing*, 15(10):1981–1998, 2011.
13. A. Bargiela and W. Pedrycz. *Granular computing: An introduction*. Kluwer Academic Publishers, Boston, Dordrecht, London, 2003.
14. A. Bargiela and W. Pedrycz. *Human-centric information processing through granular modelling*, volume 182 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, 2009.
15. F. Barrientos and G. Sainz. Interpretable knowledge extraction from emergency call data based on fuzzy unsupervised decision tree. *Knowledge-Based Systems*, 25(1):77–87, 2011.
16. A. Bastian. How to handle the flexibility of linguistic variables with applications. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(4):463–484, 1994.
17. U. Bodenhofer and P. Bauer. A formal model of interpretability of linguistic variables. In [23], pages 524–545, 2003.
18. U. Bodenhofer and P. Bauer. Interpretability of linguistic variables: A formal account. *Kybernetika*, 41(2):227–248, 2005.
19. A. Botta, B. Lazzerini, F. Marcelloni, and D. C. Stefanescu. Context adaptation of fuzzy systems through a multi-objective evolutionary approach based on a novel interpretability index. *Soft Computing*, 13(5):437–449, 2009.
20. C. J. Carmona, P. Gonzalez, M. J. del Jesus, M. Navio-Acosta, and L. Jimenez-Trevino. Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department. *Soft Computing*, 15(12):2435–2448, 2011.
21. J. Casillas, O. Córdón, and F. Herrera. COR: A methodology to improve ad hoc data-driven linguistic rule learning methods by inducing cooperation among rules. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 32(4):526–537, 2002.

22. J. Casillas, O. Cordón, F. Herrera, and L. Magdalena. *Accuracy improvements in linguistic fuzzy modeling*, volume 129 of *Studies in Fuzziness and Soft Computing*. Springer-Verlag, Heidelberg, 2003.
23. J. Casillas, O. Cordón, F. Herrera, and L. Magdalena. *Interpretability issues in fuzzy modeling*, volume 128 of *Studies in Fuzziness and Soft Computing*. Springer-Verlag, Heidelberg, 2003.
24. J. Casillas, P. Martínez, and A. D. Benítez. Learning consistent, complete and compact sets of fuzzy rules in conjunctive normal form for regression problems. *Soft Computing*, 13(5):451–465, 2009.
25. F. Cheong. A hierarchical fuzzy system with high input dimensions for forecasting foreign exchange rates. *International Journal of Artificial Intelligence and Soft Computing*, 1(1):15–24, 2008.
26. O. Cordón. A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems. *International Journal of Approximate Reasoning*, 52:894–913, 2011.
27. O. Cordón, F. Herrera, F. A. Márquez, and A. Peregrín. A study on the evolutionary adaptive defuzzification methods in fuzzy modeling. *International Journal of Hybrid Intelligent Systems*, 1(1):36–48, 2004.
28. O. Cordón, F. Herrera, and P. Villar. Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base. *IEEE Transactions on Fuzzy Systems*, 9(4):667–674, 2001.
29. P. Ducange and F. Marcelloni. Multi-objective evolutionary fuzzy systems. In A. M. Fanelli, W. Pedrycz, and A. Petrosino, editors, *Proceedings of the 9th International Workshop on Fuzzy Logic and Applications*, volume LNAI6857, pages 83–90, 2011.
30. Ch. Dujet and N. Vincent. Force implication: A new approach to human reasoning. *Fuzzy Sets and Systems*, 69(1):53–63, 1995.
31. J. Espinosa and J. Vandewalle. Constructing fuzzy models with linguistic integrity from numerical data-AFRELI algorithm. *IEEE Transactions on Fuzzy Systems*, 8(5):591–600, 2000.
32. P. Fazendeiro, J. Valente De Oliveira, and W. Pedrycz. A multiobjective design of a patient and anaesthetist-friendly neuromuscular blockade controller. *IEEE Transactions on Bio-medical Engineering*, 54(9):1667–1678, 2007.
33. M. Fazzolari, R. Alcalá, Y. Nojima, H. Ishibuchi, and F. Herrera. A review of the application of multi-objective evolutionary fuzzy systems: Current status and further directions. *IEEE Transactions on Fuzzy Systems*, 21(1):45–65, 2013.
34. A. Fiordaliso. A constrained Takagi-Sugeno fuzzy system that allows for better interpretation and analysis. *Fuzzy Sets and Systems*, 118(2):307–318, 2001.
35. M. J. Gacto, R. Alcalá, and F. Herrera. Integration of an index to preserve the semantic interpretability in the multiobjective evolutionary rule selection and tuning of linguistic fuzzy systems. *IEEE Transactions on Fuzzy Systems*, 18(3):515–531, 2010.
36. M. J. Gacto, R. Alcalá, and F. Herrera. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, 181(20):4340–4360, 2011.
37. I. Gadaras and L. Mikhailov. An interpretable fuzzy rule-based classification methodology for medical diagnosis. *Artificial Intelligence in Medicine*, 47(1):25–41, 2009.
38. A. Ghandar, Z. Michalewicz, and R. Zurbruegg. Enhancing profitability through interpretability in algorithmic trading with a multiobjective evolutionary fuzzy system. In C. A. Coello, V. Cutello, K. Deb, S. Forrest, G. Nicosia, and M. Pavone, editors, *Parallel Problem Solving from Nature*, volume LNCS7492, pages 42–51. Springer Berlin Heidelberg, 2012.
39. S. Guillaume. Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, 9(3):426–443, 2001.
40. S. Guillaume and B. Charnomordic. Interpretable fuzzy inference systems for cooperation of expert knowledge and data in agricultural applications using FisPro. In *IEEE International Conference on Fuzzy Systems*, pages 2019–2026, 2010.
41. M. M. Gupta and J. Qi. Design of fuzzy logic controllers based on generalized T-operators. *Fuzzy Sets and Systems*, 40(3):473–489, 1991.
42. F. Herrera. Genetic fuzzy systems: Taxonomy, current research trends and prospects. *Evolutionary Intelligence*, 1:27–46, 2008.

43. A. Homaifar and E. McCormick. Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. *IEEE Transactions on Fuzzy Systems*, 3(2):129–139, 1995.
44. E. Hüllermeier. Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems*, 156(3):387–406, 2005.
45. E. Hüllermeier. Fuzzy sets in machine learning and data mining. *Applied Soft Computing*, 11(2):1493–1505, 2011.
46. H. Ishibuchi, T. Murata, and I. B. Türksen. Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems. *Fuzzy Sets and Systems*, 89(2):135–150, 1997.
47. H. Ishibuchi, T. Nakashima, and T. Murata. Three-objective genetics-based machine learning for linguistic rule extraction. *Information Sciences*, 136(1-4):109–133, 2001.
48. H. Jones, B. Charnomordic, D. Dubois, and S. Guillaume. Practical inference with systems of gradual implicative rules. *IEEE Transactions on Fuzzy Systems*, 17(1):61–78, 2009.
49. J. Kiszka, M. Kochanska, and D. Sliwiska. The influence of some fuzzy implication operators on the accuracy of a fuzzy model - Parts I and II. *Fuzzy Sets and Systems*, 15:111–128, 223–240, 1985.
50. A. Kumar. Interpretability and mean-square error performance of fuzzy inference systems for data mining. *Intelligent Systems in Accounting, Finance and Management*, 13(4):185–196, 2005.
51. B.-D. Liu, C.-Y. Chen, and J.-Y. Tsao. Design of adaptive fuzzy logic controller based on linguistic-hedge concepts and genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 31(1):32–53, 2001.
52. E. H. Mamdani. Application of fuzzy logic to approximate reasoning using linguistic synthesis. *IEEE Transactions on Computers*, 26(12):1182–1191, 1977.
53. P. A. Mastorocostas, J. B. Theocharis, and V. S. Petridis. A constrained orthogonal least-squares method for generating TSK fuzzy models: Application to short-term load forecasting. *Fuzzy Sets and Systems*, 118(2):215–233, 2001.
54. C. Mencar, G. Castellano, and A. M. Fanelli. Distinguishability quantification of fuzzy sets. *Information Sciences*, 177(1):130–149, 2007.
55. C. Mencar, C. Castiello, R. Cannone, and A. M. Fanelli. Design of fuzzy rule-based classifiers with semantic cointension. *Information Sciences*, 181(20):4361–4377, 2011.
56. C. Mencar, C. Castiello, R. Cannone, and A. M. Fanelli. Interpretability assessment of fuzzy knowledge bases: A cointension based approach. *International Journal of Approximate Reasoning*, 52(4):501–518, 2011.
57. C. Mencar and A. M. Fanelli. Interpretability constraints for fuzzy information granulation. *Information Sciences*, 178(24):4585–4618, 2008.
58. R. S. Michalski. A theory and methodology of inductive learning. *Artificial Intelligence*, 20(2):111–161, 1983.
59. G. A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63:81–97, 1956.
60. M. Mucientes and J. Casillas. Quick design of fuzzy controllers with good interpretability in mobile robotics. *IEEE Transactions on Fuzzy Systems*, 15(4):636–651, 2007.
61. D. P. Pancho, J. M. Alonso, O. Cordon, A. Quirin, and L. Magdalena. FINGRAMS: Visual representations of fuzzy rule-based inference for expert analysis of comprehensibility. *IEEE Transactions on Fuzzy Systems*, DOI:10.1109/TFUZZ.2013.2245130, 2013.
62. D. P. Pancho, J. M. Alonso, and L. Magdalena. Quest for interpretability-accuracy trade-off supported by fingrams into the fuzzy modeling tool GUAJE. *International Journal of Computational Intelligence Systems*, 6(sup1):46–60, 2013.
63. W. Pedrycz and F. Gomide. *An Introduction to Fuzzy Sets. Analysis and Design*. The MIT Press, Cambridge (MA), 1998.
64. P. Pulkkinen, J. Hytonen, and H. Koivisto. Developing a bioaerosol detector using hybrid genetic fuzzy systems. *Engineering Applications of Artificial Intelligence*, 21(8):1330–1346, 2008.

65. A. Riid and E. Rüstern. Transparent fuzzy systems in modelling and control. In [23], pages 452–476, 2003.
66. A. Riid and E. Rüstern. Interpretability of fuzzy systems and its application to process control. In *IEEE International Conference on Fuzzy Systems*, pages 1–6, 2007.
67. A. Riid and E. Rüstern. Identification of transparent, compact, accurate and reliable linguistic fuzzy models. *Information Sciences*, 181(20):4378–4393, 2011.
68. T. L. Saaty and M. S. Ozdemir. Why the magic number seven plus or minus two. *Mathematical and Computer Modelling*, 38(3-4):233–244, 2003.
69. M. Setnes, R. Babuška, U. Kaymak, and H. R. van Nauta Lemke. Similarity measures in fuzzy rule base simplification. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 28(3):376–386, 1998.
70. T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modelling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:116–132, 1985.
71. E. Trillas and L. Valverde. On implication and indistinguishability in the setting of fuzzy logic. In J. Kacprzyk and R. R. Yager, editors, *Management Decision Support Systems Using Fuzzy Logic and Possibility Theory*, pages 198–212. Verlag TUV Rheinland, 1985.
72. L. Troiano, L. J. Rodríguez-Muñiz, J. Ranilla, and I. Díaz. Interpretability of fuzzy association rules as means of discovering threats to privacy. *International Journal of Computer Mathematics*, 89(3):325–333, 2012.
73. J. Valente de Oliveira. Semantic constraints for membership function optimization. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 29(1):128–138, 1999.
74. E. Van Broekhoven, V. Adriaenssens, and B. de Baets. Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: An ecological case study. *International Journal of Approximate Reasoning*, 44(1):65–90, 2007.
75. L.-X. Wang and J. M. Mendel. Fuzzy basis functions, universal approximation, and orthogonal least squares learning. *IEEE Transactions on Neural Networks*, 3:807–814, 1992.
76. L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.
77. L. A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(1):28–44, 1973.
78. L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning, Parts I, II, and III. *Information Sciences*, 8,8,9:199–249,301–357,43–80, 1975.
79. L. A. Zadeh. From computing with numbers to computing with words - from manipulation of measurements to manipulation of perceptions. *IEEE Transactions on Circuits and Systems - I: Fundamental theory and applications*, 45(1):105–119, 1999.
80. L. A. Zadeh. Is there a need for fuzzy logic? *Information Sciences*, 178(13):2751–2779, 2008.
81. S.-M. Zhou and J. Q. Gan. Low-level interpretability and high-level interpretability: A unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets and Systems*, 159(23):3091–3131, 2008.