FIR Forecasting Strategies Able to Cope with Missing Data: A Smart Grid Application

Sergio Jurado[a], Àngela Nebot[b], Fransisco Mugica[b] and Mihail Mihaylov[c]

[a] *Sensing & Control Systems, Aragó 208-210, 08011 Barcelona, Spain*
Email: sergio.jurado @sensingcontrol.com
Email: s.juradogomez@gmail.com
Phone: +34 605 565 303;
[b] *Soft Computing research group, Technical University of Catalonia, Jordi Girona 1-3, 08034 Barcelona, Spain*
Email:{angela,fmugica}@lsi.upc.edu
[c] *AI lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*
Email: mmihaylo@vub.ac.be

Abstract: Dealing with missing data is of great practical and theoretical interest in forecasting applications. In this study, we deal with the problem of forecasting with missing data in smart grid and smart home applications, where the information from home area sensors and/or smart meters is sometimes missing, which may hinder or even prevent the forecasting of the next hours and days. In concrete, we focus in a Soft Computing technique called Fuzzy Inductive Reasoning (FIR) and its improved version that can cope with missing information in the forecasting process: flexible FIR. In this article eight different strategies for flexible FIR forecasting are defined and studied taking into account: causal relevance of input variables, consistency of predictions, inertia criterion and K-Nearest Neighbours. Furthermore, we evaluate the implications of prediction accuracy and number of registers predicted, when the number of Missing Values (MVs) in the training dataset is increased progressively. To this end, a real smart grid forecasting application, i.e. electricity load forecasting, has been chosen in this study. The results show that all eight strategies proposed are able to cope with MVs and take advantage of the inherent information in the data, with better results in those strategies making use of causal relevance. In addition, the robustness of flexible FIR and its eight strategies are proved taking into account that the percentage of registers predicted is on average *96.15%* when the *%MVs* in training dataset was around *73%.*

Keywords: Soft Computing, Fuzzy Inductive Reasoning, Entropy-based Feature Selection, Prediction with Missing Values, Energy Modelling

1. Introduction

The problem of missing data is of great practical and theoretical interest in forecasting applications. It is important to know how to react to certain situations where Missing Values (MVs) are present both during the model generation and the offline/online forecasting. As an example, when a sensor fails in a production process, it might not be necessary to stop everything if sufficient information is implicitly contained in the remaining sensor data. Furthermore, in economic forecasting, one might want to continue to use a predictor even when an input variable becomes meaningless (for example, due to political changes in a country). In real breast cancer problems [1], missing data imputation is an important task in cases where it is crucial to use all available data and not discard records with missing values.

In this study, we deal with the missing data problem in smart grid or smart home forecasting applications. The information that arrives from the different sensors in the home area network and/or the smart meters, may contain missing data, which may hinder or even prevent the forecasting of the next hours and days. This issue is observed in projects such as iURBAN [2] and GreenCom [3], where either the smart metering infrastructure or the smart home gateway, occasionally, does not send data correctly. This may be caused by a loss in the Internet connection of the gateway, fail in the communication between smart meters and concentrators, or issues between the database interfaces.

In many studies, the problem of MVs is treated from a pre-processing perspective; conventional missing data imputation techniques, such as the substitution with the mean for an unknown feature is studied in [4], which can lead to solutions that are far from optimal. In [5] Lakshminarayan et al., explore the use of machine-learning based alternatives to standard statistical data completion methods, for dealing with missing data. Barladi et al. [6] propose a novel method for missing data reconstruction by fuzzy similarity. Other studies where the missing data problem is approached from a pre-processing point of view are [7] and [8]. However, conventional missing data deletion techniques like list-wise and pair-wise have been observed to present several limitations, such as discarding a great deal of potentially usable data or wrong correlation matrices respectively. Moreover, in smart grid or smart home applications where online predictions are needed, systems must deal continuously with the problem of unknown features.

In this paper, we address the missing data problem found out in a Soft Computing technique called Fuzzy Inductive Reasoning (FIR) [9][10]. Although its popularity is not comparable to other Soft Computing techniques such as Neural Networks (NN), this methodology has been proved to model real complex systems with high accuracy compared to other typical Artificial Intelligence (AI) and statistical techniques [11][12][13][14].

There are few studies about how MVs affect the model generation and prediction in Soft Computing techniques. Most of them focus on neural networks [15][16] or genetic algorithms [17] but, to the best of our knowledge, not in FIR. While FIR displays high prediction accuracy in complex systems, it has several limitations when missing data are present in the forecasting process. In [18] we introduced the concept of flexible FIR that can cope with missing information in the input pattern as well as in situations where patterns are not found in the behaviour matrix. However, that study is limited in the analysis of output forecast strategies and the evolution of flexible FIR prediction accuracy against the number of MVs in the dataset.

The contributions of this article can be summarized as follows. We perform a detailed analysis of an improved version of FIR, which can cope with missing information in the input values, as well as during the prediction phase. Eight different FIR forecasting strategies are proposed and studied taking into account several features to perform the prediction: causal relevance of input variables, consistency of predictions, inertia criterion and K-Nearest Neighbours. Furthermore, we evaluate the implications in prediction accuracy and number of instances[1] predicted, when the number of MVs in the training dataset is increased progressively. To this end, a real smart grid forecasting application, i.e. electricity load forecasting, has been chosen in this study.

The reason behind these experiments is to select the most appropriate *Output Forecast* strategy, which is not highly affected by the dispersion of the output classes and preserves confidence in prediction. In concrete, we suggest 8 different strategies that take into account different parameters and features for this calculation.

The paper is structure as follows: in section 2 the standard FIR methodology is summarized and the problem of MVs in this methodology is identified. Then, section 3 analyses a solution to cope with the missing data problem in FIR, called flexible FIR Prediction, and proposes 8 different output forecasting strategies. Next, section 4 presents the datasets used, which come from eight different buildings belonging to the UPC (Technical University of Catalonia), the experiments and the discussion of the results encountered when applying standard and flexible FIR Prediction to the datasets. Finally, section 5 points out the conclusions of our research and the near future work.

2. Fuzzy Inductive Reasoning

2.1 Standard FIR

The conceptualization of the FIR methodology arises of the General System Problem Solving (GSPS) approach proposed by Klir [19]. This methodology of modelling and simulation has the ability to describe systems that cannot be easily described by classical mathematics or statistics, i.e. systems for which the underlying physical laws are not well understood [9]. A FIR model is a qualitative non-parametric model based on fuzzy logic. The FIR model consists of its structure (relevant variables or selected features) and a pattern rule base (a set of input/output relations or history behaviour) that are defined as if-then rules.

The FIR methodology is a modelling tool that is able to infer the model of the system under study very quickly and it is a good option for real time forecasting. Moreover, it is able to deal with missing data as it has been already proved in a large number of applications [13]. However, its capacity to deal with

[1] Values corresponding to an input and/or output variable before fuzzification

missing data decreases significantly when the complexity of the mask is big, because it implies the generation of a big number of pattern rules in the behaviour matrix containing MVs. In Section 3 we deal with this issue.

Before starting the process of finding a FIR model in order to make predictions from the data, it is necessary to fuzzify the data in order to reduce the search space and speed up the optimization process [9]. FIR embraces a slightly different approach than other fuzzy techniques to solve the uniqueness problem. Rather than mapping into multiple fuzzy rules, FIR only maps into a single rule with the largest likelihood. To this end, FIR converts each quantitative value into a qualitative fuzzy triple, i.e. the *class*, the *membership* and the *side* values. The class value represents a discretization of the original real-valued variable. The fuzzy membership value denotes the level of confidence, expressed in the class value chosen to represent a particular quantitative value. The side value indicates whether the data point is to the left or the right of the peak of the corresponding fuzzy membership value. Thus, a single quantitative value is recoded into a qualitative triple which contains exactly the same information as the original quantitative value and it is thus possible to regenerate the quantitative value precisely. For a deeper insight into the fuzzification process refer to [9].

The process of obtaining a FIR model structure corresponds to a Feature Selection Process (FSP). The model structure holds the relevant features and it is represented by a mask through which the causal relations (both spatial and temporal) between input and output variables x are described. Table 1 presents an example of mask for a system with four inputs (u_1, u_2, u_3, u_4) and one output (y) variables.

Table 1. Example of mask for a system with four inputs (u_1, u_2, u_3, u_4) and one output (y) variables

x t	u_1	u_2	u_3	u_4	y
$t - 3\delta t$	-1	0	0	0	-2
$t - 2\delta t$	0	-3	0	0	0
$t - \delta t$	0	0	0	-4	0
t	0	0	0	0	+1

Each negative element in the mask is called mask input (*m*-input). It exhibits a causal relation with the output, i.e. it influences the output up to a certain degree. The single positive value denotes the mask output (*m*-output). In the example of Table 1, the prediction of the output at the current time, $y(t)$, is directly related to the variables u_1, u_2, u_4 and y in different times t, i.e. $u_1(t-3\delta t)$, $y(t-3\delta t)$, $u_2(t-2\delta t)$ and $u_4(t-\delta t)$.

The optimal mask function of FIR is used to obtain the best mask, i.e. the best FIR structure, for the system under study [9]. The procedure consists in finding the mask that best represents the system by computing a quality measure for all possible masks, and selecting the one with the highest quality. The process starts with the definition of a so-called mask candidate matrix encoding an ensemble of all possible masks from which the best is to be chosen. Table 2 shows and example of mask candidate matrix for the same system example of Table 1.

The mask candidate matrix contains elements of value -1, where the mask has potential causal relations. Elements of value +1, where the mask has its output can also be found. Finally, elements of value 0 denote forbidden connections.

Table 2. Example of mask candidate matrix for a system with four inputs (u_1, u_2, u_3, u_4) and one output (y)

x t	u_1	u_2	u_3	u_4	y
$t - n\delta t$	-1	-1	-1	-1	-1
…	…	…	…	…	…
$t - 2\delta t$	-1	-1	-1	-1	-1
$t - \delta t$	-1	-1	-1	-1	-1
t	-1	-1	-1	-1	+1

The number of rows of the mask candidate matrix is called the depth of the mask. It represents the temporal domain that can influence the output. Each row is delayed relative to its successor by a time interval of δt representing the time lapse between two consecutive samplings. δt may vary from one application to another. In the study presented in this paper, a value of δt equal to 1 hour is used, due to the data characteristics.

The optimal mask function of FIR offers the possibility to specify an upper limit to the acceptable mask's *complexity*, i.e. the largest number of non-zero elements that the mask may contain. Starting from the candidate matrix with minimum complexity two, i.e. 1 input and the output, the qualitative model identification process looks for the best out of the legal masks. Then it is proceed, by searching through all legal masks of complexity three, i.e. all masks with two inputs and an output, and find the best of these. It continues in the same way until the maximum complexity has been reached. This strategy corresponds to an exhaustive search of exponential complexity. However, suboptimal search strategies' of polynomial complexity can also be used, i.e. genetic algorithms [20].

Each possible mask is compared to the others with respect to its potential merit. The optimality of the mask is evaluated with respect to the maximization of its forecasting power that is quantified by means of the quality measure. Let us focus on the computation of the quality of a specific mask. The overall quality of a mask, Q_m, is defined as the product of its uncertainty reduction measure, H_r and its observation ratio, O_r, as described in equation 1.

$$Q_m = H_r \,.\, O_r \qquad (1)$$

The uncertainty reduction measure is defined in equation 2.

$$H_r = 1 - {}^{H_m}\!/_{H_{max}} \qquad (2)$$

Where H_m is the overall entropy of the mask and H_{max} the highest possible entropy. H_r is a real number in the range between 0.0 and 1.0, where higher values usually indicate lower uncertainty and therefore an improved forecasting. The highest possible entropy H_{max} is obtained when all probabilities are equal. Zero entropy is encountered for totally deterministic relationships. The overall entropy of the mask is then computed as described in equation 3.

$$H_m = -\sum_{\forall i} p(i)\,.\,H_i \qquad (3)$$

Where $p(i)$ is the probability of that input state to occur and H_i is the Shannon entropy relative to the i^{th} input state. The Shannon entropy relative to the i^{th} input state is calculated from equation 4.

$$H_i = \sum_{\forall o} p(o|i).\, log_2 p(\,o|i) \qquad (4)$$

Where $p(o|i)$ is the 'conditional probability' of a certain output state o to occur, given that the input state i has already occurred. The term probability is meant in a statistical rather than in a true probabilistic sense. It denotes the quotient of the observed frequency of a particular state in the episodically behaviour divided by the highest possible frequency of that state. The observation ratio, O_r, measures the number of observations for each input state. From a statistical point of view, according to [21], every state should be observed at least five times. If every legal input state has been observed at least five times, O_r is equal to 1.0. If no input state has been observed at all (no data are available), O_r is equal to 0.0. The optimal mask is the mask with the largest Q_m value, being the one that generates forecasts with the smallest amount of uncertainty, and, therefore, the features that compose the structure of this model are the ones selected as the most relevant ones.

Once the most relevant features are identified, they can be used in any modelling methodology. In FIR the mask is used to obtain the pattern rule (called behaviour matrix) from the fuzzified training dataset. Each pattern rule is obtained by reading out the class values through the 'holes' of the mask (the places where the mask does not have zero values), and it places each class next to each other to compose the rule. The class values are the mapped continuous variables into categorical.

Once the behaviour matrix and the mask are available, a prediction of future output states of the system can take place using the FIR inference engine, as described in Figure 1. This process is called qualitative simulation. The FIR inference engine is based on the K-Nearest Neighbour (KNN) rule, commonly used in the pattern recognition field. The forecast (of the output variable) is obtained by means of composition of the potential conclusion that results from firing the *k* rules, whose antecedents have best matching with the current input state.

As can be seen in the left hand side of Figure 1, the mask is placed on top of the qualitative data matrix (fuzzified test set), in such a way that the output matches the first element to be predicted. The values of the inputs are read out from the mask. And the behaviour matrix (pattern rule base) is used, as it is explained later, to determine the future value of the output, which can then be copied back into the qualitative data matrix. The mask is then shifted one position down to predict the next output value. This process is repeated until all the desired values have been forecast.

The fuzzy forecasting process works as follows: the input pattern of the new input state is compared with those of all previous recordings of the same input state contained in the behaviour matrix. For this purpose, a normalization function is computed for every element of the new input state and an Euclidean distance formula is used to select the KNN. They are the ones with smallest distance, which are used to forecast the new output state. The contribution of each neighbour to the prediction of the new output state's estimation is a function of its proximity. This is expressed by giving a distance weight to each neighbour, as shown in Figure 1. The new output state values can be computed as a weighted sum of the output states of the previously observed *k* nearest neighbours. In section 3.1, the standard and new fuzzy forecasting strategies are explained in detail.

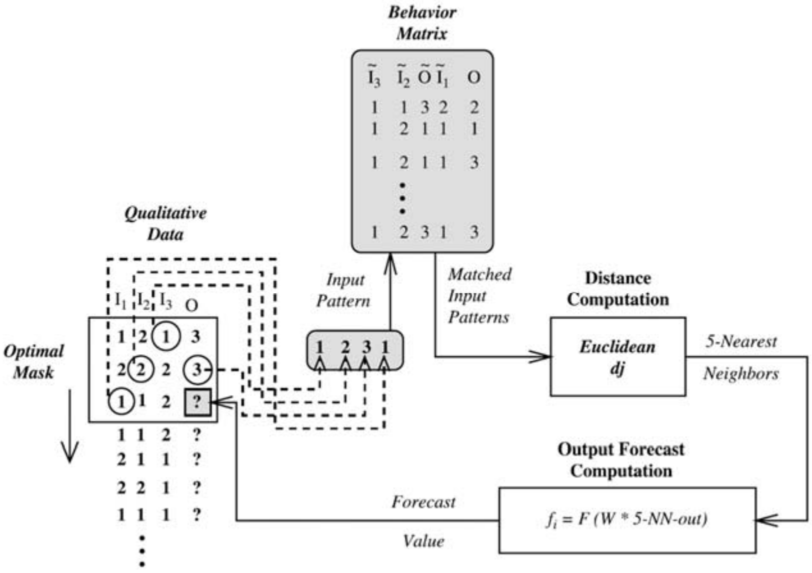


Figure 1. Qualitative prediction process diagram (with an example containing three inputs and one output) of Standard FIR

2.2 *FIR model with missing data*

As explained in detail in the previous section, a model in FIR methodology is composed of the mask (model structure) and a set of pattern rules, called behaviour matrix. The mask defines the causal and temporal relations between the inputs and output variables, i.e. it contains the variables selected as relevant. Once the best mask has been identified, it can be applied to the qualitative data obtained from the system resulting in a particular pattern rule base. This is a set of rules that represents pseudo-static relationships and that contains the system's behaviour.

The FIR process is illustrated in Figure 2, for one of the subsets with MVs. The left hand side of this figure shows an excerpt of the qualitative data matrix that stores the class values. In this example a class value of *-9* indicates a missing value. The dashed box symbolizes the mask that is shifted downwards along the class value matrix. The circled values in the mask denote the positions of the relevant selected features, whereas the square indicates the position of the output. The class values are read out from the class value matrix through the circled values of the mask, and are placed next to each other in the behaviour matrix that is shown on the right hand side of the figure. Here, each row represents one position of the mask along the class value matrix. Each row of the behaviour matrix represents one pseudo-static qualitative state or a pattern rule.

From the example illustrated in Figure 2, it can be seen that for a qualitative data matrix of *358* registers[2] containing *24* consecutive MVs, and a mask depth of *168,* it is possible to generate up to *191* pattern rules that contain at least one missing element. This can become a huge problem due to the fact that the current prediction process of FIR methodology discards the pattern rules containing MVs, and, therefore, the valid pattern rule base available is reduced significantly. This implies that the FIR prediction process is barely able to predict a new input pattern due to the fact that it does not exist in the behaviour matrix.

Due to the fact that the FIR prediction process discards those pattern rules that contain one or more missing elements, it performs the inference using only the set of pattern rules that are complete, i.e. do not contain any MV. In consequence, a lot of information is lost.

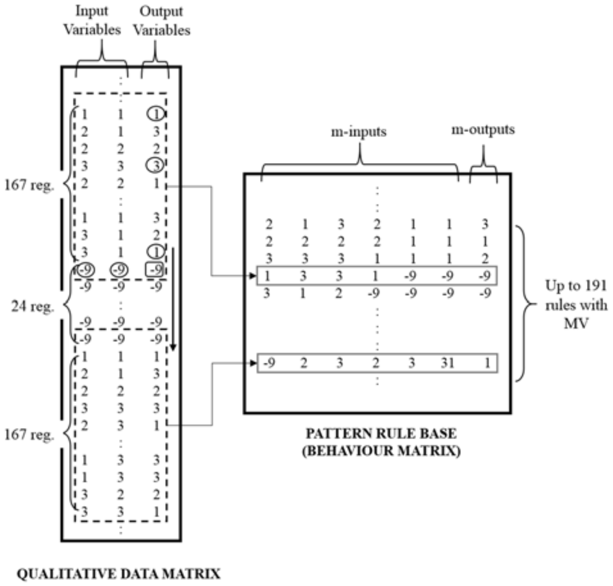


Figure 2. FIR process that generates the pattern rule base starting from the qualitative training dataset and the optimal mask. Example that contains 24 consecutive MV registers (reg.) in the input data. The figure illustrates the high proportion of generated pattern rules (191) that contains MVs.

In addition to this problem, there is a second weakness of FIR when it faces MVs, in particular when the input pattern contains MV. It may happen, especially with online predictions, where input pattern generated may contain MVs: failure in the communication between smart meter and concentrator, the battery of a home area sensor is depleted, loss of internet connection, etc. As it is shown in Figure 3, in the standard version of FIR, the input pattern is searched in the behaviour matrix. However, it cannot be found because, as previously explained, i) the behaviour matrix discards pattern rules containing missing elements and ii) the input pattern could represent several states because it contains a missing value. So, the *m*-inout could belong to any class of the fuzzified variable.

[2] Fuzzified instance

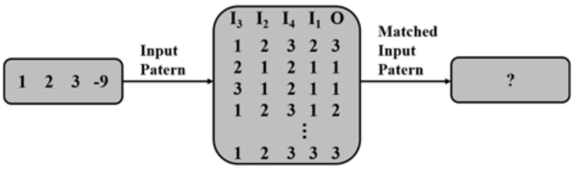


Figure 3. Input pattern containing MV (-9 value), which is unable to find a matched pattern in the behaviour matrix and therefore unable to predict

3. Flexible FIR Prediction

The enhancement proposed in this paper is to develop an algorithm that makes the inference process be flexible in a dynamic way. The idea is to use the traditional algorithm when there exist in the behaviour matrix rules that have the same input pattern (free of MVs) as the one to be predicted. When this is not the case, the algorithm will select the set of pattern rules that have the same input pattern, but relaxing one of its m-inputs. That is, the same input pattern, but allowing one of the m-inputs to be missing. Basically, this means that we are using a different mask than the one selected by FIR in the modelling process. Therefore, we are ignoring one of the variables selected as relevant in the feature selection process. It is possible that the “new” mask is suboptimal, but very close to the quality of the optimal mask. However, it can also happen that the “new” mask is a bad one.

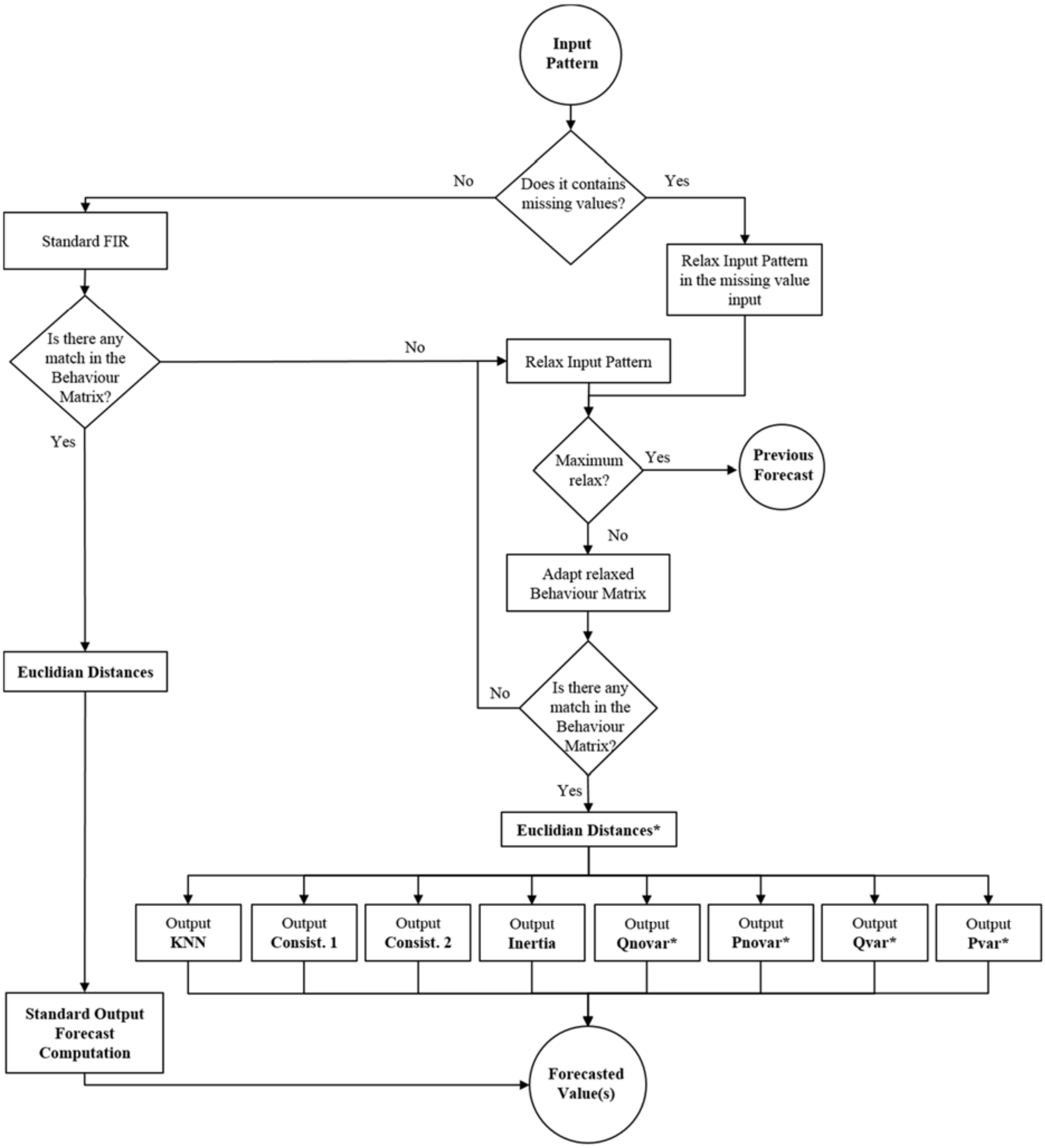
The new FIR flexible prediction algorithm is depicted in Figure 4. After the training of the model, when a new (online) class input pattern arrives, it is checked whether it contains MVs. If not, it continues with the standard FIR prediction. If it contains MVs, the new algorithm relaxes the input pattern, particularly, the positions of MVs.

At this point, a Maximum Relax Parameter (MRP) is required. The MRP specifies the maximum number of relaxed parameters of the input pattern. The greater the number of m-input patterns relaxed, the larger the search space. And, therefore, more noise is added to the final prediction. The total number of missing elements allowed in the pattern rules to perform the inference process is determined from the size of the input pattern:

$$MRP = (\textit{Size Input Pattern}) / 2 \qquad (5)$$

Considering the example of Figure 3, the *Size Input Pattern* is *4* and therefore the MRP is *2*, which means that the input pattern will be relaxed up to 2 m-inputs. If the maximum relaxation is reached and there is none match in the behaviour matrix, the forecast value assigned is the same as the previous forecasted value. In hourly energy forecasting problems this solution is usually more convenient than not predicting, because current and previous hourly consumptions are highly correlated.

In the *MRP* formula, we consider that dividing *Size Input Pattern* by more than *2* the relaxed input pattern may differ too much from the original one because the information lost is too high.



** Output forecast computations that involve Qvar, Qnovar, Pvar and Pnovar also imply different computation of Euclidian distance. More details in Section 3.1*

Figure 4. Flow diagram of the FIR flexible prediction algorithm

If the MRP is not reached, the behaviour matrix is relaxed to include all the pattern rules without the m-input parameter relaxed. As Figure 5 shows (top), the input pattern remains the same but a missing m-input is allowed. A different mask from the one selected by FIR in the modelling process is used, ignoring, therefore, one of the variables selected as relevant in the feature selection process. Similarly, when the input pattern is not found in the behaviour matrix, *n* Adapted Behaviour Matrices are generated, *n* being all possible combinations of relaxation (bottom of Figure 5 with *n*=4). Afterwards, it is verified if there are any matches of the new relaxed input pattern in the relaxed behaviour matrices.

If there is no match of the new relaxed input pattern in any of the *n* adapted behaviour matrices, the algorithm will look for patterns with two missing elements. The total number of missing elements allowed in the pattern rules to perform the inference process will be determined in function of the size of the input pattern and the quality of the suboptimal masks associated to the input patterns.

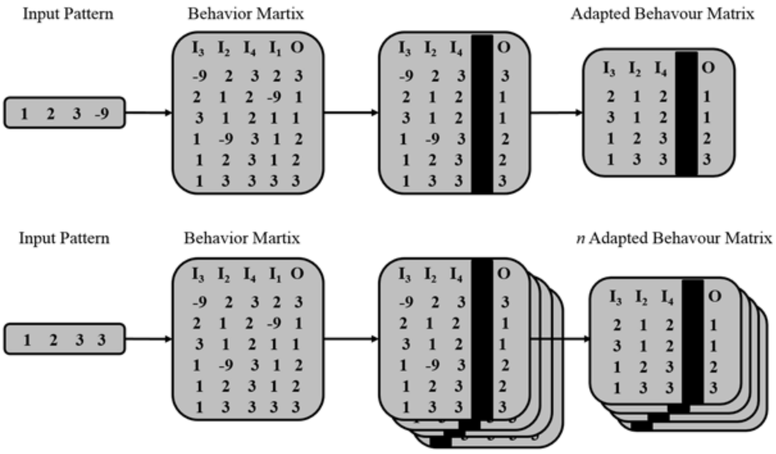


Figure 5. Adaptation of the behaviour matrix when there are MVs in the input pattern (top) or the input pattern is not found in the original behaviour matrix (down)

3.1 Output Forecast Computation

As explained in section 2.1, the fuzzy forecasting compares the current input pattern with previous recordings that have the same input pattern in the behaviour matrix. Then an Euclidean distance formula is used to select the KNN that are used to forecast the new output state. The contribution of each neighbour to the estimation of the new output state prediction is a function of its proximity.

In the new approach, flexible FIR prediction may result in a dispersion of the output classes, due to the loss of information in the input pattern and in the behaviour matrix when they have been relaxed. This could result in a difficult prediction with low confidence. To cope with this possible problem, in this research we want to investigate different *Output Forecast* strategies (In Figure 4 they are represented beneath the computation of the Euclidean Distances). The reason behind these experiments is to select the most appropriate *Output Forecast* strategy, which is not highly affected by the dispersion of the output classes and preserves confidence in prediction. In concrete, we suggest 8 different strategies that take into account different parameters and features for this calculation.

3.1.1 Classic FIR KNN (aKnn)

In this case, the output is computed using the Euclidean distance and the specific KNN algorithm of FIR.

In FIR the membership and side functions of the new input state are compared with those of all previous recordings of the same input state contained in the class behaviour matrix [9]. For this purpose, a normalization function is computed for every element of the new input state:

$$p_i = Side_i \cdot B \cdot \sqrt{\ln(Memb_i)} + 0.5 \qquad (6)$$

where $B = (4 \cdot \ln(0.5))^{-1/2}$, and *Memb*$_i$ and *Side*$_i$ are the membership and side values respectively of the input *i*.

For the extreme classes the functions are calculated in the following way. For the left class:

$$p_i = C \cdot \sqrt{\ln(Memb_i)} \qquad (7)$$

where $C = \ln(0.5))^{-1/2}$. For the right class:

$$p_i = 1 - C \cdot \sqrt{\ln(Memb_i)} \qquad (8)$$

Irrespective of whether an original signal is, the corresponding p_i signal ranges exactly from 0 to 1.

The p_i values are quantitative real valued variables that can be used to represent the relative magnitude of a particular qualitative triple. However they are not regenerations of the original quantitative signals. They are normalized variables. Consequently different p_i signals can be compared to each other or can be summed up without weighing them relative to each other. Something that would not be meaningful using the original or regenerated signals. The normalization function is a transformation from a qualitative triple to a quantitative variable but this variable lives in a different space from the original quantitative variable.

The p_i values corresponding to the different variables of an input state are then concatenated to form the vector:

$$p = [p_1, p_1, \dots p_t] \qquad (9)$$

assuming the state contains t inputs (in case of the example, $t = 3$ because one variable is missing). We call the vector p the *norm image* of the original input state.

The analysis proceeds by computing norm images for every previous recording of the same input state. Let us call these vectors p_k. Every p_k vector is a little different, since only the class values of the recorded input states are identical, but not their membership or side function values.

Finally the norm of the differences between the p vector representing the new norm image and the p_k vectors representing all previous recordings of the same input state are computed:

$$d_j = \sqrt{\sum_{i=1}^{N} (p_i - p_{ik})^2} \qquad (10)$$

and the k previous recording with the smallest d_j are identified (if at least k such recordings are found in the class behaviour matrix). The KNN are used to forecast the new output state.

The contribution of each neighbour to the estimation of the prediction of the new output state is a function of its proximity. The new output state values can be computed as a weighted sum of the output states of the previously observed k nearest neighbours.

Notice that this strategy corresponds to the classic KNN of standard FIR but using an adapted Behaviour Matrix.

3.1.2 Use of the causal relevance

In this strategy the causal relevance (CR) is incorporated in the calculation of the Euclidean distance [22]. The idea behind this strategy is simple and can be addressed through the following questions: How much does each m-input influence the prediction of the output? Is it possible to compensate the loss of causality of the m-input I_2? If it is possible to quantify the importance of each m-input with respect to the output, then it is possible to assign a higher weight to those m-inputs with a higher importance.

The *Output Forecast* is based in the KNN algorithm of FIR, as in the previous case.

Similarly to the previous strategy, d_j is computed for each state, but adding a new distance measure that takes the CR into account. Therefore, the CR is used as a weight in the modified Euclidean distance formula, as presented in equation 11:

$$d_j = \sqrt{\sum_{i=1}^{N} ((R_{dis_i}) * (p_i - p_{ik})^2)} \qquad (11)$$

Where R_{disi} is the quantified CR of the i_{th} input. Using the weight-modified distance formula of equation 2, the distances of the most relevant inputs exert a stronger influence on the overall distance of this specific input pattern, whereas the influence of the less relevant inputs is reduced with respect to the classical Euclidean distance formula, which has been used previously.

3.1.2.1 Causal Relevance Qnovar (bQnv)

In [22] several formulae have been proposed, studied and used for the prediction of different types of applications, i.e. medical, biological and linear systems. The best weight equation that behaves consistently well in all the aforementioned applications is:

$$R_{dis_i} = 1 - Q_{novar_i} \qquad (12)$$

The CR Q_{novari} is based on the quality of the mask Q_m. Q_{novari} is the quality of the mask that is identical to the optimal mask except for excluding the i_{th} m-input from the set of mask inputs. Q_{novari} can be interpreted as an indirect causal correlation between the specific m-input under study and the output to be predicted. It quantifies the amount of information that will be lost when the i_{th} m-input is eliminated from the model that will be used to predict the system.

3.1.2.2 Causal Relevance Qvar (bQv)

Similarly, as described in [22], the best weight equation found for the CR Q_{var} is:

$$R_{dis_i} = Q_{var_i} \tag{13}$$

Q_{vari} is the quality of the mask, of complexity two, that contains only the m-output and the i_{th} m-input. Q_{vari} can be interpreted as a direct causal correlation between the i_{th} m-input and the output to be predicted.

3.1.2.3 Causal Relevance Pnovar (bPnv)

In this strategy the CR is not based on the quality of the mask but on the Mean Squared Error (MSE) prediction:

$$R_{dis_i} = 1 - P_{novar_i} \tag{14}$$

P_{novari} is the normalized MSE obtained in a validation dataset utilizing the mask that is identical with the optimal mask except for excluding the i_{th} m-input [23][24].

$$P_{novar_i} = \overline{MSE_{M(opt-\imath)}} = \frac{MSE_{M(opt-i)}}{\sum_{j=1}^{n} MSE_{M(opt-j)}} \tag{15}$$

where n is the number of m-inputs of the optimal mask and $M_{(opt\text{-}j)}$ is the optimal mask excluding the i_{th} m-input.

3.1.2.4 Causal Relevance Pvar (bPv)

Similar to 3.1.2.3

$$R_{dis_i} = P_{var_i} \tag{16}$$

The P_{var} is the normalized error (inverse) obtained in a validation dataset utilizing the mask of complexity two that contains only the m-output and the i_{th} m-input [23][24].

$$P_{var_i} = \overline{MSE_{M\imath}} = \left(\frac{\sum_{j=1}^{n} MSE_{M(j)}}{MSE_{M(i)}}\right) \Big/ \sum_{\mathrm{k}=1}^{n} \left(\frac{\sum_{j=1}^{n} MSE_{M(j)}}{MSE_{M(i)}}\right)_k \tag{17}$$

where n is the number of m-inputs of the optimal mask and $M_{(j)}$ is the mask that contains only the j_{th} m-input and the m-output.

3.1.3 Inconsistency in Forecast

The dispersion of the output classes of the neighbours obtained in the behaviour matrix indicates a difficult prediction with low confidence. An indirect measure of the matched input patterns' confidence is the variance of the outputs O_I ($\mathrm{O}_{\mathrm{variance}}$):

$$O_{variance} = variance\ (O_I) \tag{18}$$

The CR for the calculation of Euclidean distance could not help enough in those cases, therefore, here we present three possible strategies to deal with high variances in O_I.

3.1.3.1 Consistency forecast 1 (cCf1)

If $O_{variance}$ is different than 0 and the first two neighbours have the same Euclidean distance but they belong to different classes, then take the KNN for the first and second classes. If $O_{variance}$ is 0, classic FIR KNN is used.

3.1.3.2 Consistency forecast 2 (cCf2)

If $O_{variance}$ is different than 0, then from all the instances that matches the input state, use the KNN of the most repeated output class. If $O_{variance}$ is 0, classic FIR KNN is used.

3.1.3.3 Inertia Criterion (cIn)

For any value of $O_{variance}$, use the previous value as the predicted value if a *confidence measure* is low.

In electricity load forecasting the previous and current value are usually highly correlated, thus the idea with this strategy is that if there is a high probability to perform a bad prediction due to the dispersion of the possible solutions, then use the previous value as the predicted value. To this end, a confidence measure is computed [25] and, if it is equal or below *0.5,* i.e. a confidence below or equal to 50%, the previous value is used as the predicted value. If not, *aKnn* is used.

In order to compute the confidence of a prediction the position value, Pos_o, of the *m*-output is computed as follows:

$$Pos_o = Class_o + Side_o \cdot (1 - Memb_o) \qquad (19)$$

where $Class_o$, $Side_o$ and $Memb_o$ is the qualitative triple representing the *m*-ouput. Then, for each of the KNNs the position value of the outputs is computed. The maximum, Pos_{max}, and minimum, Pos_{min}, position values are then used to compute the *confidence measure*:

$$confidence\ measure = 1 - (Pos_{max} - Pos_{min}) \qquad (20)$$

4. Experiments and Results

4.1 Dataset

Data of 8 buildings of the Universitat Politècnica de Catalunya (UPC) was obtained for this study, in order to have a training/test sample with high diversity of consumptions. They have different profiles of usage (sports centre, library, administration building, restaurant...), belong to five different campuses and are located in different cities. Thus, affecting different climatology (temperature, humidity, solar radiation, etc.), consumption patterns, schedules and working days. The buildings included are: 1) the Library of the ETSEIAT[3] faculty in Terrassa; 2) one campus building also in ETSEIAT faculty; 3) the Library of EPSEM[4] faculty in Manresa; 4) the Library of EPSEVG[5] faculty in Vilanova 5) the bar of the EETAC[6] faculty in Castelldefels; 6) and 7) two buildings with different classrooms and labs at FIB[7] faculty in Barcelona; and 8) the sports centre in Campus Nord also in Barcelona. The energy consumptions of these 8 buildings have been collected through a remote metering system every hour. Therefore, there are 24 recordings per day and per location.

For all 8 buildings, the dataset comprises a whole year of electricity consumption, from 13/11/2013 to 12/11/2014, from which we separated 91% for training and the remaining 9% for testing. The testing data comprises 35 different days (i.e. 35 test sets) distributed equally through the whole year; meaning around 9 days per season and taking into account the seven days of the week (from Monday to Sunday). By choosing these days we pretend to evaluate the models against the changes caused by seasonal period(s) and day of the week.

Experiments

In our experiments we perform 24 hour predictions with 9 different FIR forecasting strategies; the standard FIR prediction and the aforementioned flexible FIR Prediction (8 output forecast strategies). On the one hand, the aim is to understand which output's forecast strategy can cope better with the dispersion of output forecast classes and low confidence in predictions (due to the presence of MVs in data). And on the other hand, to analyse the evolution of flexible FIR Prediction accuracy against the number of MVs in the dataset.

To add MVs in the experiments, we substitute a data point by an unaccepted value for the model. In our current implementation of FIR we have set the missing value to *-9*.

[3] ETSEIAT: School of Industrial and Aeronautic Engineering of Terrassa
[4] EPSEM: School of Engineering of Manresa
[5] EPSEVG: School of Engineering of Vilanova
[6] EETAC: School of Telecommunications and Aerospatiale of Castelldefels
[7] FIB: Barcelona School of Informatics

Our data does not contain MVs because we need the real consumptions to evaluate standard and flexible FIR. Instead, MVs are added artificially (more details in sections 4.1.1 and 4.1.2).

4.1.1 Experiment 1: Flexible FIR Strategy evaluation

In the first experiment (Figure 6) the aim is to understand how possible it is to perform predictions with partial information in the input pattern and how it is correlated with the number of input variables and depth of the mask. To do so, we have identified three branches with different number of input variables and depths. In the first branch we only take into account a basic variable in load forecasting; the output variable, i.e. historical consumptions. In concrete, the four most relevant electricity past consumption selected in the FIR FSP. The second branch includes the past values of the electricity consumptions (four variables), and a binary variable that specifies whether it is a working day. Therefore, the second experiment has five inputs. Finally, the third branch includes all five inputs of the previous experiment plus the hours of the day variable. The FSP has been applied to different depths: 24, 72 and 24 + 24. The reason why these depths are chosen is explained in section 4.1.3.

The 9% of data used for testing is substituted by MVs (only in *historical consumptions* output variable) during the training process of the model (see Table 3). With this modification in data, we are emulating failures in gathering the data.

Table 3. Number of MVs in *Experiment 1* and *Experiment 2*

EXPERIMENT 1

% of MVs in training dataset (only *historical consumption* output variable)	Number of MVs
9	6720

EXPERIMENT 2

% of MVs in training dataset (all input and output variables)	Number of MVs
9	20160
18	39168
27	58176
36	77184
45	96192
54	115200
63	134208
72	153216
81	172224

4.1.2 Experiment 2: Robustness and reliability of Flexible FIR

The second experiment consists of increasing the number of MVs in data progressively and evaluate the accuracy of the predictions. Figure 7, shows the scheme of the experiment, which is slightly different to the first experiment: the training test (without MVs) is used to compute the 4 more relevant past consumptions based on FIR FSP. Then *n*, the percentage of MVs in data, is initiated (an initial value is given) and we randomly substitute *n*% of the training data by MVs. Right after, the process of FIR model creation and FIR forecasting (standard and flexible) followed the scheme of the first experiment but always with six variables: the 4 most important past consumptions, whether it is working day and the hour of the day. Finally, the prediction error is computed based on the 35 test days. Then, *n* is increased, and again we randomly substitute *n*% of training data by MVs. The experiment is repeated until a maximum *n*%. Notice that FIR FSP is computed only once and when data do not contain MVs.

For the purpose of the experiments described in this paper, we assumed that data was missing at random, i.e. there was no correlations between the occurrence of missing values for different variables, and ignored the mechanism of missing data. .

Table 3 shows the number of MVs inserted in the training dataset. In this experiment, they are inserted to all input and output variables. Therefore, the number of MVs with *n* = 9% is three times more in the *experiment 2* than 1. The number the MVs are randomly selected and it increases around 9% in each round.

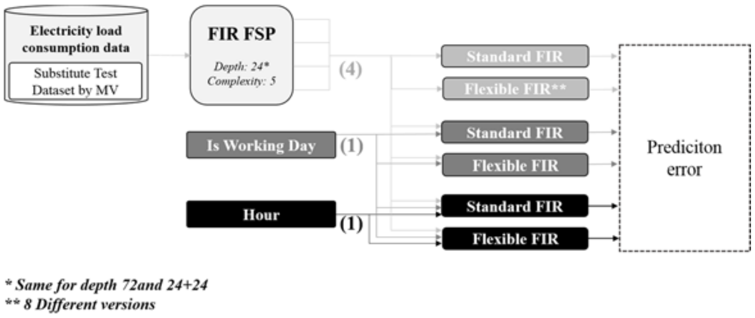


Figure 6. Scheme of Experiment 1 when depth is equal to 24. The same scheme applies to depths 72 and 24+24. The numbers in parentheses stand for the number of variables added in the model.

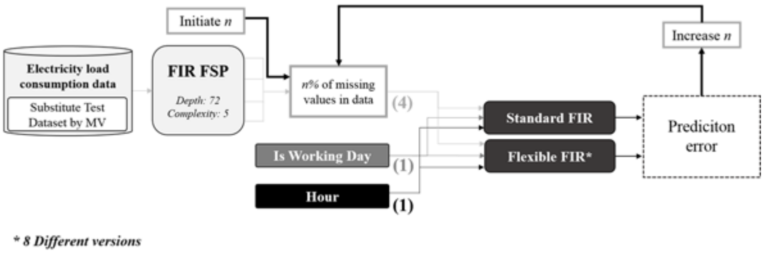


Figure 7. Scheme of Experiment 2 to evaluate implications of MVs in the training dataset with flexible FIR Prediction. The numbers in parentheses stand for the number of variables added in the model.

4.1.3 Model parameters

In this study the modelling process consists of: 1) feature selection process 2) use of the relevant features to derive a FIR model and 3) selection of the parameters that will allow a more flexible prediction of FIR in case of missing data. This process is repeated for each location and each depth studied. It is decided to build up one model that predicts electricity consumptions one day ahead for each season instead of 4 independent models, i.e. one per season. The main reason is that the data studied do not present a clear trend; hence no deseasonal pre-processing is applied because we want to study the capacity of the different methodologies to obtain generic models.

The FSP is applied only to the historical consumption data and not to the hourly and daily information. It is decided to follow this strategy because the hourly and daily information contains only the hour of the day and if it is or not a working day, respectively. Therefore, the valuable information is gained with the actual value not with the previous ones. However, previous consumptions contain information patterns from where important knowledge could be extracted.

The selection of the depth and number of variables is a crucial issue that can affect those methods that are more sensitive to the curse of dimensionality. When the FSP of FIR is used, when increasing the number of variables (i.e. complexity) and depth, the quality of the results also increases until the optimal values are reached. After that, increasing the number of variables may add noise to the system and end up with a lower-quality result. But it may also imply an unfulfillment of the observation ratio and end up with an impossibility of prediction. It has been empirically determined, for electrical load consumption applications, that more than four variables and depths higher than 72, do not increase significantly the quality of the FSP of FIR, while computational cost (in terms of time) does exponentially [18].

4.1.3.1 Model parameters Experiment 1

To capture the most relevant previous consumptions in the electric load series, different depths are studied: i) previous 24 hours, ii) previous 72 hours and iii) 24 + 24: previous 24 hours and the past 24 hours of the previous week (48 past values in total that corresponds to a depth of the mask of 168).

With regard to the complexity of the mask, the four past consumptions selected by FIR are taken into account in all the models developed in this work and the mask increases to five and six when the variables working day and hour are added.

Regarding the fuzzification parameters, three classes and the equal frequency partition algorithm have been used to discretize the electrical load consumption and hour of the day variable. Working day variable is binary and, therefore, it has been discretized into two classes.

Finally, in the *Output Forecast,* the parameter set for the KNN is *k*=5, because as we have seen in other experiments with same application [11][12][18] it has good results.

4.1.3.2 Model parameters Experiment 2

Here the focus is not on the best model configuration for predictions, but in the prediction error evolution increasing the number of MVs in the training dataset. So then, the depth and complexity selected are those with better results in *experiment 1*, which are *48* and *6* respectively (Figure 8 and Figure 9).

In the *Output Forecast, k* remains the same as *experiment 1.*

4.1.4 Flexible FIR Prediction parameters

As the standard FIR *Output Forecast*, the parameter set for the KNN in the new strategies will be also *k*=5. Apart from that, there are other few parameters to be determined in the new flexible FIR Prediction.

4.1.4.1 bPnv and bPv

As reviously pointed out in section 3.1, the P_{novar} and P_{var} are computed using a validation dataset. The validation dataset is taken from the training dataset (test dataset is not used to compute P_{novar} and P_{var}, they are independent procedures), which is a total 7896 data points; 80% is used for training and 20% for testing.

4.2 Evaluation Criteria

There are many measures of forecast's accuracy in the literature [26]. We require a statistical quality measure, which is able to compare the different forecasting methods in buildings with different average loads.

The Normalized Mean Squared Error (NMSE), described in equation 21, is used as the error measure to evaluate the forecasted results.

$$NMSE = \frac{1}{N}\sum_{t=1}^{N}\left[\left(y_r(t) - y_f(t)\right)\right]^2 / var\left(y_{training}(t)\right) \qquad (21)$$

where y_r and y_f are real and forecast electric consumptions, respectively, and $var(y_{training}(t))$ is the variance of the real electric consumptions used in the training data. N is the number of elements in the test dataset.

The Mean Squared Error is not suitable to evaluate the performance of the model when values in the datasets differ in magnitude. For example, values predicted in some buildings are in the order of 132 kWh, whereas, for instance, the Castelldefels Bar is in the order of 5 kWh. Hence, we decided to use the MSE divided by the variance of the training dataset, in order to standardize the data.

The NMSE is not the only criterion and there are some other commonly used evaluation criteria. The present research also considered to use the Mean Absolute Percentage Error (MAPE) to offer a forecasting performance from a multi-dimensional perspective. The reason to choose MAPE is that it can be used to compare the performance on different datasets, because it is a relative measure. The MAPE, is described in equation (22).

$$MAPE = 100 * \frac{1}{N}\sum_{t=1}^{N}\left|\left(y_r(t) - y_f(t)\right)/y_r(t)\right| \qquad (22)$$

However, MAPE puts a heavier penalty in negative errors than in positive errors. For example, if $y_r(1) = 4.20\ kWh$ and $y_f(1) = 6.03\ kWh$ (negative error) the MAPE is *43.57%.* Whereas if $y_r(1) = 6.03\ kWh$ and $y_f(1) = 4.20\ kWh$ (positive error) the MAPE is *30,34%.* In some applications it is interesting to penalize negative errors, because higher predictions than actual values can lead to an extra purchase of assets. In our study, we are only interested in the forecasting accuracy and it should penalize equally negative and positive errors.

This observation led to the use of the so-called "symmetric" Mean Absolute Percentage Error (sMAPE) [27] defined by

$$sMAPE = 200 * \frac{1}{N}\sum_{t=1}^{N} \left|(y_r(t) - y_f(t))\right| / (y_r(t) + y_f(t)) \qquad (23)$$

In the example above, with sMAPE, the error obtained with both negative and positive error is *35.77%*. Therefore, sMAPE has been chosen as the second measure of forecast's accuracy.

In addition, it has to be highlighted that measures based on percentage errors have the disadvantage of being undefined when *y(t) = 0* for any *t* in the period of interest, and having an extremely skewed distribution for values of *y(t)* very close to zero. In our experiments, it has been verified that none of the consumption data points *y(t)* is equal or very close to zero.

5. Results

5.1 Experiments

Experiment 1

Table 1 shows the results obtained by standard FIR and each of the eight flexible FIR strategies for the three different depths studied, and for the three experiments with different number of input variables. The results obtained are an average of the error prediction in the 35 test days considering the 8 buildings of study.

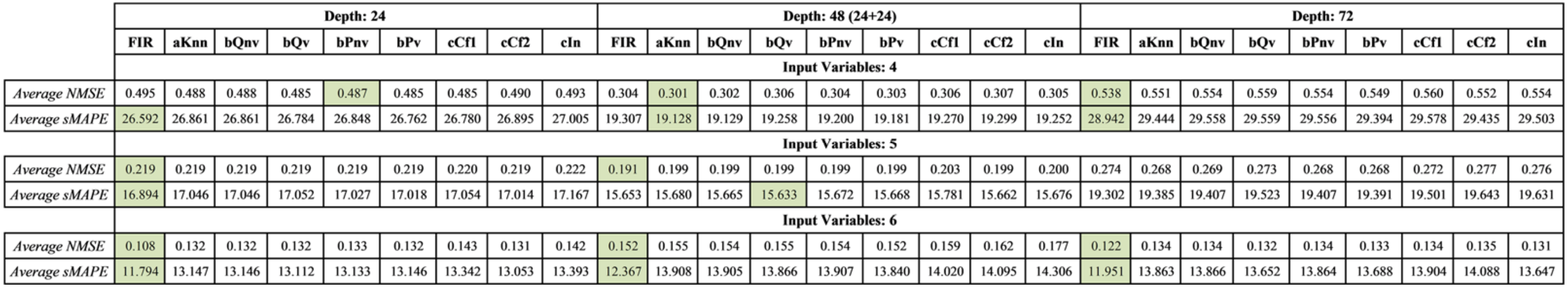
The lowest error in each experiment is highlighted. The average prediction error of the standard FIR and our new implementation are equivalent and sometimes the error is lower with flexible FIR Prediction. These are promising results because although the new version of FIR performs predictions with MVs, both in input pattern and in behaviour matrix, the NMSE and sMAPE average is around 0.108 and 11% respectively. It is important to emphasize that the average error performed with the standard FIR does not contain those registers with the missing value problem because it cannot predict them, which means it only takes into account part of the predictions to compute NMSE and sMAPE.

The main reason why high error values appear in some cells of Table 4 is because the table shows the average of 35 test days. If in those cases the prediction of a single day is quite bad then the average error is considerably increased.

Table 5 shows the results obtained only when standard FIR was not able to perform a prediction, i.e. when there were MVs in the input pattern and/or the input pattern was not found in the behaviour matrix. When it is analysed, it can be seen that the NMSE and sMAPE averages are, in the best cases, around 0.116 and 11% respectively, which is similar to the average error prediction of Table 4, although missing information is present.

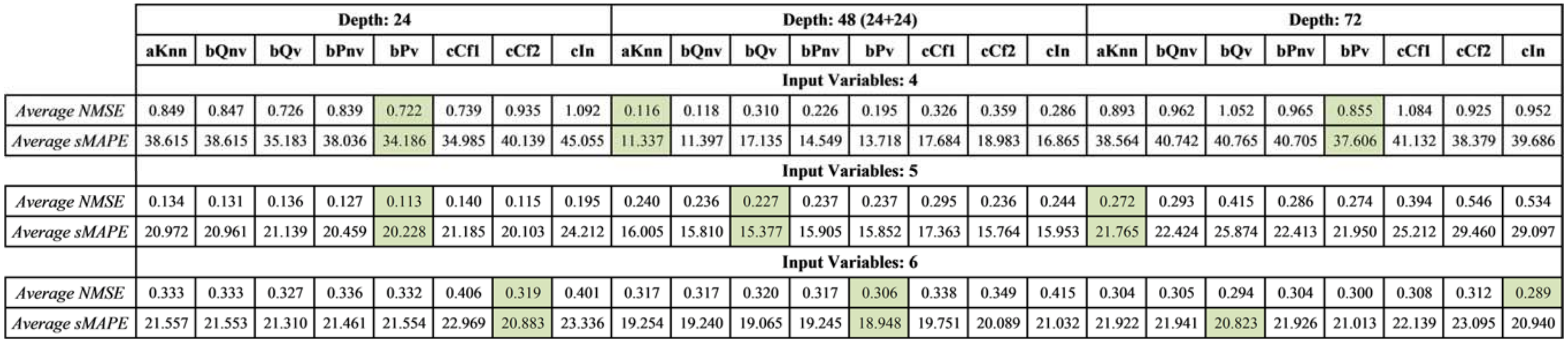
Figure 8 and Figure 9 show the error evolution with sMAPE in different FIR model configurations of standard FIR and flexible FIR (8 strategies). The average is also computed considering the prediction results in the 8 buildings of this study. In Figure 8 the prediction errors (of 4, 5 and 6 input variable) are considered to compute the average and we differentiate by the depth of the mask. Figure 9 shows the average of all prediction errors (with depth's mask 24, 48 and 72) differentiated by the number of input variables in the FIR models.

Table 4. NMSE and sMAPE Average (%) of the 35 test days representing an entire season year, in 8 different buildings, obtained by means of standard FIR Prediction and flexible FIR Prediction (F1-2-3: Output Forecast 1,2,3)



	Depth: 24									Depth: 48 (24+24)									Depth: 72								
	FIR	**aKnn**	**bQnv**	**bQv**	**bPnv**	**bPv**	**cCf1**	**cCf2**	**cIn**	**FIR**	**aKnn**	**bQnv**	**bQv**	**bPnv**	**bPv**	**cCf1**	**cCf2**	**cIn**	**FIR**	**aKnn**	**bQnv**	**bQv**	**bPnv**	**bPv**	**cCf1**	**cCf2**	**cIn**
	Input Variables: 4																										
Average NMSE	0.495	0.488	0.488	0.485	0.487	0.485	0.485	0.490	0.493	0.304	0.301	0.302	0.306	0.304	0.303	0.306	0.307	0.305	0.538	0.551	0.554	0.559	0.554	0.549	0.560	0.552	0.554
Average sMAPE	26.592	26.861	26.861	26.784	26.848	26.762	26.780	26.895	27.005	19.307	19.128	19.129	19.258	19.200	19.181	19.270	19.299	19.252	28.942	29.444	29.558	29.559	29.556	29.394	29.578	29.435	29.503
	Input Variables: 5																										
Average NMSE	0.219	0.219	0.219	0.219	0.219	0.219	0.220	0.219	0.222	0.191	0.199	0.199	0.199	0.199	0.199	0.203	0.199	0.200	0.274	0.268	0.269	0.273	0.268	0.268	0.272	0.277	0.276
Average sMAPE	16.894	17.046	17.046	17.052	17.027	17.018	17.054	17.014	17.167	15.653	15.680	15.665	15.633	15.672	15.668	15.781	15.662	15.676	19.302	19.385	19.407	19.523	19.407	19.391	19.501	19.643	19.631
	Input Variables: 6																										
Average NMSE	0.108	0.132	0.132	0.132	0.133	0.132	0.143	0.131	0.142	0.152	0.155	0.154	0.155	0.154	0.152	0.159	0.162	0.177	0.122	0.134	0.134	0.132	0.134	0.133	0.134	0.135	0.131
Average sMAPE	11.794	13.147	13.146	13.112	13.133	13.146	13.342	13.053	13.393	12.367	13.908	13.905	13.866	13.907	13.840	14.020	14.095	14.306	11.951	13.863	13.866	13.652	13.864	13.688	13.904	14.088	13.647

Table 5.NMSE and sMAPE Average (%) of the days where standard FIR Prediction could not predict due to MVs



	Depth: 24								Depth: 48 (24+24)								Depth: 72							
	aKnn	**bQnv**	**bQv**	**bPnv**	**bPv**	**cCf1**	**cCf2**	**cIn**	**aKnn**	**bQnv**	**bQv**	**bPnv**	**bPv**	**cCf1**	**cCf2**	**cIn**	**aKnn**	**bQnv**	**bQv**	**bPnv**	**bPv**	**cCf1**	**cCf2**	**cIn**
	Input Variables: 4																							
Average NMSE	0.849	0.847	0.726	0.839	0.722	0.739	0.935	1.092	0.116	0.118	0.310	0.226	0.195	0.326	0.359	0.286	0.893	0.962	1.052	0.965	0.855	1.084	0.925	0.952
Average sMAPE	38.615	38.615	35.183	38.036	34.186	34.985	40.139	45.055	11.337	11.397	17.135	14.549	13.718	17.684	18.983	16.865	38.564	40.742	40.765	40.705	37.606	41.132	38.379	39.686
	Input Variables: 5																							
Average NMSE	0.134	0.131	0.136	0.127	0.113	0.140	0.115	0.195	0.240	0.236	0.227	0.237	0.237	0.295	0.236	0.244	0.272	0.293	0.415	0.286	0.274	0.394	0.546	0.534
Average sMAPE	20.972	20.961	21.139	20.459	20.228	21.185	20.103	24.212	16.005	15.810	15.377	15.905	15.852	17.363	15.764	15.953	21.765	22.424	25.874	22.413	21.950	25.212	29.460	29.097
	Input Variables: 6																							
Average NMSE	0.333	0.333	0.327	0.336	0.332	0.406	0.319	0.401	0.317	0.317	0.320	0.317	0.306	0.338	0.349	0.415	0.304	0.305	0.294	0.304	0.300	0.308	0.312	0.289
Average sMAPE	21.557	21.553	21.310	21.461	21.554	22.969	20.883	23.336	19.254	19.240	19.065	19.245	18.948	19.751	20.089	21.032	21.922	21.941	20.823	21.926	21.013	22.139	23.095	20.940

Table 6. Number of registers (reg.) predicted by standard FIR Prediction and flexible FIR Prediction

	Depth: 24		**Depth: 72**		**Depth: 48 (24+24)**	
	FIR	**Flex FIR**	**FIR**	**Flex FIR**	**FIR**	**Flex FIR**
	Input Variables: 4					
reg.	6613	6720	6597	6720	6530	6720
%	**98.41**	**100**	**98.17**	**100**	**97.17**	**100**
	Input Variables: 5					
reg.	6521	6720	6438	6720	6574	6720
%	**97.04**	**100**	**95.80**	**100**	**97.83**	**100**
	Input Variables: 6					
reg.	5984	6720	5719	6720	5861	6720
%	**89.05**	**100**	**85.10**	**100**	**87.22**	**100**

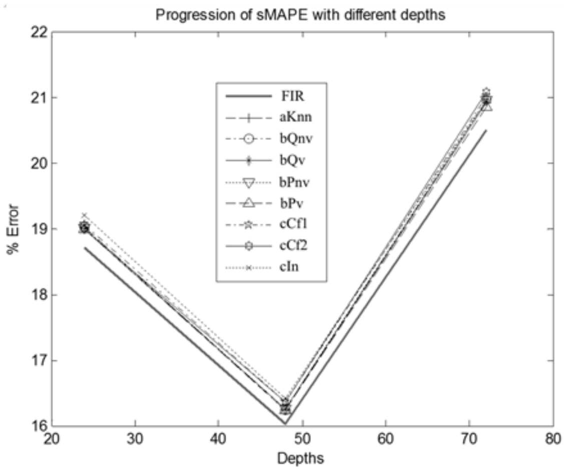


Figure 8. sMAPE error evolution against different depths (24, 48 and 72)

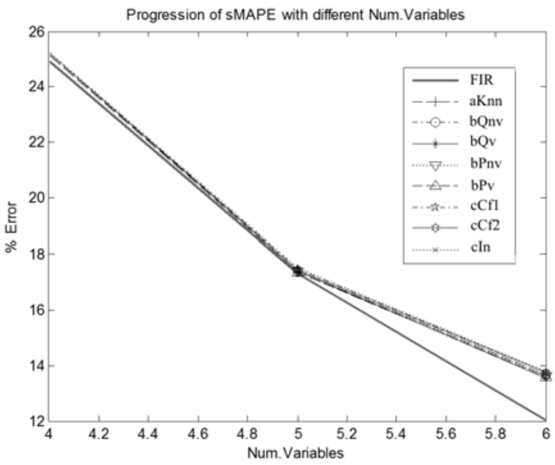


Figure 9. sMAPE evolution against different number of input variables (4, 5 and 6)

Similarly, Figure 10 and Figure 11 show the error evolution with sMAPE in different flexible FIR Prediction model configurations, when standard FIR cannot predict due to the missing data problem.

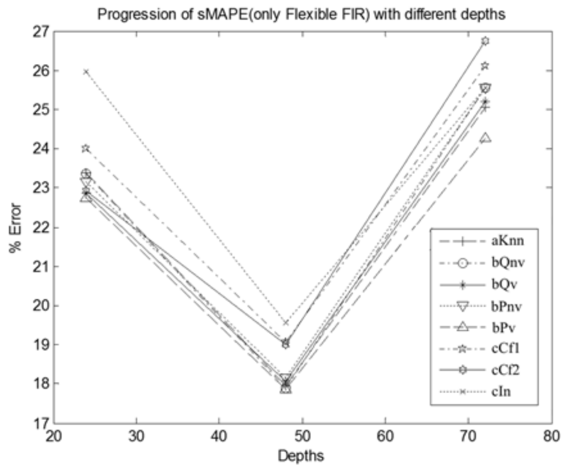


Figure 10. sMAPE error evolution against different depths (24, 48 and 72) when standard FIR cannot predict

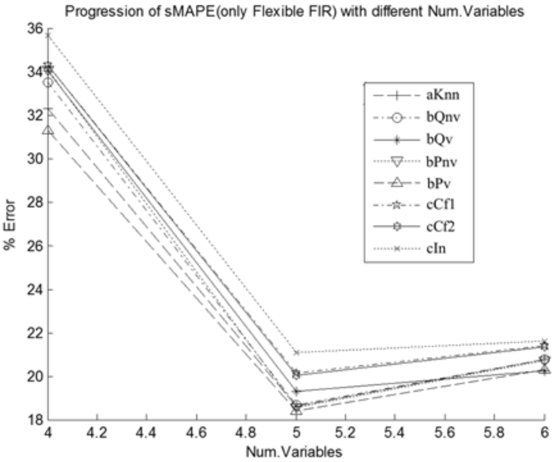


Figure 11. sMAPE error evolution against number of variables (4, 5 and 6) when standard FIR cannot predict

In Table 6 the total average registers predicted in *experiment 1* by the different methodologies are shown. When the flexible FIR Prediction was used, all the registers (from the test dataset) were predicted, while for the standard FIR decreases up to 85% with some configuration parameters.

Experiment 2

In Figure 12 a progression of the percentage of registers predicted versus percentage of MVs in the training process is shown. The *y* axis indicates the total number (in percentage) of registers that have been possible to predict from the test dataset. The *x* axis represents the percentage of MVs inserted in the training dataset (details of the experiment are explained in section *4.1.3.2 Model parameters Experiment 2*). In addition, different points of the curves have a numerical value associated, which is the average of the 8 building's sMAPE. Those percentages in grey colour belong to standard FIR, while the rest are an average among the eight flexible FIR strategies. The percentage of registers predicted with flexible FIR remains almost the same (around 100%), even though percentage of MVs in training data is more than 70%. It is observed a decrease of registers predicted with 81% of MVs. On the other hand, the percentage of registers with standard FIR decline to less than 10%, when MVs are present in around 20% of training data.

As it can be expected the lowest errors are observed when the percentage of MVs is also low. sMAPE of standard FIR remains low (between *7.75%* and *18.18%*) but with few registers predicted (from *84.77%* to *0%*). As it happens in *Experiment 1*, sMAPE of standard FIR does not contain those registers with the missing value problem, i.e. registers that standard FIR is not able to predict at all. On the other hand, flexible FIR strategies on average have an accuracy from *86.14%* (*% of MVs in training data= 9%*) to *75.13%* (*% of MVs in training data= 72%*). However, unlike standard FIR, all strategies of this new approach are able to predict almost *100%* of the registers except when *% of MVs in training data* is really very high, i.e. 81%.

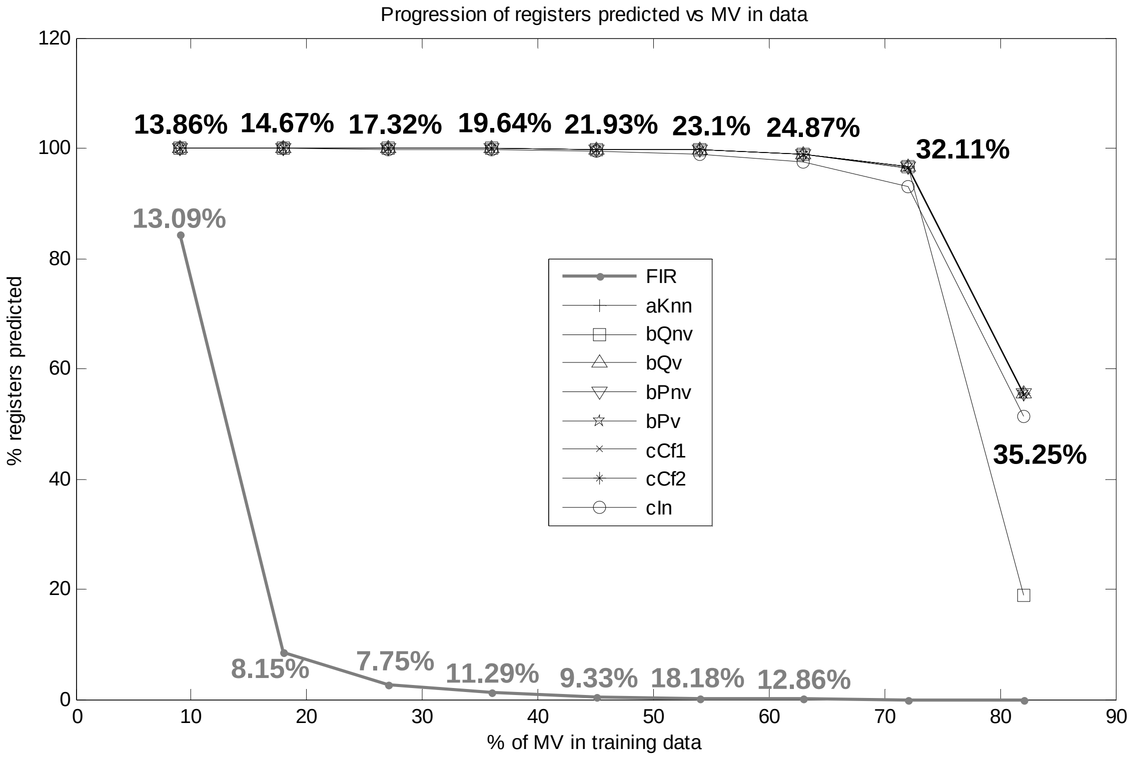


Figure 12. Prediction error (sMAPE) progression when % of MVs is increased in Standard and flexible FIR's training process

In Figure 13. Prediction error (sMAPE) progression when % of MVs is increased only in flexible FIR's training processTwo out of the eight flexible FIR strategies performance are lower: *Consist.2* (cCf1) and *Inertia* (cCf2). As it can be observed, the prediction error of 6 out of the 8 strategies are very similar and it cannot be seen the difference among them. However, the aim of these two figures is to show the tendencies of the strategies.

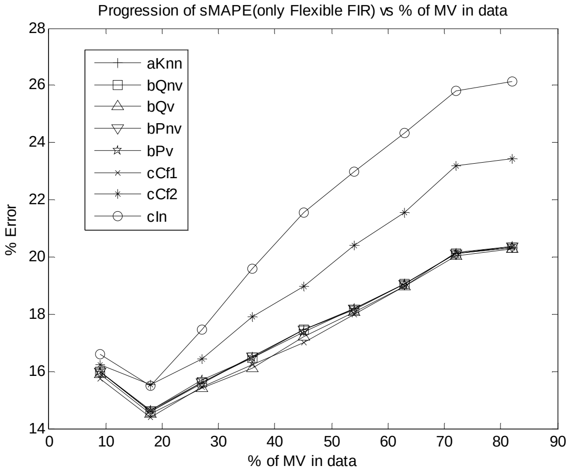


Figure 13. Prediction error (sMAPE) progression when % of MVs is increased only in flexible FIR's training process

5.2 Discussion

Based on Figure 8 and Figure 9 it is clear that, with this data, the optimal configurations of standard and flexible FIR are when mask depth is 48 and the number of input variables is 6. Whereas based on Figure 10 and Figure 11, where standard FIR cannot predict, the optimal configuration for flexible FIR is a depth of 48 and 5 input variables. From these results we can get that adding more information to the FIR model (more input variables) both standard and flexible FIR improve the prediction accuracy, which means that variables working day and hour of the day add valuable information to the models. When the predictions are only in registers affected by the presence of MVs, the variable hour of the day is not contributing to improve prediction accuracy.

As it can be seen in Table 4 prediction errors of flexible FIR Prediction are very close to standard FIR, although the last one discards predictions with the missing data problem. When the focus is on the results of registers that could not be predicted by standard FIR (Table 5), *bPv* (use of causal relevance Pvar) is the strategy with better results in more model configurations. Moreover, in 6 out of 9 model configurations, the strategy with better result utilise the causal relevance. These results suggest that the use of causal relevance in flexible FIR mitigates the impact of dispersion of output classes and the loss of causality due to removing of partial information from the behaviour matrix.

It is important to highlight that both *bPv* and *bPnv* compute the causal relevance based on the MSE of the prediction of a validation dataset, instead of being based on the quality of the mask (as it is done with *bQnv* and *bQv*). This means that the computational cost to obtain the causal relevance P_{var} and P_{novar} is higher than Q_{var} and $Q_{novar,}$ because in the first case the whole FIR forecasting process is needed, while in the second only the computation of mask's quality. Thus, we believe that for a future industrial applications where computational cost is an important variable to consider, *bQnv* and *bQv* are the best strategies to be implemented.

cCf2 and *cIn* accuracy decreases faster than the rest of strategies, when MVs are present in the training dataset, as it can be observed in Figures 10 and 11. *cCf1, cCf2* and *cIn* make use of the $O_{variance}$ to check if there is dispersion of the possible output classes. When we compare *cCf2* with *aKnn,* which is the default output forecast strategy of FIR, we can see that (on average) *aKnn* predicts better than *cCf2*. This becomes more noticeable when the percentage of MVs in data increases (Figure 12 and Figure 13). Similarly, *cIn* does not improve the prediction accuracy compared to *aKnn.* This does not mean that the *confidence measure* is an incorrect approximation to measure the probability to perform a prediction, but the use of this parameter in the strategy is not well integrated.

Unlike *cCf2* and *cIn,* prediction error in *cCf1* strategy does not increase so fast for higher percentage of MVs in the data. Nevertheless, looking at Table 4 and Table 5, on average, *aKnn* also predicts better than *cCf1*. Thus *cCf1* strategy does not contribute significantly to improve the forecasting when dispersion of output classes is present.

In consequence, it is better to respect the distances (d_j) between *norm images* (p_i) than to ignore the dispersion of possible output classes, or use the previous value as the predicted value, if there is a high probability to perform a bad prediction.

Based on the results from *experiment 2,* the robustness of flexible FIR and its eight strategies are demonstrated taking into account that the percentage of registers predicted is around *96%* when the *% of MVs in training data* was around *73%*. A significant decrease in the number of registers predicted can be observed when MVs are present in *82%* of training dataset, which is high number of MVs. Furthermore, the difference of prediction error when *% of MVs in training data* is *9%* (sMAPE = *13.86%*) and 6*3%* (sMAPE = *24.87%*) is around *11%*, which is not a big difference taking into account the volume of information lost. Thus, flexible FIR is able to keep a good compromise between information lost and prediction accuracy.

6. Conclusions and Future Work

In this work a comparison of FIR forecasting strategies is performed when MV is present in data. Due to some limitations in standard FIR, it is not able to perform predictions when it contains missing data in the input pattern and also it discards pattern rules with MVs. To cope with this situation we proposed an improved version of the FIR fuzzy forecasting, which is called flexible FIR Prediction. Eight different strategies of prediction for this new approach are proposed and compared to understand: how loss of causality of a variable, dispersion of output classes and increase of missing data in the training process affect the forecasting accuracy. This study defines a more reliable and robust FIR forecasting process, and could help for instance, in load modelling applications where missing data is present in online predictions.

The experiments presented contain MVs during the training process and the online prediction. All eight strategies proposed are able to cope with them and take advantage of the inherent information in the data, even though it contains MVs. In this research, we demonstrate that flexible FIR is able to predict electricity consumptions accurately, making use of partial information, whereas standard FIR cannot.

On average, the strategy with lower prediction errors is *bPv*, which uses the causal relevance P_{var}. Whilst *cCf2* and *cIn* are strategies with higher prediction errors. The lowest prediction error is found with strategy *aKnn* when depth is equal to 48 and with 4 input variables.

The outcome of our study sheds light on robust Soft Computing methodologies for smart home and smart grid applications. As future work, we plan to integrate flexible FIR, and some of the proposed strategies, in a smart home Internet of Things (IoT) platform that will provide home context awareness such as prediction of events or patterns detection. This is an interesting application due to the intermittent communication of home area sensors, which generate significant number of MVs. In a first integration of flexible FIR in the smart home IoT results are encouraging compared to standard prediction and pattern recognition frameworks such as Weka or RapidMiner.

Acknowledgement

We would like to thank Narcís Avellana for many fruitful discussions and his support in this research.

References

[1] Jerez, J.M., Molina, I., García-Laencina, P.G., Alba, E., Ribelles, N., Martín, M., Franco, L. 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem, *Artificial Intelligence in Medicine*, vol. 50(2), pp. 105-115.

[2] www.iurban-project.eu

[3] www.greencom-project.eu

[4] Tresp, V., Ahmad, S. and Neuneier, R. 1994. Training Neural Networks with Deficient Data. *Advances in Neural Information Processing Systems 6*, pp. 128-135

[5] Lakshminarayan, K., Harp, S.A., Goldman, R., Samad, T. 1996. Imputation of missing data using machine learning techniques, in *Proceedings: Second International Conference on Knowledge Discovery and Data Mining*, pp. 140-145

[6] Baraldi, P., Di Maio, F., Genini, D., Zio, E. 2015. Reconstruction of missing data in multidimensional time series by fuzzy similarity, *Applied Soft Computing*, vol. 26, pp. 1-9

[7] Lei, K.S., Wan, F. 2010. Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau, *in Automation and Logistics (ICAL)*, pp.418-422

[8] Luo, X., Zhou, M., Xia, Y., Zhu, Q., Ammari, A.C., Alabdulwahab, A. 2015. Generating Highly Accurate Predictions for Missing QoS Data via Aggregating Nonnegative Latent Factor Models, *Neural Networks and Learning Systems*, vol.PP(99), pp.1-1

[9] Nebot, A. and Mugica, F. 2012. Fuzzy Inductive Reasoning: a consolidated approach to data-driven construction of complex dynamical systems. *International Journal of General Systems*, vol. 41(7), pp. 645-665.

[10] Escobet, A., Nebot, A., Cellier, F.E. 2008. Visual-FIR: A tool for model identification and prediction of dynamical complex systems, *Simulation Modelling Practice and Theory*, vol. 16, pp. 76-92.

[11] Jurado, S., Nebot, A., Mugica F. 2015. Hybrid methodologies for electricity load forecasting: Entropy-Based Feature Selection with Machine Learning and Soft Computing Techniques. *Energy,* vol. 86, pp 276-291.

[12] Jurado, S., Peralta, J., Nebot, A., Mugica, F., and Cortez, P. 2013. Short-term electric load forecasting using computational intelligence methods, *FUZZ-IEEE 2013: 2013 IEEE International Conference on Fuzzy Systems*. doi: 10.1109/FUZZ-IEEE.2013.6622523.

[13] Nebot, A., Mugica, F., Cellier, F., Vallverdú, M. 2003. Modeling and Simulation of the Central Nervous System Control with Generic Fuzzy Models, *Transactions of The Society for Modeling and Simulation*, vol. 79(11), pp. 648-669.

[14] Gómez, P., Nebot, A., Ribeiro, S., Alquézar, R., Mugica, F. and Wotawa, F., 2003, Local maximum ozone concentration prediction using soft computing methodologies. Systems Analysis Modelling Simulation, 43(8), pp. 1011–1031.

[15] Brouwer, R.K., Pedrycz, W. 2003. Training a feed-forward network with incomplete data due to missing input variables, *Applied Soft Computing*, vol. 3(1), pp. 23-36

[16] van Lint, J.W.C., Hoogendoorn, S.P., van Zuylen, H.J. 2005. Accurate freeway travel time prediction with state-space neural networks under missing data, *Transportation Research Part C: Emerging Technologies*, vol. 13(5-6), pp. 347-369

[17] Leke, C., Twala, B., Marwala, T. 2014. Modeling of missing data prediction: Computational intelligence and optimization algorithms. *Systems, Man and Cybernetics (SMC)*, pp.1400-1404.

[18] Jurado, S., Nebot, A. and Mugica, F. 2015. A flexible fuzzy inductive reasoning approach for load modelling able to cope with missing data. *In Proceedings of the 8th International Conference on Simulation Tools and Techniques (SIMUTools '15)*, pp. 349-356.

[19] Klir, J.,Elias, D. (2002). Architecture of Systems Problem Solving, 2nd. Ed., Plenum Press, New York.

[20] Jerez, A. and Nebot, A., 1997. Genetic algorithms versus classical search techniques for identification of fuzzy models, *Proceedings EUFIT'97, 5th European Congress on Intelligent Techniques and Soft Computing*. Aachen, Germany, pp. 769-773.

[21] Law, A.M. and Kelton, W.D. 1991. *Simulation Modeling and Analysis*, vol. 2, McGraw-Hill.

[22] Nebot, A., Mugica, F., Castro, F. 2009. Causal relevance to improve the prediction accuracy of dynamical systems using inductive reasoning. *International Journal of General Systems*, vol. 38:3, pp. 331-358

[23] Castro, F., Nebot, A., Mugica, F. 2008. A Soft Computing Decision Support Framework to Improve the e-Learning Experience. *MSE'08: Modeling and Simulation Education – Spring Simulation Multiconference*, pp. 781-788.

[24] Castro, F., Nebot, A., Mugica, F. 2007. Causal relevancy approaches to improve the students' performance in an e-learning environment, *MICAI'07: 6th Mexican Internacional Conference on Artificial Intelligence*, IEEE Computer Society Highlights, pp. 342-351.

[25] Cellier, F. E., López, J., Nebot, À. and Cembrano, G. 2010. Confidence measures for predictions in fuzzy inductive reasoning, *International Journal of General Systems*, vol. 39:8, pp. 839-853

[26] Hyndman, R.J., Koehler, A.B. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting*, vol. 22, pp. 679-688.

[27] Makridakis, S. 1993. Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, vol. 9, pp. 527 – 529