Published journal article is available online at https://doi.org/10.1016/J.INS.2014.05.023

Comparison and Design of Interpretable Linguistic vs. Scatter FRBSs: GM3M Generalization and New Rule Meaning Index for Global Assessment and Local Pseudo-Linguistic Representation☆

Marta Galende[a], María José Gacto[b], Gregorio Sainz[a,c], Rafael Alcalá[d]

[a]CARTIF Centro Tecnológico. Parque Tecnológico de Boecillo, 47151 Boecillo (Valladolid), Spain.
[b]Dept. of Computer Science, University of Jaén, 23071 Jaén, Spain.
[c]Dept. of Systems Engineering and Control, University of Valladolid, 47011 Valladolid, Spain.
[d]Dept. of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain.

Abstract

This work is devoted to defining more general interpretability indexes to be applied to any scatter or linguistic model implemented by any type of membership functions. They are based on metrics that should take into account the semantic and inference issues: the semantic issue in order to preserve the meaning of the linguistic labels and the inference issue since this can influence the behavior of the rules. On the other hand, these metrics have been designed to be intuitive in order to support the analysis or selection of a final model and to favor a low computational cost within an optimization process.

In order to check their usefulness, a multi-objective evolutionary algorithm, simultaneously performing a rule selection and an adjustment of the fuzzy partitions, is guided by the proposed indexes on several benchmark data sets to obtain models with different degrees of accuracy and interpretability. In addition, using these metrics, a local analysis can be carried out between models of a different nature. This local analysis through the model components, gives support to the user to make the best choice from amongst the models.

Keywords:
Fuzzy rule-based systems, Linguistic or Scatter fuzzy systems, Interpretability, Semantic indexes

1. Introduction

The generation and use of models as a way to capture real-world notions is inherent to human progress from its origins [53]. According to [26], models can be a representation of a selected part of the world, either a phenomena or data. Specifically, a data model is defined as *"a corrected, rectified, regimented, and in many instances idealized version of the data we gain from immediate observation, the so-called raw data"*. Nowadays, considering a general point of view, a model can be featured by two main achievements:

- To reproduce accurately the behavior of the real-world notion modelled.
- To explain the knowledge learnt/captured about the real-world involved in the model.

These properties are generally known as the **accuracy** and **interpretability** of the model. To achieve models with both features is a challenge since, according to the Principle of Incompatibility of Zadeh [57],

☆Supported by the European Research Council GA no. 314031, Spanish Project no. DPI2012-39381-C02-02 and no. TIN2012-33856 and the Andalusian Excellence Regional Project no. P10-TIC-6858.
Email addresses: margal@cartif.es (Marta Galende), mgacto@ujaen.es (María José Gacto), gresai@cartif.es,gresai@eii.uva.es (Gregorio Sainz), alcala@decsai.ugr.es (Rafael Alcalá)

both aspects are contradictory. Therefore, when interpretability becomes essential for a given problem, it is necessary to find a reasonable balance between accuracy and interpretability, and fuzzy modelling, i.e. system modelling by Fuzzy Rule-Based Systems (FRBSs), can help to improve this balance [14, 13] since FRBSs make use of a descriptive language based on fuzzy logic with fuzzy predicates.

However, it is still necessary to analyze how to define and evaluate both features within FRBSs. Accuracy, or the capacity to faithfully represent the real system/world, is a well-known concept based on widely accepted error measures. But the definition of interpretability and how this can be evaluated is currently an open question [29, 7, 58, 46].

An organized framework can be established based on several reviews of the state-of-the-art devoted to the interpretability of FRBSs: A complete analysis of the main interpretability constraints and their associated formulations was carried out by Mencar and Fanelli in [46]; a taxonomy of interpretability at two levels, *Low-level* and *High-level*, for FRBSs is proposed by Zhou and Gan in [58]; another general review focused on the interpretability of the FRBS was published by Alonso et al. in [7]; finally, an exhaustive review including all the existing methodologies and metrics to assess the interpretability of linguistic FRBSs was presented by Gacto et al. in [29], proposing a taxonomy for two types of interpretability, based on *Complexity* and based on *Semantic*, to be evaluated on two main components of a FRBS, the Rule Base (RB) and the fuzzy partitions or Data Base (DB).

Taking into account the previous reviews and after the exhaustive study presented in [29], the authors pointed out that there is no single comprehensive measure to quantify the interpretability of linguistic FRBSs. This is, to get a good global interpretability assessment it would be necessary to consider appropriate measures to quantify each of these four aspects (*Complexity* and *Semantic* at RB and DB levels, respectively), since they take into account different (even contradictory) interpretability properties that should be required to easily interpret these kinds of systems. Moreover, the authors showed that while there are widely accepted metrics to quantify *Complexity* for both RB and DB (as number of rules, number of conditions or a maximum number of linguistic terms), there is no agreement regarding the choice of appropriate measures to quantify the *Semantic* interpretability for both the RB and DB since the existent ones are not able to consider the problem context when it can be expressed by an expert or they were devoted to particular types of Membership Functions (MFs), rules or inference systems. Moreover, from these reviews we can conclude that linguistic-based, or Mamdani, FRBSs have never been effectively compared to scatter-based or approximative FRBSs. Linguistic FRBSs [33] are based on linguistic rules, in which the antecedent and the consequent make use of linguistic variables comprised of linguistic terms and the associated fuzzy sets defining their meanings, while scatter-based FRBSs [33] differ from the linguistic ones in the use of fuzzy variables, i.e. fuzzy sets without a pre-established associated meaning. They are quite difficult to compare due to the lack of general interpretability indexes that could be applied independently of the type of FRBS, the type of inference system, and even the type of MFs, to quantify both semantic interpretability types that should be considered in a given FRBS, at the fuzzy partition level and the RB level [29].

This contribution presents a first approximation to face these problems, proposing more general *Semantic* interpretability metrics for both, fuzzy partition and RB level, that could be combined with any of the well-established complexity metrics and used with any type of FRBS, independently of its linguistic or scatter nature, and the shape of the MFs used, whether they be triangular, trapezoidal or gaussian. These metrics should take into account the semantic meaning of the linguistic labels provided by an expert, when this is available, as well as the inference system, since this can influence the rule behavior. All of this should be managed by intuitive metrics in order to favor a low computational cost within an optimization process and to allow a further analysis for comparing or selecting the final FRBS. The main aim is to allow a unified complexity-semantic treatment of both linguistic and scatter-based FRBSs.

In order to address these objectives, this new approximation is based on two main proposals: a generalization of the Gm3m relative index [28] for the semantic interpretability at the fuzzy partition level, and the definition of a new semantic interpretability index for the RB level, namely Rule Meaning Index (Rmi). The new metrics proposed to define these indexes could be used, in combination with the appropriate complexity metrics, to systematically approach the design of more interpretable FRBSs or to perform a local analysis of the interpretability of the system components, which allows several FRBSs obtained by any modeling approach to be compared independently of their nature, and helps the user to make the final

FRBS choice. To do this, we also propose a local-pseudolinguistic representation of the scatter-based FRBSs, which allows a possible local comparison with the linguistic-based ones.

In order to validate the proposed metrics we consider a possible example derivation and comparison scenario. Nevertheless, any FRBS modeling technique could be considered to obtain and compare different models by combination of the proposed metrics with the appropriate widely accepted complexity and accuracy metrics, or by following a multi-objective framework. In this way, different derivation techniques could be considered such as, neural [47, 12], genetic [40], swarm intelligence [43], iterative rule learning [18], etc., as well as different modeling approaches such as the use of DNF-type MFs or feature selection [8]. Because of simplicity and for better focusing on showing the utility of the proposed metrics we will consider standard MF-type and a multi-objective evolutionary approach without feature selection in order to optimize accuracy and interpretability of FRBSs obtained by different derivation techniques. Since no feature selection is adopted for the study, we can consider the most commonly used complexity metric, i.e., the number of rules. In this sense, a Multi-Objective Evolutionary Algorithm (MOEA), simultaneously performing a rule selection and an adjustment of the fuzzy partitions, has been developed and guided by an error measure, the said complexity metric and the proposed indexes, in order to obtain FRBSs with different degrees of accuracy and interpretability. The aim is not to obtain a new winning algorithm, but rather to show the usefulness of the proposed indexes when they are applied with common algorithms of the state-of-the-art.

The proposed MOEA devoted to improving accuracy, together with the complexity and the proposed indexes, has been applied to nine real-world problems from the KEEL dataset repository on the initial models obtained by four fuzzy modeling algorithms, two linguistic-based ones, NefProx (*Neuro-Fuzzy Function Approximation*) [47] and L-IRL (*Linguistic Iterative Rule Learning*) [17], and two scatter-based ones, FasArt (*Fuzzy Adaptive System ART based*) [12] and S-IRL (*Scatter Iterative Rule Learning*) [18]. All these algorithms generate initial FRBSs of different types, so the proposed indexes can be validated for the global assessment of interpretability in the learning process at different application scenarios. Finally, local pseudo-linguistic representation is used to locally compare a given scatter-based model vs. a given linguistic-based one in two example problems, as the proposed methodology, to evaluate and compare two different fuzzy models for the final user selection.

The rest of the paper is organized as follow: In Section 2, a brief summary of interpretability concepts in FRBSs is given. Then, in Section 3, the two new indexes to measure the semantic interpretability are proposed. In Section 4, the MOEA used to optimize these indexes is described, and in Section 5, the results of the experiments are discussed. Finally, in Section 6, the most interesting conclusions are set out.

2. An Introduction to the Interpretability of FRBSs

Different terminology has been used by the authors to refer to the concept of interpretability in fuzzy systems. Concepts such as readability, transparency, intelligibility, comprehensibility, understandability, etc., have been widely associated to the idea of interpretability [46]. In the scientific literature we can find many definitions of "interpretability": Bodenhofer and Bauer in [10] define interpretability as the "*possibility to estimate the system's behavior by reading and understanding the rule base only*"; Mencar and Fanelli in [46] establish that "*A model is interpretable if its behavior is intelligible, i.e. it can be easily perceived and understood by a user*"; later, Gacto et al. in [29] define interpretability as "*the capacity to express the behavior of the real system in an understable way*".

According to these definitions, there is not a single global definition of interpretability for a FRBS. The concept of interpretability is in part subjective and it directly depends on the person in charge of dealing with the system. So, it is not possible to address the concept of interpretability in a single way. In the last decade, several works have analyzed the interpretability challenge for FRBSs, looking for interpretability measures that could be universally accepted by the research community [32, 48, 54]. This effort has continued in recent years and some review papers [46, 58, 7, 29] provide a well-established framework devoted to interpretability concepts and formulations.

In particular, Gacto et al. present an exhaustive review including all the existing methodologies and metrics to assess the interpretability of linguistic FRBSs in [29]. As a consequence, they propose a tax-

onomy based on four quadrants, which represent different aspects that should be considered to assess the interpretability of linguistic FRBSs (see Table 1): Complexity at the RB level (Q_1), Complexity at the fuzzy partition level (Q_2), Semantics at the RB level, (Q_3) and Semantics at the fuzzy partition level (Q_4). Thereby, the different measures or constraints described in the specialized literature are fitted into a common framework.

Table 1: A taxonomy to analyze the interpretability of linguistic FRBSs

	Rule Base level	Fuzzy Partition level
Complexity-based Interpretability	Q_1 Number of rules Number of conditions	Q_2 Number of Membership Functions Number of Features
Semantic-based Interpretability	Q_3 Consistency of rules Rules fired at the same time Transparency of rule structure (rule weights, etc.) Cointension	Q_4 Completeness or Coverage Normalization Distinguishability Complementarity Relative measures

Early works used the **number of rules** [39, 38] as a measure to reduce the complexity of the model and to obtain a good, or better, trade-off between accuracy and complexity [3, 27]. The **number of conditions** (sometimes used in combination with the number of rules) has also been used to minimize the length of the rules [40, 16, 2, 8]. Related to the complexity of the fuzzy partition, the **number of membership functions** [19, 1] and the **number of features** [5] have also been considered.

However, recent works propose additional complementary metrics taking into account the semantics related with different components of the FRBS. Regarding the **semantic-based interpretability at the RB level**, the most commonly used metrics are based on rule consistency, in terms of redundant and inconsistent rules [5, 52, 6, 50]. More recent works use other aspects like the number of rules fired at the same time [42, 49] or the cointension [45, 44].

Regarding the **semantic-based interpretability at fuzzy partition level**, the classic semantic restrictions of distinguishability, natural zero positioning, normality and coverage have already been proposed by Oliveira in [48]. These restrictions had been widely used by other authors [24, 51]. In the last few years, the metrics to assess these properties have become more complex [11, 50], although most of them get their best absolute value when the DB is composed of strong uniformly distributed fuzzy partitions (which satisfy most of the above properties).

On the other hand, some relative measures have been recently proposed, such as GM3M [28] and the integrity index I [9]. These metrics consider that interpretability is dependent on the problem context and user perceptions, so they allow accuracy improvements while trying to keep partitions and meanings as much as possible to their original values (an interpretable definition of the MFs provided either by an expert or by a machine learning process probably based on absolute measures or directly considering strong uniformly distributed fuzzy partitions). To do so, these metrics have also been combined with classic complexity measures such as the number of rules [28] or the number of conditions [9].

Summarizing all of this, it seems evident that the complexity-based measures are traditionally the most commonly used and accepted measures, but they cover only a part of the concept of interpretability since they do not consider semantic aspects. On the other hand, semantic-based interpretability is nowadays an open problem almost exclusively addressed to linguistic FRBSs. However, as stated in [29], to get a good global assessment it would be necessary to consider appropriate measures from all of the four quadrants, in order to take into account the different interpretability properties required for these kinds of systems together.

Additionally, an important aspect also affecting the interpretability of a FRBS is the **inference mechanism** used by the FRBS and the associated fuzzy operators. In [40] it has already been mentioned that the interpretability of the FRBS depends on, among other factors, the simplicity of fuzzy reasoning. To our knowledge, there are not measures taking into account the influence of these mechanisms, although they affect the semantic-based interpretability at the RB level, running the way in which the rules interact.

Finally, there is an upcoming open problem related to the interpretability of type-2 FRBSs. While there are many works related to tackle the interpretability of type-1 FRBSs, the interpretability of type-2 FRBSs has been vaguely taken into account since nowadays they still represent an emerging new trend where some works considering the footprint of uncertainty or optimal granularity allocation [15, 35] represent interesting approximations for designing clear and compact type-2 fuzzy inference systems. In our opinion, further extensions for the applicability of the existent interpretability metrics will be an interesting upcoming challenge.

3. A Proposal for Assessing the Semantic Interpretability of FRBSs: Linguistic vs. Scatter

In order to effectively evaluate and compare the interpretability of any FRBS, it is necessary to define metrics to measure both types of semantic interpretability independently of the system's fuzzy nature (linguistic or scatter), the type of MFs or the inference system used. Taking into account the analisys from the previous section, we have that: The most novel interpretability metrics at the fuzzy partition level are relative metrics such as the Gm3m index [28], but this is exclusively defined for linguistic FRBSs based on triangular MFs. The smaller number of contributions are for semantic interpretability at the RB level, where we can usually find computationally expensive measures that do not take the inference system into account.

This section proposes a generalization of the Gm3m index (subsection 3.1) for measuring semantic interpretability at the level of fuzzy partitions for any type of MFs and for scatter-based FRBSs, as well as a new intuitive and easy to compute index for semantic interpretability at the RB level (subsection 3.2), namely Rmi, that takes the inference system into account. Finally, a methodology to locally compare linguistic and scatter-based RBs is introduced by considering the latter as pseudo-linguistic FRBSs, subsection 3.3.

3.1. Generalizing the GM3M Index for Relative Semantic Interpretability at the Level of Fuzzy Partitions

Gm3m [28] is a knwon index devoted to quantifying the interpretability of the tuned DB of a linguistic fuzzy model based on triangular MFs with respect to a previous interpretable linguistic partition (obtained from experts, automatic methods or, as was considered in [28], by a strong uniformly distributed fuzzy partition). In this work, this index is extended to measure the semantic interpretability of MFs, regardless of their shape or type: triangular, trapezoidal, gaussian, linguistic or scatter-based, etc. In order to present this extension, a brief overview regarding MF tuning and the Gm3m index is performed and, finally, the generalization proposed for this index is introduced.

3.1.1. Preliminaries: Tuning Membership Functions and original Gm3m Index

Tuning MFs involves refining the MF shapes from a previous definition once the remaining FRBS components have been obtained [34, 41]. The classic way to refine the MFs is to change their definition parameters. For example, if the triangular-shape MF shown in Figure 1 is considered, changing the basic parameters — a, b, and c — will vary the shape of the fuzzy set associated with the MF, thus influencing the FRBS performance. The same is true for other shapes of MFs (trapezoidal, Gaussian, etc.).

In the case of linguistic FRBSs, tuning involves fitting the characterization of the MFs associated with the primary linguistic terms considered in the system. Thus, the meaning of the linguistic terms is changed from a previous definition (an initial DB comprised of the semantic concepts and the corresponding MFs giving meaning to them). In order to preserve the initial semantic interpretability of linguistic fuzzy systems throughout the MFs optimization process, a relative index (namely *GM3M*) has been proposed in [28].

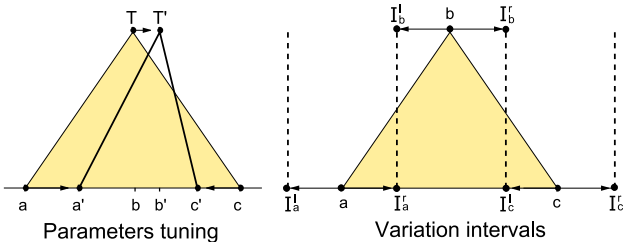


Figure 1: Tuning by changing the basic MF parameters and the variation intervals

GM3M is defined as the geometric mean of three metrics, and its values range between 0 (the lowest level of interpretability) and 1 (the highest level of interpretability). The index is defined as:

$$GM3M = \sqrt[3]{\delta \cdot \gamma \cdot \rho}$$

where δ, γ and ρ are three complementary metrics to measure interpretability when a tuning is performed on the MFs. The geometric mean ensures small values of *GM3M* when any of the metrics actually has low values (interpretability). Each metric was proposed for working with triangular MFs within a linguistic framework to measure: MFs displacement (δ), MFs lateral amplitude rate (γ) and MFs area similarity (ρ), respectively.

Let us represent the definition parameters of the original and the tuned triangular MFj as (a_j, b_j, c_j) and (a'_j, b'_j, c'_j), and their variation intervals as $[I^l_{a_j}, I^r_{a_j}]$, $[I^l_{b_j}, I^r_{b_j}]$ and $[I^l_{c_j}, I^r_{c_j}]$, respectively. These intervals determine the maximum variation for each parameter and could be defined in a different way for different problems.

The δ metric is used to control the displacements in the central point of the MFs. It is based on computing the normalized distance between the central points of the tuned and the original MF, and it is calculated through obtaining the maximum distance from all the MFs. For each MFj in the linguistic fuzzy partition, we define $\delta_j = |b_j - b'_j|/I$, where $I = (I^r_{b_j} - I^l_{b_j})/2$ represents the maximum variation for each central parameter. Thus δ^* is defined as $\delta^* = max_j\{\delta_j\}$ (the worst case). δ^* takes values between 0 and 1 (values near to 1 show that the MFs present a great displacement). The following transformation is made so that this metric represents proximity (maximization): Maximize $\delta = 1 - \delta^*$.

The γ metric is used to control the MF shapes. It is based on relating the left and right parts of the support of the original and the tuned MFs. Let us define $leftS_j = |a_j - b_j|$ and $rightS_j = |b_j - c_j|$ as the amplitude of the left and the right parts of the original MF support, and $leftS'_j = |a'_j - b'_j|$ and $rightS'_j = |b'_j - c'_j|$ as the corresponding parts in the tuned MFs. γ_j is calculated using the following equation for each MFj: $\gamma_j = min\{leftS_j/rightS_j\ ,\ leftS'_j/rightS'_j\}/max\{leftS_j/rightS_j\ ,\ leftS'_j/rightS'_j\}$. Values near to 1 mean that the left and right rates are highly maintained in the tuned MFs. Finally, γ is calculated by obtaining the minimum value of γ_j (the worst case): Maximize $\gamma = min_j\{\gamma_j\}$.

The ρ metric is used to control the area of the MF shapes. It is based on relating the areas of the original and the tuned MFs. Let us define A_j as the area of the triangle representing the original MFj, and A'_j as the new area. ρ_j is calculated using the following equation for each MF: $\rho_j = min\{A_j, A'_j\}/max\{A_j, A'_j\}$. Values near to 1 mean that the original area and the tuned area of the MFs are more similar (less changes). The ρ metric is calculated by obtaining the minimum value of ρ_j (the worst case): Maximize $\rho = min_j\{\rho_j\}$.

3.1.2. Gm3m *Index Generalization*

Following the philosophy set out when Gm3m was defined, the first step is to associate a given MF (independently of its type or shape) to its corresponding original or interpretable MF (defined by experts, automatic methods or considering strong uniformly distributed linguistic partitions). Since now different types of MFs (trapezoidal, gaussian, etc.) or FRBSs (scatter or linguistic) should be able to be considered, this association must be redefined to calculate GM3M. This is done using the mid points of the $\alpha - cut$ from both MFs with $\alpha = 0.5$. A given MF is associated to the original MF whose $0.5 - cut$ mid point is the nearest to its $0.5 - cut$ mid point.

In Figure 3, some examples are shown. In the first example, the MF "?" is associated with "M" since the distance between the mid points of their $0.5-cut$ is the lowest with respect to the remaining MF candidates. This association is preferable to directly using the cores because of the similarity between the given MF and the fact that "M" is higher that of "S". In the second example, MF "?" is associated to "M".

Thus for each given MF' of the FRBS involved in the calculation of the GM3M value, its corresponding original or interpretable MF is associated. As is shown, several given MF' can now be associated with the same original or interpretable MF, so that the association is no longer one to one. Therefore, in the rest of the section, we remove subindices in the formulation, using only primes for the given MFs in the FRBS.

In accordance with GM3M methodology, the next step is to determine the feature points and the variation intervals of the MFs (see Figure 2) in order to calculate δ and the remaining GM3M metrics.

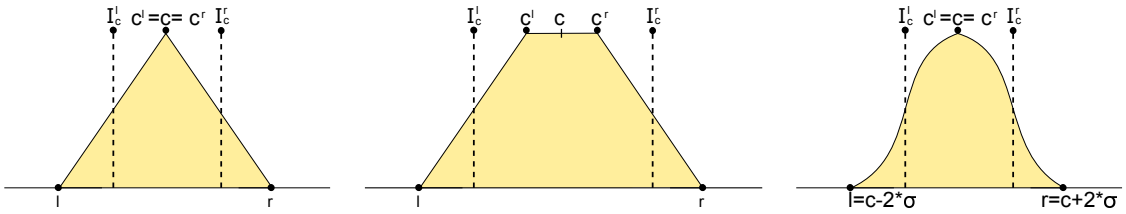


Figure 2: Definition parameters and core interval for the different MF types

To do this, the associated original MF must be parameterized according to its shape:

- Triangular: (l, c, r) are the left, central and right definition points of the MF, while c^l and c^r are the position of the left and right definition points of the MF core. In this case both are equal to the central point $c^l = c^r = c$, i.e., the characteristic position of the MF.
- Trapezoidal: (l, c^l, c^r, r) are respectively the left, left central, right central and right definition points of the MF, and $c = (c^l + c^r)/2.0$ is the centre of the core, i.e., the characteristic position of the MF.
- Gaussian: (c, σ) are respectively the central point and width of the MF, while c^l and c^r are the position of the left and right definition points of the MF core. In this case, both are equal to the central point $c^l = c^r = c$, i.e., the characteristic position of the MF. Finally, let us define $l = c - 2\sigma$ and $r = c + 2\sigma$.

Once the original MF has been parametrized, the corresponding variation interval $[I_c^l, I_c^r]$ is defined by $I_c^l = (l + c^l)/2.0$ and $I_c^r = (c^r + r)/2.0$.

For a given MF', $\delta' = |c - c'|/I$, $I = max(c - I_c^l, I_c^r - c)$ represents the maximum allowed variation for the central parameter. Thus δ^* is defined as $\delta^* = max\{\delta'\}$ (the worst case). Notice that δ^* could take negative values. The following transformation is made so that this metric represents proximity (maximization): Maximize $\delta = 1 - \delta^*$. In order to quantify the interpretability of a particular MF, we can consider 0 as a limit, but we should maintain negative values for optimization purposes.

The γ' value is computed in the same way, taking into account the slopes of the corresponding MFs. Let us define $leftS = |c^l - l|$ as the amplitude of the left part of the associated MF support, and $rightS = |r - c^r|$ as the right part amplitude; and $leftS' = |c'^l - l'|$ and $rightS' = |r' - c'^r|$ as the corresponding parts in the new MF' under consideration. Then, γ is calculated by the same formula in section 3.1.1: Maximize $\gamma = min\{\gamma'\}$.

The ρ metric is modified to allow non singleton cores, such as trapezoidal MFs. Let us consider A_s as the area of the slopes and A_c as the area of the core. Since they represent different conceptual parts, they should be considered separately in order to detect changes in any of them. Notice that A_c will be equal to zero for the case of triangular and gaussian MFs. Thus, ρ' is calculated for each MF' using the following equation: $\rho' = (min\{A_s, A'_s\} + min\{A_c, A'_c\})/(max\{A_s, A'_s\} + max\{A_c, A'_c\})$. Values near 1 mean that the original MF area and the new area are more similar (less changes). The ρ metric is calculated by obtaining the minimum value of ρ' (the worst case): Maximize $\rho = min\{\rho'\}$.

In this way, we can compare FRBSs with any kind of MFs. That is, we could have a triangular based partition defined by an expert and measure how different the trapezoidal MFs obtained by the learning

algorithms are. If trapezoidal MFs are not so different to triangular ones, GM3M will present high values. See Figure 3 with an example of different cases. Both examples include detailed calculations of GM3M for a better understanding. In Example 1, the GM3M value is 0.545, representing a displacement of $\delta = 0.562$, with a lateral amplitude rate of $\gamma = 0.330$ and an area similarity of $\rho = 0.872$, which means that the tuned MF is not totally different from its original MF. In Example 2, the GM3M value is 0.000, which shows that MF M' has a poor semantic interpretability because of a large displacement (distance) between MFs ($\delta = 0.000$).

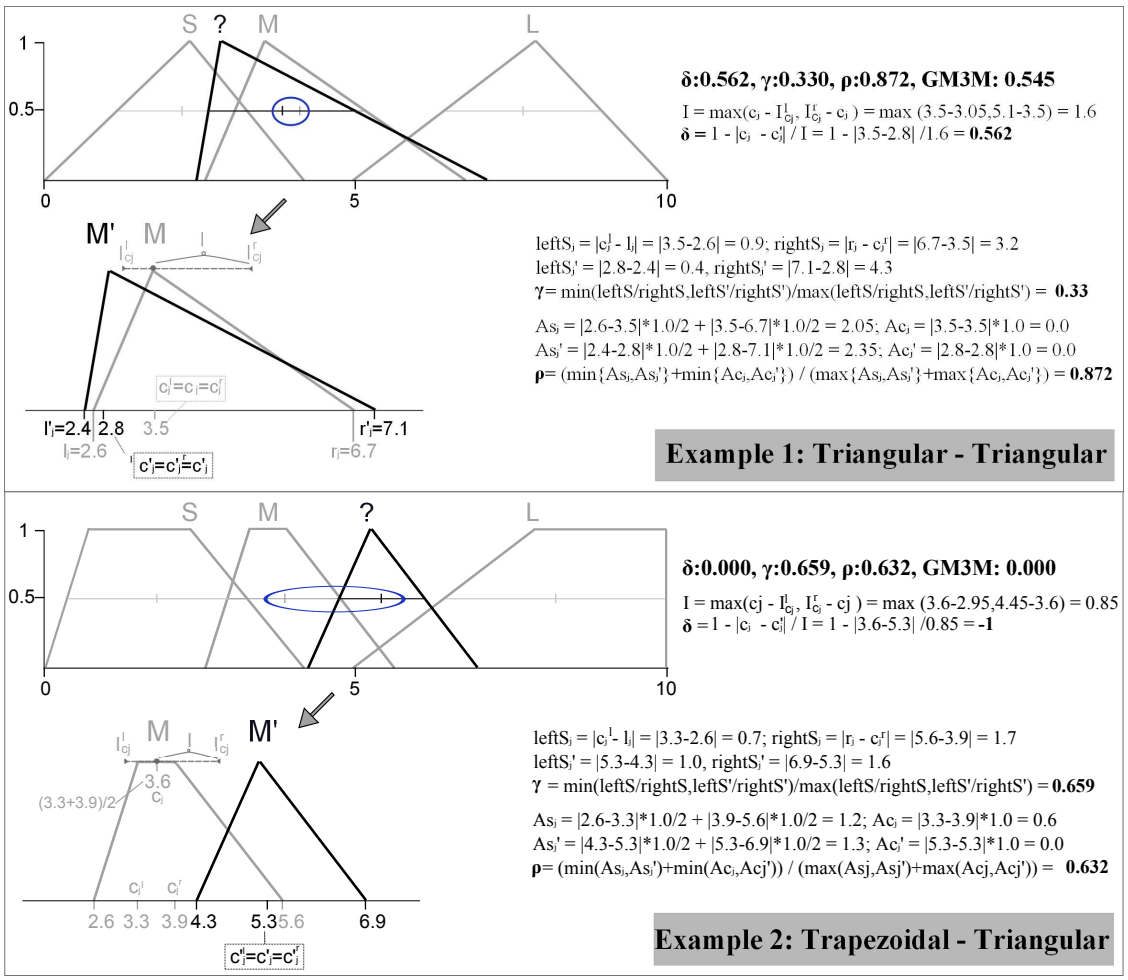


Figure 3: GM3M example cases

In order to analyze the performance of this proposal, the following studies will be focused on triangular MFs. Moreover, we also consider original strong fuzzy partitions with uniform MFs for these studies because they are usually considered the most interpretable ones [29].

3.2. A New Semantic Interpretability Index at the Level of Rule Base: The Rule Meaning Index (RMI)

Semantic-based interpretability at the RB level is usually controlled by properties such as consistency or number of rules simultaneously fired [29]. Here, our objective is to propose a new index to evaluate the semantic interpretability of any type of (linguistic or scatter) RB, but taking into account the fact that the real problem is not the number of rules interacting, but how they are contradictory to the system output when they are fired. That is, redundancy (rules simultaneously fired with similar consequents) is not a semantic problem but a complexity one, since each of these rules is consistent with the global output of the system at

its activation region. Moreover, as previously said, this should also take into account the inference system, since this affects the system output, and the computational cost should be low.

This section is devoted to proposing this new index, taking all of this into account in order to measure the semantic interpretability at RB level: the Rule Meaning Index (RMI). This index is based on computing an individual RMI value for each rule, R_i in the RB (RMI(R_i)) in order to calculate the final global RMI value for a given FRBS. The goal of RMI(R_i) is to evaluate the degree of reliability of the rule R_i in comparison to what the complete model would say for the activation area of that rule, which takes into account the FRBS inference through the estimated output.

The structure of each rule R_i is defined as follows:

$$R_i : \text{If } X_{i1} \text{ is } A_{i1} \text{ and ... and } X_{in} \text{ is } A_{in} \text{ then } Y \text{ is } B_i \text{ (or } C_i \text{ in the case of classification)},$$

with X and Y being the input and output variables respectively, n the number of antecedents and A_{i1} to A_{in} and B_i fuzzy sets (or C_i being a class).

Let us define O_{R_i} as the FRBS output, which makes use of all the rules when the input is defined as the core of the fuzzy set of each antecedent A_{ik} ($k = 1..n$) for rule R_i. Then the definition of the index for regression or control problems is:

$$\text{RMI}(R_i) = \mu_{Cons_{R_i}}(O_{R_i})$$

where $\mu_{Cons_{R_i}}(O_{R_i})$ is the degree of membership of O_{R_i} to the consequent MF B_i of the rule R_i. Finally, the global RMI index for a FRBS is defined as the minimum value amongst the RMI(R_i) values:

$$\text{RMI} = min_i(\text{RMI}(R_i)), \forall\ 1 \leq i \leq RuleNumber$$

RMI is defined in $[0, 1]$: 0 implies the lowest interpretability and 1 the highest. High values for RMI mean a rule well-defined, because the estimated output using its antecedent has a high level of matching with the rule consequent. Low values for RMI means a rule not well-defined, or even incoherent, in which the estimated output has a low level of matching with the rule consequent.

In the case of classification problems and following the same philosophy, RMI(R_i) is defined as follows, $\text{RMI}(R_i) = \begin{cases} 0, \text{ if } O_{R_i} \neq C_i \\ 1, \text{ if } O_{R_i} = C_i \end{cases}$, and RMI as follows, $\text{RMI} = \sum_{i=1}^{RuleNumber} \text{RMI}(R_i)/RuleNumber$.

Since this contribution is mainly focused on regression problems, from now on we will consider the first RMI definition. In the following, the steps to calculate RMI are described in detail. To summarize, for each rule R_i, its RMI value is as follows:

Definition of the input: based on the n antecedents of the rule R_i.

↓

Estimation of the output $\mathbf{O_{R_i}}$: in order to take into account the inference of the system.

↓

Calculation of RMI: worst or average case depending on the problem type.

1 Definition of the input. In order to calculate the RMI value, the FRBS input must be defined as an n-dimensional fuzzy set by the cores of the MFs in the n antecedents of R_i. In Figure 2, some cases for different MF shapes are shown. Triangular or gaussian MF: the core of the function is the central point c. Trapezoidal MF: the core of the function is a new fuzzy set defined between the points c^l and c^r.

This input is itself a type of fuzzy set: singleton, rectangular, etc. The fuzzy set defined by the core of antecedent k of rule i, A_{ik}, is denoted by $Core(A_{ik})$.

2 Estimation of the output $\mathbf{O_{R_i}}$, considering the input generated in the previous step.

- In the case of a singleton fuzzy set as input, the output is estimated as usual by considering each value as a crisp value.

- In the case of a fuzzy set as input, first the degree of activation for each antecedent k is computed, then the output is estimated.
  The degree of activation is computed as the height of the fuzzy set resulting from the intersection between: the input defined by the core of the antecedent k in rule i ($Core(A_{ik})$), and the MF of antecedent k in rule j (A_{jk}) (see Figure 4). The associated equation is:

$$\begin{aligned} &ActivationDegree(Core(A_{ik}), A_{jk}) = \\ &= Height(Core(A_{ik}) \wedge A_{jk}) = \\ &= \max\{h | h = \mu_{Core(A_{ik}) \wedge A_{jk}}\} = \\ &= \max\{h | h = (\min\{\mu_{Core(A_{ik})}, \mu_{A_{jk}}\})\} \end{aligned}$$

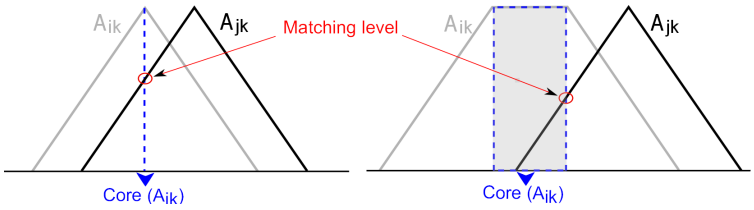


Figure 4: Matching level between $Core(A_{ik})$ and A_{jk} in two example cases, triangular-triangular and trapezoidal-triangular

  Once the degree of activation for each antecedent is computed as a crisp value, then the inference process to estimate the output, O_{R_i}, is as usual. In this second step, the Rmi metric takes into account the FRBS inference, so that if the inference changes, the value of the metric changes depending on the computed output.

3 **Calculation of the RMI value** as the matching between the estimated output from step 2 and the consequent MF of R_i (regression and control): to measure how different the system and the local R_i outputs are.

Figure 5 shows an example to calculate Rmi for rules R_1 and R_3, considering a classic FRBS inference, min-max and defuzzification by center of gravity. Lines in black are the MFs of a rule and gray lines are the MFs from the rule whose value RMI is calculated. As previously discussed, first the FRBS input is defined by the cores of these MFs (shaded areas and arrows in the antecedents), then the output is estimated through the activation of the MFs in the consequents (wavy areas) and finally the RMI value is calculated as the matching between the estimated output and the consequent MF of the rule.

The previous definition of Rmi can be generalized (extended): using α-cuts of the antecedent MFs to implement different levels of intensities for the inputs of the FRBS. The universe of discourse of the core fuzzy set of intensity α used as input is defined as follows:

$$\mu_{Core^{\alpha}(A_{ik})} = \begin{cases} \mu_{A_{ik}} & \text{, if } \mu_{A_{ik}} \geq \alpha \\ 0 & \text{, if } \mu_{A_{ik}} < \alpha \end{cases}$$

The value of α must be greater than 0.5 for normalized systems, lower values are meaningless. Figure 6 shows some examples when the intensity is 0.8.

In this study, we use triangular MFs with intensity 1.0 to measure the Rmi index for regression problems.

3.3. Assessing the Local Relative Interpretability of FRBSs: Linguistic and Pseudo-linguistic Models

These indexes of semantic-based interpretability, extended Gm3m and Rmi, can be used as global measures for FRBS optimization or learning. However, the individual metrics calculated for each MF or rule in order to compute them can also be used to locally represent and compare different FRBSs. First of all, Gm3m can quantify the semantic interpretability for each MF. In this way, a semantic meaning, defined by

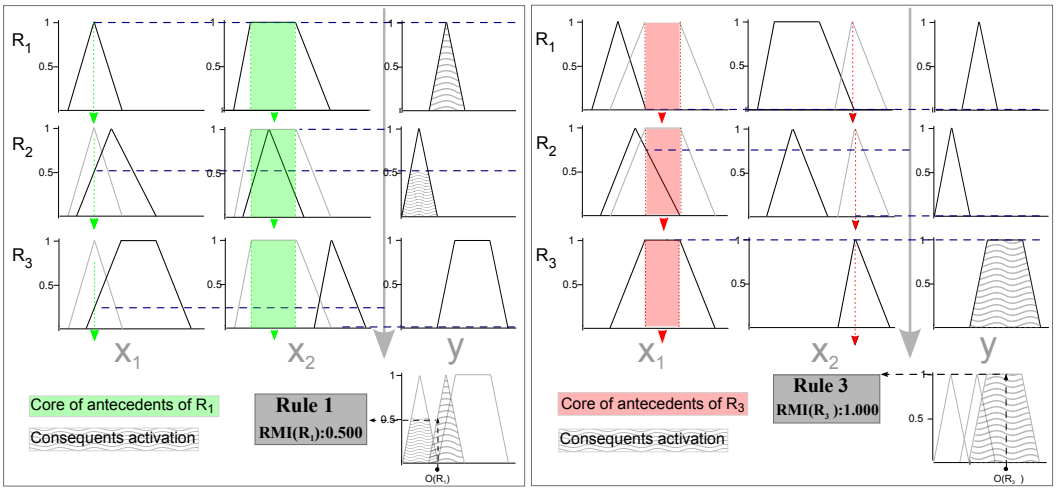


Figure 5: RMI example cases at R_1 and R_3 in a FRBS with three rules

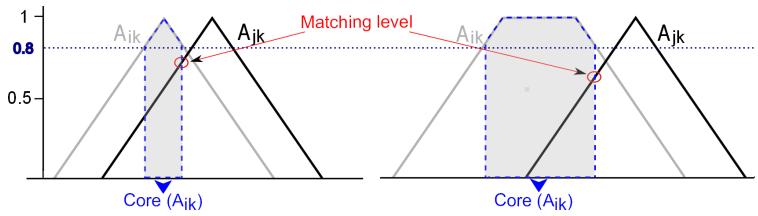


Figure 6: Matching level between $Core(A_{ik})$ and A_{jk} with $\alpha = 0.8$

an original fuzzy partition, can be associated with a given new fuzzy partition, linguistic or scatter, with a given degree. Considering example 1 in Figure 3, it can be said that the semantics of the scatter MF "M'" is equivalent to the semantics of the linguistic MF "M" with degree of 0.545. Then a fuzzy rule antecedent, or a consequent, can be described in two ways:

1. Using the notation of a scatter MF: x_1 is M'
2. Associating the MF to the equivalent linguistic MF, whose semantics have been previously defined by considering the local GM3M degree: x_1 is $M^{0.545}$. Alternatively, using the graphical representation shown in Figure 7, the MFs could be represented together with their degrees in order to make its local interpretation easier. We call the scatter-based models represented in this way as pseudo-linguistic models.

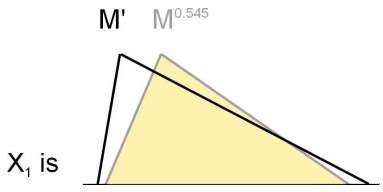


Figure 7: Graphical representation for pseudo-linguistic models

On the other hand, RMI(R_i) can quantify the semantic interpretability, or reliability, of each individual fuzzy rule. Taking the example of Figure 5: R_3 has the highest degree of reliability (RMI= 1) because the output of the FRBS in its activation region is consistent with it. Meanwhile, the degree for R_1 is 0.5 because

there is another very similar rule (R_2) whose consequent is a little different. This implies some level of incoherence (R_1 does not perfectly explain the real system behavior).

By using these quantitative measures, different types of FRBSs can be compared and the user can select the most convenient Knowledge Base (KB) in each case: pseudo-linguistic or linguistic.

4. Multi-Objective Evolutionary Tuning and Rule Selection for Finding the Different Trade-Offs on Linguistic and Scatter FRBSs

The proposed algorithm performs a fuzzy rule selection and a MFs tuning to improve the system's accuracy, reducing the model's complexity and preserving the semantic interpretability at the MF and the RB level. Here, two new versions of the MOEA called *SPEA*2 for Semantic Interpretability [28] have been implemented for application on scatter and linguistic models, $L-TS_{SP2-SEM}$ and $S-TS_{SP2-SEM}$ respectively. Since both versions are based on common specifically designed components, we will describe them as a single algorithm. A particular description is provided for the parts that are different in each version. In the following, the common and specific components are presented.

4.1. Objectives

Each chromosome is associated with a four-dimensional vector containing the fulfilment degree for each one of the following objectives:

1. Extended GM3M maximization: to preserve or improve the MF semantic interpretability.
2. Rule Meaning Index (RMI) maximization: to preserve or improve the rule semantic interpretability.
3. Number of Rules (NR) minimization: to reduce the model complexity.
4. Mean Squared Error (*MSE*) minimization: to reduce the system error defined as $MSE = \frac{1}{|D|}\sum_{l=1}^{|D|}(F(x^l) - y^l)^2$, where $|D|$ is the dataset size, $F(x^l)$ is the output of the FRBS when the l-th example is an input, and y^l is the known desired output.

If one solution does not cover some examples, then the fitness objectives are penalized so that the solution is dominated by others.

4.2. Coding Scheme and Initial Gene Pool

A double coding scheme for *rule selection* (C_S) and *tuning* (C_T) is used: $C^p = C_S^p C_T^p$. The coding scheme for $C_S^p = (c_{S1}, \ldots, c_{Sm})$ consists of binary-coded strings with size m (number of initial rules). Depending on whether a rule is selected or not, values '1' or '0' are respectively assigned to the corresponding gene. Real coding is used for C_T, but taking into account the two different FRBS types studied in this work (linguistic and scatter), there are two different coding schemes for the C_T part. The following subsections show these coding schemes. Anyway, the maximum number of rules in the encoding is determined from the initial RBs since no feature selection is performed.

4.2.1. C_T Coding for Linguistic-based Modeling

The real coding scheme for linguistic-based modeling (L-$TS_{SP2-SEM}$) has the following form, m^i being the number of labels of each of the n variables in the DB.

$$C_T^p = C_1 C_2 \ldots C_n;$$
$$C_i = (a_1^i, b_1^i, c_1^i, \ldots, a_{m^i}^i, b_{m^i}^i, c_{m^i}^i), \quad i = 1, \ldots, n \ .$$

4.2.2. C_T Coding for Scatter-based Modeling

The real coding scheme for scatter-based modeling (S-$TS_{SP2-SEM}$) is very similar to the previous one, but now each MF of each variable in each rule is encoded. If the FRBS has m initial rules with n system variables ($n-1$ input variables and 1 output variable), then the coding scheme is:

$$C_T^p = C_1 C_2 \ldots C_m;$$
$$C_i = (a_1^i, b_1^i, c_1^i, \ldots, a_n^i, b_n^i, c_n^i), \quad i = 1, \ldots, m \ .$$

4.2.3. Initial Population

The initial population have all individuals with value '1' in their genes for C_S. The MFs of the initial model (linguistic or scatter-based) are included as the C_T part of the first individual. Further, the linguistic DB, given as original reference in order to compute the GM3M index, is also included as the C_T part of the second individual. The rest of individuals are generated at random according to the variation intervals defined in section 3.1.2.

On the other hand, on some very particular occasions the initial scatter models are not able to cover a small number of the training data. These models are repaired by adding rules to the RB using the given reference linguistic DB and the Wang & Mendel algorithm [56] on these uncovered examples. Of course, these rules are also considered for rule selection and scatter-based tunning.

4.3. Crossover and Mutation

The intelligent crossover and mutation operators used are based on previous experience of rule selection and tuning ([27]). The steps to obtain each offspring are as follows:

- BLX-0.5 [23] crossover is applied to obtain the C_T part of the offspring.
- Once the offspring C_T part has been obtained, the binary part C_S is obtained based on the C_T parts (MFs) of parents and offspring. For each gene in the C_S part which represents a concrete rule:
  1. The MFs involved in such a rule are extracted from the corresponding C_T parts for each individual involved in the crossover (offspring and parents 1 and 2). Thus, we can obtain the specific rules that each of the three individuals are representing.
  2. Euclidean normalized distances are computed between the offspring rule and each parent rule by considering the center points (vertex) of the MFs comprising such rules. The differences between each pair of centers are normalized by the amplitudes of their respective variation interval.
  3. The parent with the closer rule to the one obtained by the offspring is the one that determines if this rule is selected or not for the offspring by directly copying its value in C_S for the corresponding gene.

This process is repeated until all the C_S values are assigned for the offspring. Four offspring are obtained repeating this process four times (after considering mutation, only the two most accurate are taken as descendants). Once an offspring is generated, the mutation operator changes a gene value at random in the C_T part (making its value equal to initial DB given as reference) and directly sets to zero a gene selected at random in the C_S part (one gene is modified in each part) with probability P_m. Applying these operators, two problems are solved: First, crossing individuals with very different rule configurations is more productive. Second, this favors rule extraction since mutation is only engaged to remove unnecessary rules.

4.4. Specific Mechanisms for Handling the particular Trade-off between Accuracy and Interpretability

The proposed algorithm uses the *SPEA*2 selection mechanism [59]. However, in order to improve the search algorithm ability, the following changes are considered:

- A mechanism for incest prevention based on the concepts of CHC [22] to avoid premature convergence in C_T. $L - TS_{SP2-SEM}$ for linguistic-based modeling uses a mechanism implemented according to [28]: only those parents whose Hamming distance divided by 4 is higher than a threshold are crossed. Since we consider a real coding scheme (only C_T parts are considered), we have to transform each gene using a Gray Code with a fixed number of bits per gene (*BGene*) determined by the system expert. In this way, the threshold value is initialized as $L = (\#C_T * BGene)/4$, where $\#C_T$ is the number of genes in the C_T part of the chromosome. At each generation of the algorithm, the threshold value is decremented by one, which allows closer solutions to be crossed. $S - TS_{SP2-SEM}$ for the scatter-based modeling has two minimal differences: the threshold value is initialized as $L = (maxD/4) + 1$, where $maxD$ is the maximum Hamming distance in C_T for external population; and for each algorithm

generation, the decrement in the threshold value is proportional to the number of variables. In this way the algorithm avoids many iterations until the first cross.

- The restarting operator forces the external population to be empty, generating a new initial population. Before removing them from the external population, this new initial population includes a copy of the best individual with the best accuracy value, and the two best individuals with the best different values for the rest of the objectives. The remaining individuals take the values of the most accurate individual for C_S, and values generated at random for C_T. Restarting is applied once the most accurate solution improves from the previous restart and if 50 percent of crossovers are detected at any generation (the required ratio can be defined as $\%_{Required} = 0.5$). This condition is updated each time restarting is performed as $\%_{Required} = (1+\%_{Required})/2$. The restarting is not applied at the end and it is disabled if it has never been applied before reaching the mid-point of the total number of evaluations. After the restarting operator, the corresponding variation intervals (for GM3M index) are recalculated based on the most accurate solution.

- During environmental selection, when the size of the current nondominated set exceeds the size of the external population ($\overline{N}$), the SPEA2 truncation procedure is modified according to: The $\overline{N}/2$ most accurate solutions are marked as non-removable, since this is the most difficult objective. Also, the best and the second best of the remaining objectives are included. If there are solutions that have the same value for these objectives, only the most accurate is picked up. Thus, the second best solution represents a different value in the objective.

- At each stage of the algorithm (between restarting points), the number of solutions in the external population ($\overline{P}_{t+1}$) that is considered to form the mating pool is progressively reduced by focusing only on those with the best accuracy. To do this, the solutions are sorted from the best to the worst (considering accuracy as the criterion), and the number of solutions that are considered for selection is reduced progressively from 100% at the beginning to 50% at the end of each stage. This is done by taking into account the value of L. In the last evaluations when restart is disabled, this mechanism for focusing on the most accurate solutions is also modified to focus on the best individuals alternatively for each objective.

- For scatter-based modeling, $S - TS_{SP2-SEM}$, the MFs associated to the rules selected in C_S are the only ones considered to compute the GM3M index and the Hamming distance in the mechanism for incest prevention.

5. Experimental Study

To evaluate the usefulness of the proposed approach, we have used nine real-world problems from the KEEL dataset repository [4] (*http://www.keel.es/*). Table 2 summarizes the main characteristics of these datasets. To ease the analysis and application of the statistical test, only three representative points of the Pareto front (from the most accurate to the most interpretable) have been considered on each plane (accuracy-complexity, accuracy-MF semantic and accuracy-rule semantic). This section is organized as follows:

1. Subsection 5.1 presents the experimental set-up.
2. Subsection 5.2 analyzes the results on the most accurate solution.
3. Subsection 5.3 presents an analysis of the median solutions in the different objective planes.
4. Subsection 5.4 analyzes the results on the most interpretable solutions in the different objective planes.
5. Subsection 5.5 includes a global analysis where the most relevant conclusions are set out.
6. Finally, Subsection 5.6 shows how to use the indexes to do a local analysis on some particular KBs obtained in the experiments.

Table 2: Data sets considered for the experimental study

Datasets	Name	Variables	Patterns	Datasets	Name	Variables	Patterns
Plastic Strength	PLA	3	1650	Weather Ankara	WAN	10	1609
Quake	QUA	4	2178	Weather Izmir	WIZ	10	1461
Electrical Maintenance	ELE	5	1056	Mortgage	MOR	16	1049
Abalone	ABA	9	4177	Treasury	TRE	16	1049
Stock prices	STP	10	950				

Available at `http://sci2s.ugr.es/keel/datasets.php`

5.1. Experimental Set-up

Several fuzzy models are generated based on different algorithms for each dataset considered. In order to evaluate the accuracy-interpretability trade-off of linguistic and scatter based approaches, two different algorithms are used to obtain an initial set of candidate rules for each fuzzy modeling approach. The linguistic algorithms used in this work are the Linguistic Iterative Rule Learning (L-IRL) [17] and Neuro-Fuzzy Function Approximation (NEFPROX) [47]. The scatter algorithms are the Scatter Iterative Rule Learning (S-IRL) [18] and a Neuro-Fuzzy System based on the Adaptative Resonance Theory (FASART) [12]. Each algorithm has its own fuzzy inference system and parameters:

L-IRL and S-IRL *Center of gravity weighted by the matching* strategy as a defuzzification operator and the *minimum t-norm* as implication and conjunctive operators. The used parameters are: nLT (the number of linguistic terms for initial linguistic partitions), ϵ (minimum covering degree), ω (covering for positive examples), K (percentage of negative examples), P (population size), Gen (for the number of generations), a and b (crossover and mutation), P_c (crossover probability) and P_m (mutation probability). Further, in the case of S-IRL, the evolutionary strategy (ES) is applied until there is no improvement in 50 generations over a percentage $\alpha = 20\%$ of the individuals of the population.

NEFPROX *max-min* inference and defuzzification by *mean of maximum*. The only parameter for this model is: nLT (the number of linguistic terms for initial linguistic partitions).

FASART Fuzzification by *single point*, Inference by *product*, and Defuzzification by *average of fuzzy set centers*. The parameters for this model are: ρ (the vigilance parameter) and γ (the fuzzification rate).

The specific parameters used to generate the initial KBs are shown in table 3 and depend on the number of variables in each dataset. Two cases are distinguished: datasets with a number of variables smaller than 9, and the remaining ones. Setting this criterion, the initial KBs obtained show a reasonable number of rules for the more complex datasets. Once the initial KBs are generated, the different post-processing algorithms are applied. According to the specifications explained in Section 4, there are two versions of the proposed MOEA: $L-TS_{SP2-SEM}$ for linguistic-based modeling, which will be applied on KBs generated by linguistic-based algorithms; and $S-TS_{SP2-SEM}$ for scatter-based modeling, which will be applied on KBs generated by scatter-based algorithms. The parameters to run the algorithms are: population size of 200, external population size of 61, 100000 evaluations, 0.2 as mutation probability, and 30 bits per gene for the Gray codification. The application of both versions on the initial KB generation algorithms gives way to four possible combinations. Table 4 summarizes these four cases considered for the experiments: LING1, LING2, SCAT1 and SCAT2.

Table 3: Parameters for the initial KBs

#var	L-IRL and S-IRL	NEFPROX	FASART
< 9	$nLT = 5, \epsilon = 1.5, \omega = 0.05, K = 0.1, P = 61, Gen = 100$	$nLT = 5$	$\rho = 0.7, \gamma = 8$
	$a = 0.35, b = 5, P_c = 0.6, P_m = 0.1, ES = 50, \alpha = 30\%$		
≥ 9	$nLT = 3, \epsilon = 1.5, \omega = 0.05, K = 0.1, P = 61, Gen = 100$	$nLT = 3$	$\rho = 0.7, \gamma = 6$
	$a = 0.35, b = 5, P_c = 0.6, P_m = 0.1, ES = 50, \alpha = 30\%$		

Table 4: Methods considered for analyzing the interpretability-accuracy trade-off

Method	Ref.	Description	Objectives
		Methods for generating the initial KBs	
L-IRL	[17]	Linguistic Iterative Rule Learning (Initial Linguistic KB Generation)	—
NEFPROX	[47]	Neuro-Fuzzy Function Approximation (Initial Linguistic KB Generation)	—
S-IRL	[18]	Scatter hard constrained Iterative Rule Learning (Initial Scatter KB Generation)	—
FASART	[12]	Neuro fuzzy system based on ART (Initial Scatter KB Generation)	—
		Multi-Objective Evolutionary Algorithms for Post-processing	
L-TS$_{SP2-SEM}$	New	Linguistic Tuning and Rule Selection with semantic by SPEA2	*MSE* / NR / Gm3m / Rmi
S-TS$_{SP2-SEM}$	New	Scatter Tuning and Rule Selection with semantic by SPEA2	*MSE* / NR / Gm3m / Rmi
		Studied Combinations (Two Linguistic and Two Scatter)	
Ling1	—	L-IRL + L-TS$_{SP2-SEM}$	*MSE* / NR / Gm3m / Rmi
Ling2	—	NEFPROX + L-TS$_{SP2-SEM}$	*MSE* / NR / Gm3m / Rmi
Scat1	—	S-IRL + S-TS$_{SP2-SEM}$	*MSE* / NR / Gm3m / Rmi
Scat2	—	FASART + S-TS$_{SP2-SEM}$	*MSE* / NR / Gm3m / Rmi

Since these four multiobjective approaches use four objectives, we project the solutions obtained in three planes, accuracy-complexity, accuracy-Gm3m and accuracy-Rmi, subsequently removing the dominated solutions appearing from these projections. In this way, we can better analyze the existent relations between each interpretability and accuracy objective. Some researchers have also used these kinds of projections for graphical representation and statistical analysis when three objectives are optimized together [28, 9].

In the experiments, we adopted a *5-fold cross-validation model*, i.e., we randomly split the data set into 5 folds, each containing 20% of the patterns of the data set, and used four folds for training and one for testing [1]. The algorithm was applied six times, considering a different random seed for each of the possible five different partitions (training/test). Therefore, we consider the average results of 30 runs.

The approximated Pareto front is generated in the corresponding objective planes for every dataset and trial. Then, we focus on three representative points: the most interpretable one, the median one and the most accurate one for training. We compute the mean values over the 30 trials for the MSE on the training and test sets ($MSE_{tra/tst}$), the NR, the Gm3m and/or the Rmi index, depending on the objective planes involved for each of these representative points. These three points are representative positions on each plane, so they are considered to perform a statistical analysis. Anyway, the final user could select the most appropriate solution from the final Pareto front: looking for a trade-off between NR, Gm3m and Rmi depending on his/her own preferences.

In order to assess whether significant differences exist among the results, we adopt statistical analysis [21, 31, 30] based on non-parametric tests, according to the recommendations made in [21] and [31], where a set of simple, safe and robust non-parametric tests for statistical comparisons of classifiers has been introduced. In particular, we will use non-parametric tests for multiple comparison: Friedman's test [25], Iman and Davenport's test [37] and Holm's method [36]. In the tests the null hypothesis means the similarity regarding to the linguistic and scatter improved FRBSs. A detailed description of these tests can be seen in `http://sci2s. ugr.es/sicidm/`. To perform the tests, we use a level of confidence $\alpha = 0.1$.

The averaged results of the initial FRBSs obtained by Ling1, Ling2, Scat1 and Scat2 in the five folds are shown in Table 5, mean squared error for training and test ($MSE_{tra/tst}$), the number of rules (NR), the MF semantic interpretability index (Gm3m) and the rule semantic interpretability index (Rmi).

[1]The corresponding data partitions (5-fold) for these data sets are available at the KEEL project webpage [4]: `http://sci2s.ugr.es/keel/datasets.php`

Table 5: Initial results obtained by LING1, LING2, SCAT1 and SCAT2

Dataset	LING1				LING2				SCAT1				SCAT2			
	$MSE_{tra/tst}$	NR	GM3M	RMI	$MSE_{tra/tst}$	NR	GM3M	RMI	$MSE_{tra/tst}$	NR	GM3M	RMI	$MSE_{tra/tst}$	NR	GM3M	RMI
PLA	5.25/5.26	75.4	1.00	0.00	3.40/3.38	14.8	0.77	0.83	4.05/4.16	91.4	0.00	0.00	3.79/3.82	46.8	0.46	0.00
QUA	0.06/0.06	227.6	1.00	0.00	0.04/0.04	53.6	0.73	0.96	259/260	71.4	0.00	0.00	0.05/0.05	107.4	0.42	0.00
ELE	129401/133565	88.8	1.00	0.25	407398/410115	65.0	0.97	0.99	239184/242827	38.6	0.06	0.46	109178/153749	81.8	0.39	0.64
ABA	24.79/24.72	50.2	1.00	0.25	18.63/18.59	72.0	0.77	0.86	20.66/20.70	21.5	0.02	0.62	8.13/8.59	45.6	0.41	0.11
STP	16.75/16.91	45.6	1.00	0.70	10.47/10.62	123.2	0.64	0.86	20.20/20.15	16.1	0.02	0.63	2.07/2.19	36.2	0.39	0.87
WIZ	38.41/39.48	52.4	1.00	0.60	14.43/15.13	105.4	0.77	0.90	87.03/88.25	18.7	0.00	0.58	7.02/9.97	83.4	0.38	0.72
WAN	52.74/53.53	45.6	1.00	1.00	26.17/26.96	157.4	0.79	0.78	88.31/90.32	19.8	0.00	0.56	8.96/11.65	93.6	0.35	0.75
MOR	2.00/2.00	31.4	1.00	1.00	1.99/2.01	78.2	0.90	0.97	25.17/24.84	15.5	0.00	0.58	0.45/0.50	22.6	0.41	0.83
TRE	2.67/2.68	33.0	1.00	0.53	3.68/3.72	74.4	0.54	0.82	20.53/21.38	15.9	0.01	0.64	0.82/0.86	25.0	0.40	0.82

5.2. Results and Analysis of the Most Accurate Solution

The results obtained by the studied algorithms on the most accurate solutions are shown in Table 6. This table is grouped in columns by algorithms, and shows the averaged results obtained by each algorithm for all the studied datasets. See Section 5.1 for a description of this type of table. The penultimate row shows the average for all the datasets and the last one shows the number of times that the final model improves the initial model.

Table 6: Average results of the studied algorithms on the most accurate models

Dataset	LING1				LING2				SCAT1				SCAT2			
	$MSE_{tra/tst}$	NR	GM3M	RMI	$MSE_{tra/tst}$	NR	GM3M	RMI	$MSE_{tra/tst}$	NR	GM3M	RMI	$MSE_{tra/tst}$	NR	GM3M	RMI
PLA	2.24/2.38	32.0	0.71	0.04	2.39/2.52	13.3	0.47	0.09	2.23/2.45	27.7	0.39	0.24	2.11/2.27	21.6	0.52	0.35
QUA	0.03/0.04	96.5	0.47	0.00	0.03/0.04	26.2	0.41	0.77	0.04/0.04	21.7	0.43	0.34	0.03/0.04	65.0	0.45	0.00
ELE	28247/37644	32.5	0.54	0.54	23140/30249	24.4	0.56	0.73	16079/23613	27.3	0.47	0.76	15907/19419	58.6	0.41	0.87
ABA	5.42/5.56	13.0	0.47	0.66	5.20/5.46	17.5	0.46	0.89	4.81/5.07	13.6	0.48	0.86	4.49/4.83	25.9	0.44	0.93
STP	2.98/3.28	14.1	0.47	0.76	1.44/1.94	25.8	0.48	0.88	1.07/1.46	12.6	0.51	0.91	0.63/0.87	26.7	0.39	0.99
WIZ	2.51/3.05	13.0	0.61	0.93	2.48/3.26	29.8	0.53	0.83	1.81/3.13	12.0	0.51	0.93	1.07/2.02	52.9	0.34	0.96
WAN	3.86/5.65	9.5	0.57	0.91	4.27/5.29	29.2	0.53	0.89	2.47/4.68	12.3	0.53	0.92	1.38/2.54	62.5	0.33	0.95
MOR	0.07/0.09	9.0	0.60	0.97	0.03/0.05	15.5	0.61	0.93	0.06/0.08	10.3	0.52	0.94	0.05/0.06	16.8	0.43	1.00
TRE	0.10/0.11	9.0	0.63	0.98	0.06/0.08	15.4	0.64	0.99	0.08/0.10	10.5	0.55	0.95	0.07/0.09	17.4	0.43	1.00
Av.	-	25.40	0.56	0.64	-	21.89	0.52	0.78	-	16.45	0.49	0.76	-	38.61	0.41	0.78
Win initial	9/9	9	0	7	9/9	9	1	4	9/9	9	9	9	9/9	9	6	9

Table 7 shows the rankings (through Friedman's test) of the different algorithms considered for the four measures (MST_{tst}, NR, GM3M and RMI). Iman-Davenport's tells us that significant differences exist among the results observed for all datasets, with p-values (0.0246), (5.081E-3), (6.132E-3) and (5.081E-3) on MSE_{tst}, NR, GM3M and RMI respectively. The results of Holm's test are shown in table 8 for every measure.

Table 7: Rankings obtained through Friedman's Test on the different measures for the most accurate models

Algorithm	MSE_{tst}	NR	GM3M	RMI
LING1	3.3333	2.1111	1.6667	2.4444
LING2	2.5556	2.4444	2.1111	2.1111
SCAT1	2.5556	1.7778	2.6667	3.6667
SCAT2	1.5556	3.6667	3.5556	1.7778

According to the results and non-parametric statistical tests, some comments can be given by focusing on the most accurate solutions:

- The error and rule number are significantly reduced for all the cases in comparison with the initial models. Also, the semantic interpretability of rules and MFs is improved for the scatter approaches,

Table 8: Holm's post-hoc test with $\alpha = 0.1$ on the different measures for the most accurate models

Holm on MSE_{tst}						Holm on NR						Holm on Gm3m						Holm on Rmi					
i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.
3	Ling1	2.92	0.003	0.03	Rej.	3	Scat2	3.10	0.002	0.03	Rej.	3	Scat2	3.10	0.002	0.03	Rej.	3	Scat1	3.10	0.002	0.03	Rej.
2	Scat1	1.64	0.100	0.05	Acc.	2	Ling2	1.10	0.273	0.05	Acc.	2	Scat1	1.64	0.100	0.05	Acc.	2	Ling1	1.10	0.273	0.05	Acc.
1	Ling2	1.64	0.100	0.10	Acc.	1	Ling1	0.55	0.584	0.10	Acc.	1	Ling2	0.73	0.465	0.10	Acc.	1	Ling2	0.55	0.584	0.10	Acc.

although this is not applicable to the linguistic ones, which improve the accuracy at the cost of part of the initial semantics.

- The best ranking in Friedman's test is obtained for the scatter algorithms, except for the Gm3m measure, where Ling1 is the best ranked algorithm.
- Holm's test accepts the hypothesis of equality between the three first algorithms for all measures. In terms of complexity and semantic interpretability (NR, Gm3m and Rmi), one scatter algorithm and the two linguistic ones should be considered equivalent. In terms of accuracy (MSE_{tst}), the two scatter algorithms and one linguistic (Ling2) one are equivalent too.
- We observe that Scat2 has the most accurate solutions with high rule semantic interpretability. On the other hand, Ling2 is equivalent to Scat2 according to the statistical tests on error and Rmi, but Ling2 has better semantic interpretability on NR and Gm3m measures. Otherwise, Ling2 is equivalent to Scat1 on NR and Gm3m, but Scat1 introduces overfitting for some datasets.

In accordance with these results, there is no general rule to decide the most adequate algorithm, scatter or linguistic, to achieve the best accuracy. In fact, in some cases, linguistic approaches can achieve similar levels of accuracy to the scatter algorithms.

5.3. Results and Analysis of the Median Solutions on the Different Objective Planes

This section analyzes the results of the median solutions on the three planes for all the proposed algorithms. Table 9 shows the results obtained on the different measures for each plane as shown in Table 6 of the previous section (see Section 5.2 for a detailed description). Table 10 shows the rankings (through Friedman's test) in each plane for the different algorithms considered on the four measures (MST_{tst}, NR, Gm3m and Rmi). The Iman-Davenport's test p-values in the NR plane (1.081E-4, 0.0213, 0.0126 and 0.1175 on MSE_{tst}, NR, Gm3m and Rmi respectively) imply that there are statistical differences among the results, except for Rmi. The p-values in the Gm3m plane (0.0712, 0.0394, 8.214E-5 and 0.6321 on MSE_{tst}, NR, Gm3m and Rmi respectively) imply that there are statistical differences among the results, except for Rmi. And the p-values in the Rmi plane (0.4551, 5.081E-3, 5.081E-3 and 6.249E-4 on MSE_{tst}, NR, Gm3m and Rmi respectively) imply that there are statistical differences among the results, except for MSE_{tst}. The results of Holm's test are shown in table 11.

According to the results and the non-parametric statistical tests, some conclusions can be achieved:

- The improvements achieved in comparison with the initial models are similar to those obtained for the most accurate solutions for every plane: the training and test error, and the number of rules are reduced for all the cases, except for some cases of Scat2. Also, the rule and MF semantic interpretability is improved for the scatter approaches, although this is again not applicable for the linguistic ones.
- The p-values of Iman-Davenport's test show that, in some planes, the differences between algorithms are not statistically significant (MSE_{tst} in Rmi plane and Rmi in NR and Gm3m planes). For the remaining measures and planes, the best ranking for the Friedman's test is obtained by different algorithms.

Table 9: Average results of the studied algorithms on the median models

Results on the Median point for Accuracy/NR plane																
Dataset	Ling1				Ling2				Scat1				Scat2			
	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi
PLA	2.46/2.56	21.5	0.64	0.21	2.82/2.95	10.1	0.53	0.57	2.70/2.89	21.2	0.38	0.36	2.89/2.98	12.7	0.57	0.35
QUA	0.03/0.04	87.9	0.51	0.00	0.03/0.04	22.8	0.44	0.69	0.04/0.04	20.3	0.46	0.36	0.03/0.04	55.2	0.49	0.00
ELE	51770/62262	24.5	0.54	0.56	40586/49636	20.3	0.49	0.73	58102/69947	22.5	0.49	0.69	84275/90980	48.7	0.37	0.67
ABA	5.87/5.99	9.5	0.49	0.67	5.69/5.88	11.5	0.46	0.93	6.13/6.31	9.5	0.50	0.81	6.28/6.46	19.3	0.49	0.72
STP	3.85/4.07	10.9	0.45	0.81	2.33/2.80	19.6	0.51	0.86	2.83/3.26	9.1	0.52	0.90	2.38/2.56	19.0	0.40	0.95
WIZ	3.54/3.90	8.9	0.59	0.90	3.35/4.01	21.7	0.46	0.87	3.73/5.16	8.8	0.52	0.92	6.30/7.03	29.8	0.34	0.84
WAN	4.97/6.54	7.0	0.57	0.89	5.13/6.07	23.8	0.47	0.89	6.14/7.82	8.9	0.51	0.90	10.16/11.08	32.1	0.40	0.83
MOR	0.22/0.24	6.4	0.55	0.95	0.10/0.13	10.5	0.60	0.93	0.23/0.26	7.5	0.53	0.93	0.68/0.72	10.2	0.47	0.96
TRE	0.23/0.26	6.0	0.56	0.97	0.11/0.13	10.9	0.61	0.97	0.29/0.30	7.5	0.52	0.92	0.94/0.99	11.3	0.40	0.98
Av.	-	20.27	0.55	0.66	-	16.79	0.51	0.83	-	12.80	0.49	0.75	-	26.49	0.44	0.70
Win initial	9/9	9	0	7	9/9	9	1	3	9/9	9	9	9	5/6	9	6	9
Results on the Median point for Accuracy/Gm3m plane																
Dataset	Ling1				Ling2				Scat1				Scat2			
	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi
PLA	2.26/2.39	30.8	0.75	0.05	2.46/2.58	13.0	0.53	0.30	2.40/2.61	26.9	0.44	0.36	2.15/2.31	21.3	0.55	0.33
QUA	0.03/0.04	95.5	0.57	0.00	0.04/0.04	25.9	0.57	0.71	0.00/0.04	23.1	0.53	0.29	0.03/0.04	65.0	0.49	0.00
ELE	29165/38261	32.3	0.61	0.54	36944/45671	25.5	0.63	0.71	29680/40013	27.0	0.55	0.75	42268/46000	55.6	0.47	0.76
ABA	5.57/5.71	12.6	0.57	0.68	5.55/5.76	16.5	0.53	0.91	6.21/6.38	11.7	0.58	0.85	4.56/4.90	25.8	0.49	0.90
STP	3.40/3.65	13.8	0.57	0.78	1.65/2.10	26.7	0.52	0.86	3.26/3.63	11.1	0.59	0.87	1.15/1.33	25.5	0.45	0.97
WIZ	2.79/3.24	13.0	0.68	0.94	2.83/3.57	32.0	0.60	0.83	4.59/6.05	10.6	0.59	0.91	6.38/7.11	31.6	0.48	0.87
WAN	4.41/5.86	9.4	0.65	0.92	5.20/6.20	33.2	0.60	0.90	4.52/6.30	11.8	0.59	0.91	9.97/10.67	40.3	0.49	0.84
MOR	0.07/0.09	9.0	0.66	0.95	0.06/0.07	16.8	0.67	0.95	0.20/0.22	9.4	0.60	0.93	0.54/0.61	13.2	0.52	0.98
TRE	0.11/0.12	9.0	0.69	0.98	0.08/0.11	17.4	0.68	0.97	0.31/0.35	9.6	0.61	0.92	0.41/0.43	15.7	0.49	0.98
Av.	-	25.06	0.64	0.65	-	23.00	0.59	0.79	-	15.68	0.56	0.75	-	32.67	0.49	0.74
Win initial	9/9	9	0	7	9/9	9	1	4	9/9	9	9	9	7/8	9	9	9
Results on the Median point for Accuracy/Rmi plane																
Dataset	Ling1				Ling2				Scat1				Scat2			
	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi
PLA	2.36/2.49	28.5	0.65	0.27	2.40/2.52	13.3	0.48	0.21	2.31/2.51	27.3	0.39	0.39	2.45/2.55	18.3	0.52	0.53
QUA	0.03/0.04	96.5	0.47	0.00	0.03/0.04	26.1	0.40	0.86	0.04/0.04	22.6	0.42	0.49	0.03/0.04	61.4	0.47	0.02
ELE	30563/39637	32.0	0.53	0.58	28442/36966	23.7	0.52	0.82	28150/34894	26.8	0.49	0.82	20759/25373	57.6	0.42	0.89
ABA	5.44/5.58	12.9	0.47	0.72	5.20/5.46	17.5	0.45	0.91	5.20/5.41	13.5	0.49	0.92	4.50/4.84	25.9	0.44	0.94
STP	3.16/3.44	13.8	0.46	0.82	1.45/1.95	25.8	0.47	0.92	1.08/1.48	12.6	0.52	0.95	0.74/0.97	26.2	0.38	0.99
WIZ	2.54/3.11	13.0	0.60	0.96	2.49/3.24	29.7	0.51	0.87	2.16/3.60	11.9	0.51	0.96	2.06/2.98	48.3	0.35	0.97
WAN	3.96/5.50	9.6	0.56	0.95	4.28/5.31	29.2	0.51	0.92	2.60/4.71	12.2	0.53	0.96	9.55/10.59	40.2	0.43	0.97
MOR	0.07/0.09	9.0	0.59	0.99	0.03/0.05	15.5	0.60	0.95	0.06/0.08	10.2	0.52	0.96	0.05/0.06	16.7	0.43	1.00
TRE	0.10/0.11	9.0	0.63	0.99	0.06/0.09	15.4	0.64	0.99	0.14/0.16	10.3	0.56	0.98	0.07/0.09	17.4	0.41	1.00
Av.	-	24.93	0.55	0.70	-	21.81	0.51	0.83	-	16.38	0.49	0.82	-	34.67	0.43	0.81
Win initial	9/9	9	0	7	9/9	9	1	4	9/9	9	9	9	8/9	9	7	9

Table 10: Rankings obtained through Friedman's Test on the different measures and planes for the models on the median point

Accuracy/NR plane					Accuracy/Gm3m plane					Accuracy/Rmi plane				
Alg.	MSE_{tst}	NR	Gm3m	Rmi	Alg.	MSE_{tst}	NR	Gm3m	Rmi	Alg.	MSE_{tst}	NR	Gm3m	Rmi
Ling1	2.0000	2.1667	1.5556	3.2778	Ling1	1.8889	2.2222	1.4444	2.8333	Ling1	3.0000	2.1111	1.5556	3.2222
Ling2	1.4444	2.6667	2.5556	1.8889	Ling2	2.0000	2.8889	2.1111	2.6667	Ling2	2.2222	2.4444	2.3333	3.1111
Scat1	2.8889	1.7222	2.4444	2.2222	Scat1	3.1111	1.6667	2.6667	2.4444	Scat1	2.6667	1.7778	2.5556	2.4444
Scat2	3.6667	3.4444	3.4444	2.6111	Scat2	3.0000	3.2222	3.7778	2.0556	Scat2	2.1111	3.6667	3.5556	1.2222

- Holm's test rejects the hypothesis of equality between algorithms in terms of accuracy (MSE_{tst}) for the NR plane, where linguistic algorithms outperform the scatter ones. All algorithms should be considered equivalent for the other planes.
  In terms of rule semantic interpretability (Rmi), the hypothesis of equality is only rejected for the Rmi plane, so that Scat2 outperforms the remaining ones. In terms of complexity (NR), the hypothesis of equality is only rejected in some cases (Scat2 for NR and Rmi planes; Scat2 and Ling2 for Gm3m

plane). In terms of MF semantic interpretability (GM3M), the best ranking interpretability algorithm (LING1) should be considered equivalent to the other linguistic algorithm (LING2) for all planes, and equivalent to one scatter algorithm (SCAT1) in NR and RMI planes.

Table 11: Holm's post-hoc test with $\alpha = 0.1$ on the different measures for the models on the median point

Accuracy/NR plane

Holm on MSE_{tst}						Holm on NR						Holm on GM3M						Holm on RMI					
i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.
3	SCAT2	3.65	0.000	0.03	Rej.	3	SCAT2	2.83	0.005	0.03	Rej.	3	SCAT2	3.10	0.002	0.03	Rej.	3	LING1	2.28	0.022	0.03	Rej.
2	SCAT1	2.37	0.018	0.05	Rej.	2	LING2	1.55	0.121	0.05	Acc.	2	LING2	1.64	0.100	0.05	Acc.	2	SCAT2	1.19	0.235	0.05	Acc.
1	LING1	0.91	0.361	0.10	Acc.	1	LING1	0.73	0.465	0.10	Acc.	1	SCAT1	1.46	0.144	0.10	Acc.	1	SCAT1	0.55	0.584	0.10	Acc.

Accuracy/GM3M plane

Holm on MSE_{tst}						Holm on NR						Holm on GM3M						Holm on RMI					
i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.
3	SCAT1	2.01	0.045	0.03	Acc.	3	SCAT2	2.56	0.011	0.03	Rej.	3	SCAT2	3.83	0.000	0.03	Rej.	3	LING1	1.28	0.201	0.03	Acc.
2	SCAT2	1.83	0.068	0.05	Acc.	2	LING2	2.01	0.045	0.05	Rej.	2	SCAT1	2.01	0.045	0.05	Rej.	2	LING2	1.00	0.315	0.05	Acc.
1	LING2	0.18	0.855	0.10	Acc.	1	LING1	0.91	0.361	0.10	Acc.	1	LING2	1.10	0.273	0.10	Acc.	1	SCAT1	0.64	0.523	0.10	Acc.

Accuracy/RMI plane

Holm on MSE_{tst}						Holm on NR						Holm on GM3M						Holm on RMI					
i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.
3	LING1	1.46	0.144	0.03	Acc.	3	SCAT2	3.10	0.002	0.03	Rej.	3	SCAT2	3.29	0.001	0.03	Rej.	3	LING1	3.29	0.001	0.03	Rej.
2	SCAT1	0.91	0.361	0.05	Acc.	2	LING2	1.10	0.273	0.05	Acc.	2	SCAT1	1.64	0.100	0.05	Acc.	2	LING2	3.10	0.002	0.05	Rej.
1	LING2	0.18	0.855	0.10	Acc.	1	LING1	0.55	0.584	0.10	Acc.	1	LING2	1.28	0.201	0.10	Acc.	1	SCAT1	2.01	0.045	0.10	Rej.

In general terms, again, models with a good accuracy-interpretability trade-off can be generated by any of the algorithms considered in this work, so the final choice is based on the user's preferences.

5.4. Results and Analysis of the Most Interpretable Solutions on the Different Objective Planes

The most interpretable models are shown in Table 12. This table shows the results obtained on the different measures for each plane as shown in Table 6 of the previous section (see Section 5.2 for a detailed description). Table 13 shows the rankings (through Friedman's test) for the different algorithms considered on the four measures (MST_{tst}, NR, GM3M and RMI). The Iman-Davenport's test p-values in the NR plane (1.431E-7, 0.0532, 0.9592 and 0.0459 on MSE_{tst}, NR, GM3M and RMI respectively) imply that there are statistical differences among the results, except for GM3M; in the other planes (6.132E-3, 6.132E-3, 0.0394 and 4.043E-6 on MSE_{tst}, NR, GM3M and RMI respectively in GM3M plane, and 6.249E-4, 0.0394, 0.0246 and 3.450E-5 on MSE_{tst}, NR, GM3M and RMI respectively in RMI plane) the p-values imply that there are statistical differences among the results. The result of Holm's test are shown in table 14.

According to the results and non-parametric statistical tests:

- Solutions achieved by LING1 and SCAT1 algorithms show a similar behavior to the previous ones, improving the initial models in the same way. Besides, the rule semantic interpretability is improved in LING2 while the error is increased in SCAT2, depending on the planes.
- The best ranking for Friedman's test is obtained, in general, by linguistic algorithms, although the scatter algorithms are better for some measures in some planes (NR in GM3M plane; NR and GM3M in RMI plane).
- Holm's test rejects the hypothesis of equality between linguistic and scatter in terms of accuracy (MSE_{tst}) for all the planes. In these cases, linguistic algorithms (LING1, LING2) outperform the scatter ones in general. In terms of complexity (NR), MF semantic interpretability, and rule semantic interpretability (RMI), the results are quite different depending on the different planes and measures, but in general, the linguistic approaches show a higher performance.

Table 12: Average results of the studied algorithms on the most interpretable models

Results on the most interpretable point for Accuracy/NR plane																
Dataset	Ling1				Ling2				Scat1				Scat2			
	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi
PLA	2.95/3.09	15.0	0.70	0.17	3.57/3.72	7.9	0.57	0.74	3.38/3.50	15.3	0.48	0.60	4.78/4.82	8.4	0.69	0.58
QUA	0.04/0.04	76.2	0.61	0.00	0.04/0.04	19.0	0.48	0.81	0.04/0.04	18.9	0.47	0.43	0.04/0.04	43.3	0.61	0.01
ELE	88331/98883	18.0	0.58	0.71	68855/79716	15.4	0.57	0.78	116518/131559	18.4	0.55	0.79	185888/194169	37.8	0.51	0.57
ABA	6.77/6.85	6.2	0.55	0.86	7.20/7.32	7.1	0.53	0.98	10.25/10.27	6.3	0.59	0.85	9.44/9.47	13.4	0.53	0.69
STP	5.25/5.50	7.3	0.52	0.83	4.18/4.61	13.3	0.57	0.97	9.56/9.90	6.3	0.62	0.90	5.68/5.86	12.4	0.52	0.96
WIZ	5.38/6.03	4.9	0.61	0.95	6.36/6.91	12.8	0.56	0.93	9.95/11.34	6.0	0.60	0.93	18.44/19.66	11.7	0.67	0.98
WAN	6.58/7.93	4.6	0.58	0.92	8.37/9.32	16.0	0.54	0.96	17.91/19.53	6.2	0.61	0.92	32.03/33.07	15.0	0.65	0.96
MOR	0.58/0.60	4.4	0.58	0.97	0.46/0.50	6.5	0.61	0.98	0.90/0.94	5.2	0.58	0.92	2.04/2.08	5.9	0.60	0.99
TRE	1.17/1.18	4.2	0.59	0.96	0.57/0.58	6.5	0.60	1.00	1.33/1.33	5.2	0.60	0.92	2.83/2.82	7.5	0.52	0.95
Av.	-	15.64	0.59	0.71	-	11.61	0.56	0.91	-	9.77	0.57	0.81	-	17.28	0.59	0.74
Win initial	9/9	9	0	7	8/8	9	1	6	9/9	9	9	9	1/1	9	9	8
Results on the most interpretable point for Accuracy/Gm3m plane																
Dataset	Ling1				Ling2				Scat1				Scat2			
	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi
PLA	2.54/2.59	33.6	0.94	0.01	3.69/3.74	14.2	0.95	0.93	3.97/4.01	35.8	0.86	0.25	3.97/4.02	14.6	0.94	0.25
QUA	0.04/0.04	120.9	0.95	0.00	0.04/0.04	34.0	0.89	0.94	0.05/0.05	32.0	0.78	0.24	0.04/0.04	57.5	0.91	0.00
ELE	97468/106125	53.8	0.90	0.40	354767/365093	45.1	0.96	0.99	169819/178258	26.0	0.87	0.83	219667/226300	49.5	0.86	0.47
ABA	10.05/10.19	20.8	0.87	0.49	15.66/15.89	29.4	0.93	0.98	19.53/19.46	11.5	0.93	0.86	7.17/7.28	25.3	0.76	0.31
STP	9.56/9.75	23.7	0.97	0.86	5.54/5.76	33.2	0.94	0.98	15.68/15.94	10.0	0.94	0.99	5.77/5.57	25.2	0.82	0.69
WIZ	10.01/10.92	23.8	0.94	0.89	9.06/10.30	52.8	0.94	0.97	40.24/42.96	10.5	0.87	0.95	18.52/19.75	17.9	0.91	0.80
WAN	22.67/23.19	16.0	0.92	0.97	15.39/16.16	63.6	0.94	0.97	39.48/40.32	9.9	0.87	0.93	35.86/36.95	24.1	0.89	0.74
MOR	0.81/0.83	12.8	0.89	0.98	0.83/0.82	26.4	0.87	0.97	2.16/2.20	8.6	0.86	0.97	2.94/2.90	11.3	0.90	0.76
TRE	0.79/0.79	11.6	0.82	0.87	1.29/1.33	20.5	0.79	0.97	3.03/3.16	8.3	0.80	0.88	3.12/3.15	15.3	0.82	0.55
Av.	-	35.23	0.91	0.61	-	35.47	0.91	0.97	-	16.96	0.86	0.77	-	26.73	0.87	0.51
Win initial	9/9	9	0	7	7/8	9	7	7	9/9	9	9	9	2/2	9	9	4
Results on the most interpretable point for Accuracy/Rmi plane																
Dataset	Ling1				Ling2				Scat1				Scat2			
	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi	$MSE_{tra/tst}$	NR	Gm3m	Rmi
PLA	2.63/2.72	26.6	0.72	0.55	2.71/2.82	12.6	0.59	1.00	3.82/3.92	27.1	0.73	0.90	4.93/4.98	11.7	0.81	0.92
QUA	0.03/0.04	96.5	0.47	0.00	0.04/0.04	28.3	0.48	1.00	0.05/0.05	29.0	0.66	0.65	0.03/0.04	59.3	0.46	0.04
ELE	92340/101194	29.3	0.64	0.89	274481/284773	40.7	0.82	1.00	154191/163598	26.2	0.84	0.97	68156/75396	54.0	0.52	0.92
ABA	6.74/6.80	11.4	0.55	0.98	5.23/5.49	17.7	0.45	1.00	18.63/18.55	11.5	0.82	0.99	6.20/6.56	24.7	0.43	0.95
STP	8.90/9.13	22.9	0.91	1.00	1.48/1.97	26.5	0.48	1.00	14.65/14.94	10.2	0.90	1.00	5.63/5.64	18.8	0.61	1.00
WIZ	6.08/7.06	16.9	0.76	1.00	3.73/4.70	43.5	0.57	1.00	30.48/31.83	10.3	0.77	1.00	17.99/19.32	17.3	0.79	1.00
WAN	15.40/16.44	12.7	0.78	1.00	4.66/5.69	31.7	0.50	1.00	29.14/30.68	10.7	0.77	0.99	36.03/37.64	20.2	0.80	1.00
MOR	0.32/0.33	9.2	0.68	1.00	0.04/0.05	15.6	0.59	1.00	1.47/1.55	8.7	0.73	1.00	0.82/0.90	10.4	0.46	1.00
TRE	0.31/0.32	8.1	0.64	1.00	0.06/0.09	15.5	0.64	1.00	1.77/1.90	8.9	0.69	1.00	1.97/2.03	13.7	0.42	1.00
Av.	-	25.95	0.68	0.83	-	25.78	0.57	1.00	-	15.84	0.77	0.94	-	25.55	0.59	0.87
Win initial	9/9	9	0	7	9/9	9	1	9	9/9	9	9	9	3/3	9	9	9

Table 13: Rankings obtained through Friedman's Test on the different measures and planes for the models on the most interpretable point

Accuracy/NR plane					Accuracy/Gm3m plane					Accuracy/Rmi plane				
Alg.	MSE_{tst}	NR	Gm3m	Rmi	Alg.	MSE_{tst}	NR	Gm3m	Rmi	Alg.	MSE_{tst}	NR	Gm3m	Rmi
Ling1	1.5556	1.7778	2.4444	3.0000	Ling1	1.6667	2.7778	1.7778	3.0000	Ling1	2.1111	2.2222	2.3333	2.8889
Ling2	1.5556	2.8889	2.6667	1.5556	Ling2	2.1111	3.2222	2.1111	1.2222	Ling2	1.4444	3.2222	3.2222	1.0000
Scat1	3.1111	2.1111	2.5556	3.0000	Scat1	3.5556	1.3333	3.3333	2.1111	Scat1	3.5556	1.6667	1.5556	3.2222
Scat2	3.7778	3.2222	2.3333	2.4444	Scat2	2.6667	2.6667	2.7778	3.6667	Scat2	2.8889	2.8889	2.8889	2.8889

In general, the linguistic algorithms show a better performance to improve semantic interpretability, but the scatter ones can obtain final models with a good interpretability in some cases. For example, it is possible to obtain the maximum Rmi value for the Accuracy/Rmi plane with a low number of rules and high Gm3m in some datasets (STP,WIZ,MOR,TRE) using Scat1 or Scat2. The drawback is that in these cases the accuracy obtained by Ling1 or Ling2 is better than the accuracy obtained with Scat1 or Scat2. Therefore,

Table 14: Holm's post-hoc test with $\alpha = 0.1$ on the different measures for models on the most interpretable point

Accuracy/NR plane																
Holm on MSE_{tst}						Holm on NR						Holm on Gm3m				
i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.
3	Scat2	3.65	0.000	0.03	Rej.	3	Scat2	2.37	0.018	0.03	Rej.	3	Ling2	0.55	0.584	0.03	Acc.
2	Scat1	2.56	0.011	0.05	Rej.	2	Ling2	1.83	0.068	0.05	Acc.	2	Scat1	0.37	0.715	0.05	Acc.
1	Ling1	0.00	1.000	0.10	Acc.	1	Scat1	0.55	0.584	0.10	Acc.	1	Ling1	0.18	0.855	0.10	Acc.

Holm on Rmi					
i	Alg.	z	p	α/i	Hyp.
3	Ling1	2.37	0.018	0.03	Rej.
2	Scat1	2.37	0.018	0.05	Rej.
1	Scat2	1.46	0.144	0.10	Acc.

Accuracy/Gm3m plane																
Holm on MSE_{tst}						Holm on NR						Holm on Gm3m				
i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.
3	Scat1	3.10	0.002	0.03	Rej.	3	Ling2	3.10	0.002	0.03	Rej.	3	Scat1	2.56	0.011	0.03	Rej.
2	Scat2	1.64	0.100	0.05	Acc.	2	Ling1	2.37	0.018	0.05	Rej.	2	Scat2	1.64	0.100	0.05	Acc.
1	Ling2	0.73	0.465	0.10	Acc.	1	Scat2	2.19	0.028	0.10	Rej.	1	Ling2	0.55	0.584	0.10	Acc.

Holm on Rmi					
i	Alg.	z	p	α/i	Hyp.
3	Scat2	4.02	0.000	0.03	Rej.
2	Ling1	2.92	0.003	0.05	Rej.
1	Scat1	1.46	0.144	0.10	Acc.

Accuracy/Rmi plane																
Holm on MSE_{tst}						Holm on NR						Holm on Gm3m				
i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.	i	Alg.	z	p	α/i	Hyp.
3	Scat1	3.47	0.001	0.03	Rej.	3	Ling2	2.56	0.011	0.03	Rej.	3	Ling2	2.74	0.006	0.03	Rej.
2	Scat2	2.37	0.018	0.05	Rej.	2	Scat2	2.01	0.045	0.05	Rej.	2	Scat2	2.19	0.028	0.05	Rej.
1	Ling1	1.10	0.273	0.10	Acc.	1	Ling1	0.91	0.361	0.10	Acc.	1	Ling1	1.28	0.201	0.10	Acc.

Holm on Rmi					
i	Alg.	z	p	α/i	Hyp.
3	Scat1	3.65	0.000	0.03	Rej.
2	Ling1	3.10	0.002	0.05	Rej.
1	Scat2	3.10	0.002	0.10	Rej.

in general, there is no clear winner.

5.5. Global Analysis

Theoretically the linguistic algorithms should generate more interpretable FRBSs with lesser accuracy than scatter approaches, presuming that no MF tuning has been considered. On the other hand, the scatter algorithms generate more accurate FRBS with lower interpretability presuming that no semantic matching of MFs is performed. In this current work, the four algorithms (LING1, LING2, SCAT1 and SCAT2) accomplish both MF tuning and consideration of the semantic quality on MF and RB levels, thus bringing the two approaches together from a different origin.

The results show that, for most cases, the error and the number of rules have been reduced for every experiment, so systems with better accuracy and lower complexity have been obtained. In general, the semantic interpretability of the final models is improved for the scatter approaches, while it is maintained at the initial level for the linguistic ones. Despite this, focusing only on the most accurate and most interpretable solutions, some conclusions can be achieved:

Most accurate solution Scatter algorithms achieve better overall accuracy, number of rules and rule semantic interpretability, for more complex datasets. This better accuracy is due to the fact that each rule MF is initially determined individually, and the complexity of these initial MFs is reduced by the MOEA MFs tuning, improving their interpretability and obtaining a best overall solution. The main problem is that, despite of this improvement, MFs semantic interpretability is very hard to be achieved.

Most interpretable solution Solutions introducing a low number of rules and high Gm3m interpretability are possible by means of the scatter algorithms. Low error and high Rmi interpretability are possible by means of the linguistic algorithms. In fact, the results show that it is possible to obtain accurate models by means of linguistic algorithms and interpretable models by the scatter ones. This breaks the well-established ideas about the performance of scatter and linguistic algorithms. Tentatively, the linguistic algorithms obtain the most interpretable initial models for lower complexity datasets. The accuracy of these models is highly improved by MFs tuning.

Median solution In general, the choice of an algorithm based on the solutions at the mid Pareto front is complicated, since there are no general winners in the different planes.

An important keypoint is that there are some datasets where the general rules from the previous analysis are broken. For example, to obtain an accurate model in the TRE dataset, with a low error and a good interpretability, the user should use a linguistic algorithm. This means that, finally, when we are solving a real world problem, we need to specifically analyze and compare the different models obtained. This is one of the aims of the proposed indexes, which are able to quantify the desired characteristics of a FRBS as complementary aspects to the accuracy and complexity.

Two illustrative examples about how to carry out a comparison, based on the measures proposed in this work, between scatter and linguistic models are shown in the following section.

5.6. Local Comparison between Linguistic and Scatter-based Models in some Example Problems

In the previous section, the statistical tests have shown that pre-established rules are not always valid. The type of modeling to be selected depends on many factors such as data or nature of the problem, or even the algorithms used for modeling. Thus, the scatter-based algorithms do not always generate a more accurate and less interpretable FRBS and vice versa. One of the main causes of this behavior is the complexity of the search space.

Therefore, in order to analyze and compare models with different degrees of accuracy and interpretability, the previously proposed metrics can also be applied locally to evaluate the different aspects of a FRBS, following the representation scheme presented in Section 3.3. As shown below, using the metrics proposed here, the expert can analyze and compare different models in order to take the final decision.

Example 1

First, we have selected two FRBSs obtained with the same fold of the data set ELE, one using Ling1 and another using Scat1. Both correspond to the most accurate FRBS in any of the analyzed planes. Table 15 shows, for each one of the FRBSs, the specific values of accuracy and global interpretability indexes considered in this proposal.

Table 15: FRBSs for dataset ELE

Algorithm	MSE_{tra}	MSE_{tst}	NR	Gm3m	Rmi
Ling1	29301.67	36266.97	29	0.314	0.649
Scat1	17102.27	19865.52	27	0.490	0.730

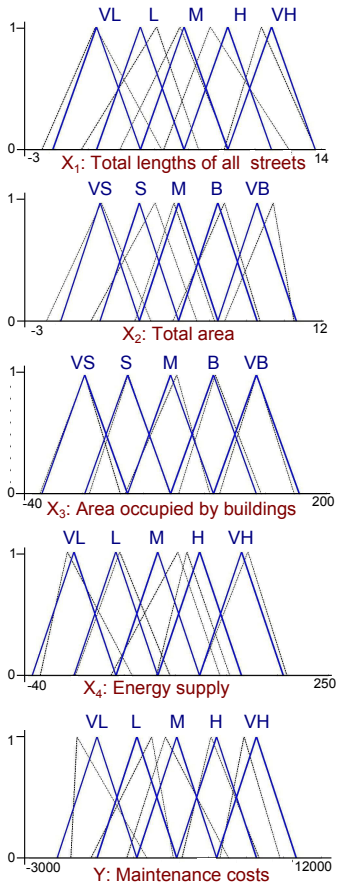
In the ELE problem, the objective is to estimate the maintenance costs of a medium voltage line (Y) from four characteristics: Sum of the lengths of all streets in the town (X_1), Total area of the town (X_2), Area that is occupied by buildings (X_3) and Energy supply to the town (X_4).

The particular KBs for these FRBSs are shown below. Figure 8 shows the DB obtained with Ling1, black line, the initial interpretable fuzzy partition, grey line (in our case, a strongly uniformly distributed fuzzy partition), and the RB of the FRBS represented as proposed in section 3.3, i.e., based on the initial fuzzy partition and showing the individual Rmi for each rule. On the other hand, Figure 9 shows the FRBS obtained with Scat1 described in the same way (pseudo-linguistic representation), and showing the Rmi value for each rule.

Observing the general features of the FRBS, from our point of view (obviously subjective), in this case the best choice is the FRBS obtained with Scat1, since it is not only more accurate, but also obtains better values in all indexes of interpretability in general. Comparing the particular features of both FRBS, it can be said that:

- In the Scat1 FRBS, all rules have high Rmi values, so the confidence of the rules is high and there is no inconsistency among them. The semantic-based interpretability at the RB level is thus high. The Rmi values in the Ling1 FRBS are smaller in almost every rule, showing that the meaning of certain rules is not fully representative of the actual system behavior.

DATA BASE (LING1, ELE)



RULE BASE (LING1, ELE)

Rule	X_1	X_2	X_3	X_4	Y	RMI
R_1	$VL^{0.83}$	$VS^{0.86}$	$VS^{0.90}$	$VL^{0.68}$	$VL^{0.34}$	0.68
R_2	$VL^{0.83}$	$VS^{0.86}$	$VS^{0.90}$	$VL^{0.68}$	$VL^{0.34}$	0.68
R_3	$L^{0.63}$	$VS^{0.86}$	$VS^{0.90}$	$VL^{0.68}$	$VL^{0.34}$	0.67
R_4	$L^{0.63}$	$VS^{0.86}$	$VS^{0.90}$	$L^{0.89}$	$L^{0.60}$	0.70
R_5	$L^{0.63}$	$VS^{0.86}$	$VS^{0.90}$	$L^{0.89}$	$L^{0.60}$	0.70
R_6	$L^{0.63}$	$S^{0.66}$	$S^{0.95}$	$VL^{0.68}$	$VL^{0.34}$	0.69
R_7	$L^{0.63}$	$S^{0.66}$	$S^{0.95}$	$L^{0.89}$	$L^{0.60}$	0.99
R_8	$L^{0.63}$	$S^{0.66}$	$S^{0.95}$	$L^{0.89}$	$M^{0.70}$	0.64
R_9	$M^{0.83}$	$S^{0.66}$	$VS^{0.90}$	$M^{0.62}$	$L^{0.60}$	0.78
R_{10}	$M^{0.83}$	$M^{0.93}$	$S^{0.95}$	$L^{0.89}$	$L^{0.60}$	0.96
R_{11}	$M^{0.83}$	$M^{0.93}$	$S^{0.95}$	$M^{0.62}$	$M^{0.70}$	0.82
R_{12}	$M^{0.83}$	$M^{0.93}$	$M^{0.81}$	$L^{0.89}$	$M^{0.70}$	0.83
R_{13}	$M^{0.83}$	$M^{0.93}$	$M^{0.81}$	$L^{0.89}$	$M^{0.70}$	0.83
R_{14}	$M^{0.83}$	$M^{0.93}$	$M^{0.81}$	$L^{0.89}$	$M^{0.70}$	0.83
R_{15}	$H^{0.62}$	$S^{0.66}$	$S^{0.95}$	$L^{0.89}$	$L^{0.60}$	0.97
R_{16}	$H^{0.62}$	$S^{0.66}$	$S^{0.95}$	$M^{0.62}$	$M^{0.70}$	0.83
R_{17}	$H^{0.62}$	$M^{0.93}$	$S^{0.95}$	$VL^{0.68}$	$L^{0.60}$	0.78
R_{18}	$H^{0.62}$	$M^{0.93}$	$S^{0.95}$	$M^{0.62}$	$M^{0.70}$	0.83
R_{19}	$H^{0.62}$	$M^{0.93}$	$S^{0.95}$	$H^{0.75}$	$M^{0.70}$	0.83
R_{20}	$H^{0.62}$	$M^{0.93}$	$M^{0.81}$	$M^{0.62}$	$H^{0.79}$	0.86
R_{21}	$H^{0.62}$	$B^{0.80}$	$M^{0.81}$	$L^{0.89}$	$M^{0.70}$	0.78
R_{22}	$H^{0.62}$	$B^{0.80}$	$M^{0.81}$	$M^{0.62}$	$H^{0.79}$	0.86
R_{23}	$H^{0.62}$	$B^{0.80}$	$B^{0.90}$	$L^{0.89}$	$H^{0.79}$	0.86
R_{24}	$H^{0.62}$	$VB^{0.60}$	$VB^{0.95}$	$M^{0.62}$	$VH^{0.70}$	0.89
R_{25}	$VH^{0.79}$	$S^{0.66}$	$S^{0.95}$	$VH^{0.86}$	$H^{0.79}$	0.86
R_{26}	$VH^{0.79}$	$S^{0.66}$	$M^{0.81}$	$L^{0.89}$	$M^{0.70}$	0.82
R_{27}	$VH^{0.79}$	$S^{0.66}$	$M^{0.81}$	$L^{0.89}$	$M^{0.70}$	0.82
R_{28}	$VH^{0.79}$	$B^{0.80}$	$M^{0.81}$	$VH^{0.86}$	$VH^{0.70}$	0.89
R_{29}	$VH^{0.79}$	$B^{0.80}$	$M^{0.81}$	$L^{0.89}$	$M^{0.70}$	0.79

Linguistic Terms
VL: Very Low, L: Low, M: Medium, H: High, VH: Very High
VS: Very Small, S: Small, M: Medium, B: Big, VB: Very Big

Accuracy
MSE_{tra} =29301.67 ; MSE_{tst}=36266.97

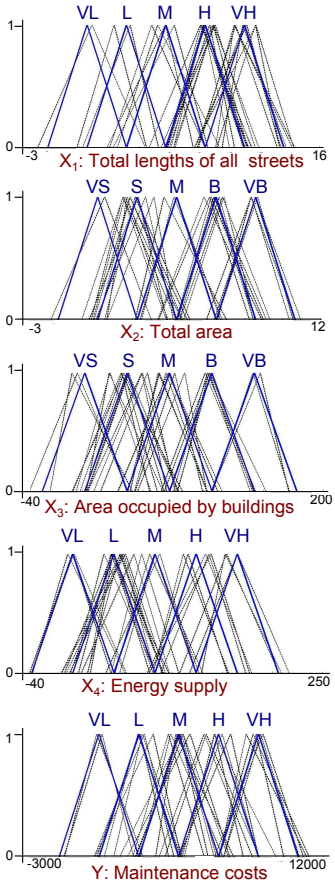
Figure 8: Linguistic model obtained with LING1 for ELE

- In the SCAT1 FRBS, most of the MFs used in the scatter-based rules can be associated with the corresponding MFs given in the initial partition with a high degree of similarity, so its semantic-based interpretability at the MF level is relatively high.
- In the LING1 FRBS, the final number of MFs per variable is lower than in SCAT1, but these MFs are further away from the initial MFs, so its GM3M value is lower in a larger number of cases.
- The number of rules of both FRBSs is similar, so it is not a determinant factor when choosing one model or another.

Example 2

For the second example, we have selected two FRBSs obtained with the same fold of the data set PLA, one using LING2 and another using SCAT2. Both correspond to the median FRBS in the Accuracy/*NR* plane. Table 16 shows, for each one of the FRBSs, the specific values of accuracy and interpretability global indexes considered in this case.

DATA BASE (SCAT1, ELE)



RULE BASE (SCAT1, ELE)

Rule	X_1	X_2	X_3	X_4	Y	RMI
R_1	$L^{0.80}$	$S^{0.76}$	$S^{0.65}$	$L^{0.78}$	$L^{0.65}$	0.88
R_2	$VL^{0.85}$	$VS^{0.82}$	$VS^{0.87}$	$VL^{0.88}$	$VL^{0.86}$	0.89
R_3	$L^{0.80}$	$S^{0.75}$	$S^{0.84}$	$L^{0.87}$	$L^{0.82}$	0.88
R_4	$M^{0.70}$	$S^{0.78}$	$S^{0.75}$	$L^{0.75}$	$L^{0.79}$	0.97
R_5	$M^{0.74}$	$S^{0.73}$	$S^{0.64}$	$L^{0.68}$	$M^{0.70}$	0.73
R_6	$H^{0.79}$	$M^{0.84}$	$S^{0.76}$	$M^{0.80}$	$M^{0.96}$	0.97
R_7	$M^{0.73}$	$B^{0.80}$	$M^{0.78}$	$L^{0.71}$	$M^{0.87}$	0.82
R_8	$VH^{0.88}$	$M^{0.93}$	$S^{0.84}$	$L^{0.77}$	$L^{0.83}$	0.99
R_9	$H^{0.81}$	$S^{0.65}$	$S^{0.60}$	$H^{0.75}$	$M^{0.74}$	0.90
R_{10}	$VH^{0.71}$	$S^{0.79}$	$S^{0.90}$	$L^{0.83}$	$M^{0.90}$	0.98
R_{11}	$H^{0.80}$	$B^{0.94}$	$B^{0.88}$	$M^{0.92}$	$H^{0.92}$	0.98
R_{12}	$H^{0.87}$	$B^{0.75}$	$M^{0.62}$	$M^{0.76}$	$H^{0.76}$	0.97
R_{13}	$H^{0.90}$	$S^{0.97}$	$S^{0.90}$	$VL^{0.95}$	$L^{0.90}$	0.94
R_{14}	$VH^{0.81}$	$B^{0.82}$	$M^{0.84}$	$L^{0.75}$	$M^{0.83}$	0.95
R_{15}	$H^{0.86}$	$B^{0.87}$	$M^{0.78}$	$H^{0.68}$	$H^{0.79}$	0.97
R_{16}	$VH^{0.61}$	$S^{0.75}$	$S^{0.77}$	$H^{0.86}$	$M^{0.76}$	0.78
R_{17}	$H^{0.91}$	$VB^{0.86}$	$VB^{0.92}$	$M^{0.81}$	$VH^{0.77}$	0.97
R_{18}	$H^{0.95}$	$B^{0.83}$	$B^{0.84}$	$L^{0.72}$	$H^{0.84}$	0.95
R_{19}	$VH^{0.70}$	$S^{0.69}$	$M^{0.90}$	$VH^{0.91}$	$H^{0.72}$	0.86
R_{20}	$H^{0.86}$	$VB^{0.94}$	$B^{0.84}$	$H^{0.79}$	$VH^{0.84}$	0.94
R_{21}	$M^{0.75}$	$S^{0.82}$	$VS^{0.62}$	$VL^{0.78}$	$VL^{0.87}$	0.94
R_{22}	$H^{0.81}$	$VB^{0.92}$	$B^{0.85}$	$L^{0.90}$	$M^{0.84}$	0.99
R_{23}	$H^{0.84}$	$VB^{0.89}$	$VB^{0.77}$	$L^{0.78}$	$VH^{0.87}$	0.82
R_{24}	$VH^{0.70}$	$M^{0.64}$	$M^{0.72}$	$L^{0.92}$	$M^{0.67}$	0.99
R_{25}	$VH^{0.71}$	$B^{0.89}$	$M^{0.92}$	$VH^{0.65}$	$VH^{0.65}$	0.87
R_{26}	$H^{0.76}$	$B^{0.82}$	$B^{0.88}$	$H^{0.78}$	$VH^{0.94}$	0.98
R_{27}	$VH^{0.89}$	$S^{0.61}$	$S^{0.80}$	$VH^{0.81}$	$H^{0.72}$	0.85

Linguistic Terms
VL: Very Low, L: Low, M: Medium, H: High, VH: Very High
VS: Very Small, S: Small, M: Medium, B: Big, VB: Very Big

Accuracy
MSE_{tra} =17102.27 ; MSE_{tst}=19865.52

Figure 9: Pseudo-linguistic model obtained with SCAT1 for ELE

Table 16: FRBSs for dataset PLA

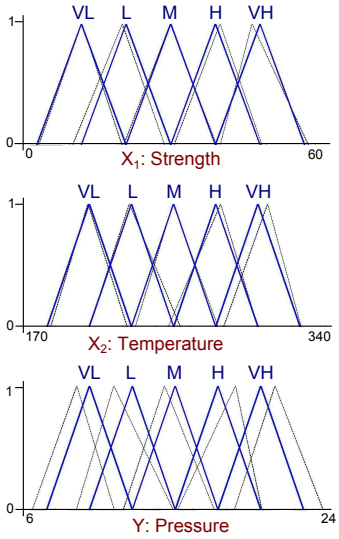
Algorithm	MSE_{tra}	MSE_{tst}	NR	GM3M	RMI
LING2	2.59	2.59	14	0.613	1.000
SCAT2	2.10	2.30	23	0.466	0.380

PLA is a regression data set where the task is to compute how much pressure (Y) a given piece of plastic can stand when some strength (X_1) is applied on it at a fixed temperature (X_2). As in the previous example, figure 10 shows the FRBS obtained with LING2, while figure 11 shows the FRBS obtained with SCAT2.

According to our own criteria (again subjective), in this case the best choice should be the LING2 FRBS, since it shows better interpretability properties at all levels. In this case, the loss of accuracy with respect to SCAT2 seems minimal. Comparing the particular features of both FRBSs we have that:

- The number of rules is lower in the LING2 FRBS, where the set of rules is small enough so it can be easily interpreted by a user.
- LING2 obtains the maximum value of semantic-based interpretability at the RB level. This value indicates that there is no inconsistency among the rules. In SCAT2, the RMI values indicate that most

DATA BASE (LING2, PLA)



RULE BASE (LING2, PLA)

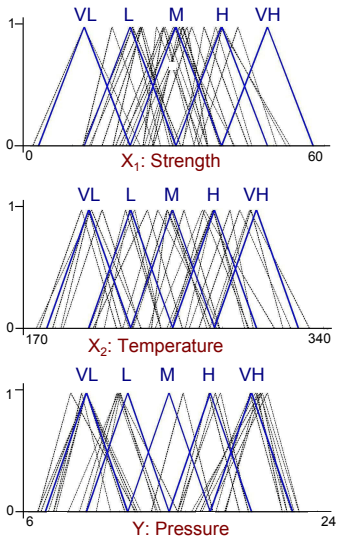
Rule	X_1	X_2	Y	RMI
R_1	$VL^{0.95}$	$VL^{0.94}$	$VH^{0.81}$	1.00
R_2	$VL^{0.95}$	$L^{0.88}$	$VH^{0.81}$	1.00
R_3	$L^{0.91}$	$VL^{0.94}$	$M^{0.83}$	1.00
R_4	$L^{0.91}$	$L^{0.88}$	$H^{0.64}$	1.00
R_5	$L^{0.91}$	$M^{0.93}$	$VH^{0.81}$	1.00
R_6	$M^{0.98}$	$VL^{0.94}$	$VL^{0.82}$	1.00
R_7	$M^{0.98}$	$L^{0.88}$	$L^{0.68}$	1.00
R_8	$M^{0.98}$	$H^{0.83}$	$H^{0.64}$	1.00
R_9	$M^{0.98}$	$VH^{0.81}$	$VH^{0.81}$	1.00
R_{10}	$H^{0.93}$	$M^{0.93}$	$VL^{0.82}$	1.00
R_{11}	$H^{0.93}$	$H^{0.83}$	$L^{0.68}$	1.00
R_{12}	$H^{0.93}$	$VH^{0.81}$	$M^{0.83}$	1.00
R_{13}	$VH^{0.77}$	$H^{0.83}$	$VL^{0.82}$	1.00
R_{14}	$VH^{0.77}$	$VH^{0.81}$	$VL^{0.82}$	1.00

Linguistic Terms
VL: Very Low, L: Low, M: Medium
H: High, VH: Very High

Accuracy
MSE_{tra} =2.59 ; MSE_{tst}=2.59

Figure 10: Linguistic model obtained with LING2 for PLA

DATA BASE (SCAT2, PLA)



RULE BASE (SCAT2, PLA)

Rule	X_1	X_2	Y	RMI
R_1	$L^{0.68}$	$VL^{0.85}$	$VH^{0.74}$	0.50
R_2	$M^{0.47}$	$M^{0.92}$	$H^{0.89}$	0.99
R_3	$M^{0.61}$	$L^{0.71}$	$VL^{0.63}$	0.69
R_4	$M^{0.84}$	$VH^{0.85}$	$VH^{0.76}$	0.97
R_5	$L^{0.76}$	$VL^{0.85}$	$L^{0.74}$	0.70
R_6	$H^{0.81}$	$VH^{0.76}$	$VL^{0.86}$	0.56
R_7	$H^{0.80}$	$H^{0.85}$	$VL^{0.90}$	0.81
R_8	$H^{0.59}$	$VH^{0.78}$	$H^{0.75}$	0.86
R_9	$L^{0.79}$	$VL^{0.83}$	$VL^{0.82}$	0.57
R_{10}	$M^{0.84}$	$H^{0.85}$	$H^{0.89}$	0.69
R_{11}	$L^{0.94}$	$M^{0.69}$	$VH^{0.78}$	0.83
R_{12}	$M^{0.63}$	$H^{0.66}$	$VH^{0.66}$	0.59
R_{13}	$M^{0.88}$	$VL^{0.87}$	$VL^{0.84}$	0.81
R_{14}	$M^{0.88}$	$H^{0.64}$	$VH^{0.90}$	0.64
R_{15}	$L^{0.79}$	$L^{0.76}$	$VH^{0.68}$	0.49
R_{16}	$M^{0.80}$	$M^{0.86}$	$VL^{0.90}$	0.57
R_{17}	$M^{0.81}$	$L^{0.83}$	$L^{0.86}$	0.55
R_{18}	$M^{0.73}$	$M^{0.84}$	$VH^{0.72}$	0.38
R_{19}	$H^{0.75}$	$H^{0.88}$	$VL^{0.89}$	0.75
R_{20}	$M^{0.76}$	$M^{0.59}$	$L^{0.83}$	0.62
R_{21}	$L^{0.76}$	$L^{0.80}$	$M^{0.77}$	0.69
R_{22}	$H^{0.84}$	$H^{0.83}$	$L^{0.73}$	0.76
R_{23}	$VL^{0.91}$	$VL^{0.90}$	$VH^{0.89}$	0.99

Linguistic Terms
VL: Very Low, L: Low, M: Medium
H: High, VH: Very High

Accuracy
MSE_{tra} =2.10 ; MSE_{tst}=2.30

Figure 11: Pseudo-linguistic model obtained with SCAT2 for PLA

of the rules have a medium level of confidence, so they are not so representative of the real behavior of the model.

- In general, the MFs obtained by LING2 are closer to the initial linguistic MFs than those obtained by SCAT2.

6. Conclusions

The objectives of this work are two fold. On the one hand, it proposes the extension of the well-known GM3M index [28] and a new index of semantic-based interpretability at the RB level named RMI in order to have more reliable semantic interpretability measures that could be applied to any type of FRBS, with any type of MF and inference system. On the other hand, a local representation for comparison between linguistic and scatter FRBSs is presented, so the user can choose the best model at each moment.

The extension of the GM3M allows the application of this index to quantify the semantic interpretability at the level of fuzzy partitions independently of the linguistic or scatter nature of the system and the type of MFs used.

The new index named RMI is based on assessing the degree of reliability of each one of the rules with respect to the others in the RB. The index is calculated taking into account the particular inference system used in the FRBS to indicate the worst case of interaction, so that the index is capable of detecting some problems like a bad choice of operators [55, 20] or those resulting from the use of weights in the rules.

Using a post-processing based on MOEAs developed ad-hoc for this proposal, the features of accuracy and interpretability of the FRBSs are improved for both linguistic and scatter approaches. The new RMI index and the GM3M extension are used to guide the post-processing based on genetic rule selection and tuning of MFs.

The checking of this proposal is carried out using nine cases of study from the KEEL dataset repository, and four fuzzy modeling algorithms, two linguistic (NefProx and L-IRL) and two scatter (FasArt and S-IRL). The algorithms generate FRBSs with different initial features of accuracy and interpretability, so that the proposal is validated in different contexts. The experimental results have shown that: there are no general trends in the performance of the algorithms considered, so it is difficult to say that any of these algorithms is better than the others in any of the listed objectives. The general rules are broken and/or do not work as expected, so it is necessary to analyze each case in particular in order to know which approach is better. To do this, the proposal also provides a local comparison support of FRBS models, regardless of their linguistic or scatter nature, which can be used by a user in a decision making process. This allows the user to select the most interesting set of rules.

References

[1] R. Alcalá, J. Alcalá-Fdez, F. Herrera, J. Otero, Genetic learning of accurate and compact fuzzy rule based systems based on the 2-tuples linguistic representation, International Journal of Approximate Reasoning 44 (2007) 45–64.

[2] R. Alcalá, P. Ducange, F. Herrera, B. Lazzerini, F. Marcelloni, A multi-objective evolutionary approach to concurrently learn rule and data bases of linguistic fuzzy rule-based systems, IEEE Transactions on Fuzzy Systems 17 (2009) 1106–1122.

[3] R. Alcalá, M.J. Gacto, F. Herrera, J. Alcalá-Fdez, A multi-objective genetic algorithm for tuning and rule selection to obtain accurate and compact linguistic fuzzy rule-based systems, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15 (2007) 539–557.

[4] J. Alcalá-Fdez, L. Sánchez, S. García, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V. Rivas, J. Fernández, F. Herrera, KEEL: A software tool to assess evolutionary algorithms to data mining problems, Soft Computing 13 (2009) 307–318.

[5] J.M. Alonso, L. Magdalena, HILK++: an interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers, Soft Computing 15 (2011) 1959 – 1980.

[6] J.M. Alonso, L. Magdalena, S. Guillaume, HILK: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism, International Journal of Intelligent Systems 23 (2008) 761–794.

[7] J.M. Alonso, L. Magdalena, G.G. Rodríguez, Looking for a good fuzzy system interpretability index: An experimental approach, International Journal of Approximate Reasoning 51 (2009) 115–134.

[8] M. Antonelli, P. Ducange, B. Lazzerini, F. Marcelloni, Multi-objective evolutionary learning of granularity, membership function parameters and rules of mamdani fuzzy systems, Evolutionary Intelligence 2 (2009) 21 – 37.

[9] M. Antonelli, P. Ducange, B. Lazzerini, F. Marcelloni, Learning knowledge bases of multi-objective evolutionary fuzzy systems by simultaneously optimizing accuracy, complexity and partition integrity, Soft Computing 15 (2011) 2335 – 2354.

[10] U. Bodenhofer, P. Bauer, A formal model of interpretability of linguistic variables, in: J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Interpretability issues in fuzzy modeling, Springer-Verlag, 2003, pp. 524–545.
[11] A. Botta, B. Lazzerini, F. Marcelloni, D.C. Stefanescu, Context adaptation of fuzzy systems through a multi-objective evolutionary approach based on a novel interpretability index, Soft Computing 13 (2009) 437–449.
[12] J. Cano Izquierdo, Y. Dimitriadis, E. Gómez Sánchez, J. López Coronado, Learnning from noisy information in FasArt and Fasback neuro-fuzzy systems, Neural Networks 14 (2001) 407–425.
[13] J. Casillas, O. Cordón, F. Herrera, L. Magdalena, Accuracy improvements to find the balance interpretability-accuracy in fuzzy modeling: An overview, in: J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Accuracy Improvements in Linguistic Fuzzy Modelling, volume 129 of *Studies in Fuzziness and SoftComputing*, Springer-Verlag, Berlin Heildelberg, 2003, pp. 3–24.
[14] J. Casillas, O. Cordón, F. Herrera, L. Magdalena, Interpretability improvements to find the balance interpretability-accuracy in fuzzy modeling: An overview, in: J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Interpretability issues in fuzzy modeling, Springer-Verlag, 2003, pp. 3–22.
[15] O. Castillo, P. Melin, W. Pedrycz, Design of interval type-2 fuzzy models through optimal granularity allocation, Applied Soft Computing 11 (2011) 5590 – 5601.
[16] M. Cococcioni, P. Ducange, B. Lazzerini, F. Marcelloni, A pareto-based multi-objective evolutionary approach to the identification of Mamdani fuzzy systems, Soft Computing 11 (2007) 1013–1031.
[17] O. Cordón, F. Herrera, A three-stage evolutionary process for learning descriptive and approximate fuzzy logic controller knowledge bases from examples, International Journal of Approximate Reasoning 17 (1997) 369–407.
[18] O. Cordón, F. Herrera, Hybridizing genetic algorithms with sharing scheme and evolution strategies for designing approximate fuzzy rule-based systems, Fuzzy Sets and Systems 118 (2001) 235 – 255.
[19] O. Cordón, F. Herrera, L. Magdalena, P. Villar, A genetic learning process for the scaling factors, granularity and contexts of the fuzzy rule-based system data base, Information Science 136 (2001) 85–107.
[20] O. Cordón, F. Herrera, A. Peregrín, Applicability of the fuzzy operators in the design of fuzzy logic controllers, Fuzzy Sets and Systems (1997) 15 – 41.
[21] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.
[22] L.J. Eshelman, The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination, in: G. Rawlin (Ed.), Foundations of genetic Algorithms, volume 1, Morgan Kaufman, 1991, pp. 265–283.
[23] L.J. Eshelman, J.D. Schaffer, Real-coded genetic algorithms and interval-schemata, Foundations of Genetic Algorithms 2 (1993) 187–202.
[24] P. Fazendeiro, J.V. de Oliveira, W. Pedrycz, A multiobjective design of a patient and anaesthetist-friendly neuromuscular blockade controller, IEEE Trans. on Biomedical Engineering 54 (2007) 1667–1678.
[25] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (1937) 675–701.
[26] R. Frigg, S. Hartmann, Models in science, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2012 ed., http://plato.stanford.edu/archives/fall2012/entries/models-science/, 2012.
[27] M.J. Gacto, R. Alcalá, F. Herrera, Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems, Soft Computing 13 (2009) 419–436.
[28] M.J. Gacto, R. Alcalá, F. Herrera, Integration of an index to preserve the semantic interpretability in the multi-objective evolutionary rule selection and tuning of linguistic fuzzy systems, IEEE Transactions on Fuzzy Systems 18 (2010) 515–531.
[29] M.J. Gacto, R. Alcalá, F. Herrera, Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Information Sciences 181 (2011) 43404360.
[30] S. García, A. Fernández, J. Luengo, F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability, Soft Computing 13 (2009) 959–977.
[31] S. García, F. Herrera, An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons, Journal of Machine Learning Research 9 (2008) 2677–2694.
[32] S. Guillaume, Designing fuzzy inference systems from data: An interpretability-oriented review, IEEE Trans. Fuzzy Syst. 9 (2001) 426–443.
[33] F. Herrera, Genetic fuzzy systems: taxonomy, current research trends and prospects, Evolutionary Intelligence 1 (2008) 27–46.
[34] F. Herrera, M. Lozano, J.L. Verdegay, Tuning fuzzy logic controllers by genetic algorithms, International J. of Approximate Reasoning 12 (1995) 299–315.
[35] D. Hidalgo, P. Melin, O. Castillo, An optimization method for designing type-2 fuzzy inference systems based on the footprint of uncertainty using genetic algorithms, Expert Systems with Applications 39 (2012) 4590 – 4598.
[36] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Statist. 6 (1979) 65–70.
[37] R.L. Iman, J.H. Davenport, Approximations of the critical region of the friedman statistic, Comm. Statist. Part A Theory Methods 9 (1980) 571–595.
[38] H. Ishibuchi, T. Murata, I.B. Türksen, Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems, Fuzzy Sets and Systems 89 (1997) 135–150.
[39] H. Ishibuchi, K. Nozaki, N. Yamamoto, H. Tanaka, Selecting fuzzy if-then rules for classification problems using genetic algorithms, IEEE Trans. Fuzzy Syst. 3 (1995) 260–270.
[40] H. Ishibuchi, T. Yamamoto, Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining, Fuzzy Sets and Systems 141 (2004) 59–88.
[41] C.L. Karr, Genetic algorithms for fuzzy controllers, AI Expert 6 (1991) 26–33.
[42] A.A. Márquez, F.A. Márquez, A. Peregrín, A mechanism to improve the interpretability of linguistic fuzzy systems with adaptive defuzzification based on the use of a multi-objective evolutionary algorithm, International Journal of Computational Intelligence Systems 5 (2012) 297 – 321.

[43] P. Melin, F. Olivas, O. Castillo, F. Valdez, J. Soria, M. Valdez, Optimal design of fuzzy classification systems using PSO with dynamic parameter adaptation through fuzzy logic, Expert Systems with Applications 40 (2013) 3196 – 3206.
[44] C. Mencar, C. Castiello, R. Cannone, A. Fanelli, Design of fuzzy rule-based classifiers with semantic cointension, Information Science 181:20 (2011) 4361 – 4377.
[45] C. Mencar, C. Castiello, R. Cannone, A. Fanelli, Interpretability assessment of fuzzy knowledge bases: A cointension based approach, International Journal of Approximate Reasoning 52 (2011) 501 – 518.
[46] C. Mencar, A. Fanelli, Interpretability constraints for fuzzy information granulation, Information Sciences 178 (2008) 4585–4618.
[47] D. Nauck, R. Kruse, Neuro-fuzzy systems for function approximation, Fuzzy Sets and Systems 101 (1999) 261–271.
[48] J.V. de Oliveira, Semantic constraints for membership function optimization, IEEE Trans. Syst., Man, Cybern. - Part A: Systems and Humans 29 (1999) 128–138.
[49] D.P. Pancho, J.M. Alonso, O. Cordón, A. Quirin, L. Magdalena, FINGRAMS: Visual representations of fuzzy rule-based inference for expert analysis of comprehensibility, IEEE Transactions on Fuzzy Systems 21 (2013) 1133 – 1149.
[50] P. Pulkkinen, J. Hytönen, H. Koivisto, Developing a bioaerosol detector using hybrid genetic fuzzy systems, Engineering Applications of Artificial Intelligence 21 (2008) 1330–1346.
[51] P. Pulkkinen, H. Koivisto, A dynamically constrained multiobjective genetic fuzzy system for regression problems, IEEE Transactions on Fuzzy Systems 18 (2010) 161–177.
[52] J.A. Roubos, M. Setnes, Compact and transparent fuzzy models and classifiers through iterative complexity reduction, IEEE Trans. Fuzzy Syst. 9 (2001) 516–524.
[53] H. Schichl, Models and history of modeling, in: Modeling languages in mathematical optimization, volume 88 of *Applied Optimization*, Springer, 2004.
[54] M. Setnes, R. Babuska, H. Verbruggen, Rule-based modeling: precision and transparency, IEEE Trans. Syst. Man Cybern. Part B Cybern. 28 (1998) 165–169.
[55] E. Trillas, A.R. de Soto, S. Cubillo, A glance at implication and t-conditional functions, in: V. Novák, I. Perfilieva (Eds.), Discovering the world with fuzzy logic, Physica-Verlag GmbH, Heidelberg, Germany, Germany, 2000, pp. 126–147. URL: `http://dl.acm.org/citation.cfm?id=357564.357576`.
[56] L.X. Wang, J.M. Mendel, Generating fuzzy rules by learning from examples, IEEE Trans. Syst., Man, Cybern. 22 (1992) 1414–1427.
[57] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Trans. Syst., Man, Cybern. 3 (1973) 28–44.
[58] S.M. Zhou, 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 (2008) 3091–3131.
[59] E. Zitzler, M. Laumanns, L. Thiele, Spea2: Improving the strength pareto evolutionary algorithm for multiobjective optimization, in: Proc. Evolutionary Methods for Design, Optimization and Control with App. to Industrial Problems, Barcelona, Spain, pp. 95–100.