Energy 86 (2015) 276–291

Contents lists available at ScienceDirect

Energy

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

Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques

Sergio Jurado [a, *], Àngela Nebot [b], Fransisco Mugica [b], Narcís Avellana [a, 1]

[a] Sensing & Control Systems, Aragó 208-210, 08011 Barcelona, Spain
[b] Soft Computing Research Group, Technical University of Catalonia, Jordi Girona 1-3, 08034 Barcelona, Spain

ARTICLE INFO

Article history:
Received 20 October 2014
Received in revised form
3 April 2015
Accepted 4 April 2015
Available online 23 May 2015

Keywords:
Building electricity forecasting
Entropy-based feature selection
Fuzzy Inductive Reasoning
Neural Networks
Random Forest
ARIMA

ABSTRACT

Scientific community is currently doing a great effort of research in the area of Smart Grids because energy production, distribution, and consumption play a critical role in the sustainability of the planet. The main challenge lies in intelligently integrating the actions of all users connected to the grid. In this context, electricity load forecasting methodologies is a key component for demand-side management. This research compares the accuracy of different Machine Learning methodologies for the hourly energy forecasting in buildings. The main goal of this work is to demonstrate the performance of these models and their scalability for different consumption profiles. We propose a hybrid methodology that combines feature selection based on entropies with soft computing and machine learning approaches, i.e. Fuzzy Inductive Reasoning, Random Forest and Neural Networks. They are also compared with a traditional statistical technique ARIMA (AutoRegressive Integrated Moving Average). In addition, in contrast to the general approaches where offline modelling takes considerable time, the approaches discussed in this work generate fast and reliable models, with low computational costs. These approaches could be embedded, for instance, in a second generation of smart meters, where they could generate on-site electricity forecasting of the next hours, or even trade the excess of energy.

1. Introduction

The fast development of new infrastructure in the electricity grid, what is known as SG (Smart Grid), is opening a wide range of opportunities. An SG is an advanced electricity transmission and distribution network (or grid) that uses information, communication, and control technologies to improve economy, efficiency, reliability, and security of the grid [1]. Nowadays, it is a priority of many governments worldwide to replace/upgrade old electricity grids from several decades ago with SG. For example, in 2010, the US government spent $7.02B on its SG initiative, while the Chinese government used $7.32B for its SG program [2].

The SG is characterized by several new trends and features: smart meters, demand response mechanisms, online customer interactions through PCs/mobile devices, dynamic electricity tariffs, online billing, incorporation of renewable energy generation (such as solar and wind energy) and electric vehicles, more reliable power transmission and distribution, dynamic load balancing, better power quality, better power security, etc.

IT (Information technology) is one of the major driving forces behind an SG, and various IT systems and techniques such as artificial intelligence, high performance computing, simulation and modelling, data network management, database management, data warehousing, and data mining could be used to facilitate smooth running of the SG [3].

Therefore, a huge gap exists between the SG and outcomes expected that has to be covered in the next years by the combination of experts in the IT and energy domain creating technologies that support these new scenarios with: (i) Novel optimisation algorithms for the energy distribution, (ii) Total control: enhancement of the monitor and control over the electricity grid and (iii) Technologies for the local production and distribution.

Most of the features and proposed scenarios are based on reliable, accurate and fast energy predictions. Distributed demand and

* Corresponding author. Tel.: +34 605 565 303.
E-mail addresses: sergio.jurado@sensingcontrol.com, s.juradogomez@gmail.com (S. Jurado), angela@lsi.upc.edu (À. Nebot), fmugica@lsi.upc.edu (F. Mugica), narcis.avellana@sensingcontrol.com (N. Avellana).
[1] Tel.: +34 931 763 520.

http://dx.doi.org/10.1016/j.energy.2015.04.039
0360-5442/

supply relies on knowing the individual consumption and/or production of public and private buildings, industries, and distribute generation power plants, too. Some of the predictions proposed in the literature are based on architectonic features such as heat loss surface, building shape factor, building heated volume and so on [4,5], or housing type and socioeconomic features such as age of the dwelling, size of the dwelling, monthly household income, number of household members, etc. [6]. However, the cost for extracting this information is very high in terms of personnel and tools. Moreover, this information can only be obtained by intrusive methods, i.e. polls at households.

Nevertheless, the massive deployment of smart metering as part of the developing SG allows energy companies to have access to the energy consumption and/or production with accuracy of minutes, in every dwelling by using remote metering. Therefore, time-series data can be collected and analysed to form predictions based on historical values. This approach is more scalable, non-intrusive and as accurate as other forecasting approaches.

To take advantage on the data generated by the smart meters and demonstrate the robustness of data driven models, we address different Machine Learning and Soft Computing methodologies for the hourly electricity forecasting in buildings, with different energy profiles, locations and weather conditions. We validate our models with real electricity consumption data and demonstrate their scalability in different consumption profiles.

A large variety of AI (Artificial Intelligent) techniques have been applied in the field of short-term electricity consumption forecasting, showing a better performance than classical techniques. Specifically, Machine Learning has been proven to accurately predict electric consumption under uncertainties. For instance, Khamis proposes in Ref. [7] a multilayer perceptron neural network to predict the electricity consumption for a small scale power system, obtaining a better performance than with traditional methods, while Marvuglia et al. consider Elman neural network for the short forecasting of the household electric consumption with prediction errors under 5% [8]. Also in Ref. [9] a study of electric load forecasting is carried out with CART (Classification and Regression Trees) and other soft computing techniques obtaining again better results than classical approaches.

Large scale studies for comparing machine learning and soft computing tools have focused on the classification domain [10]. On the other hand, very few extensive studies can be found in the regression domain. In Ref. [11] Nesrren et al. carried out a large scale comparison of machine learning models for time series forecasting. The study includes techniques such as KNN (K-Nearest Neighbours), CART regression trees, multilayer perceptron networks, Support Vector Machines, Gaussian processes, Bayesian Neural Networks and radial basis functions. The research reveals significant differences between the methods studied and concludes that the best techniques for time series forecasting are multilayer perceptron and Gaussian regression when applied on the monthly M3 time series competition data (a thousand time series) [12]. Moreover, in Ref. [13] an empirical comparison of regression analysis is carried out, as well as decision trees and ANN (artificial neural networks) techniques for the prediction of electricity energy consumption. The conclusion was that the decision tree and the neural network models perform slightly better than regression analysis in the summer and winter phases, respectively. However, the differences between the three types of models are quite small in general, indicating that the three modelling techniques are comparable when predicting energy consumption.

In contrast to the these approaches where offline modelling takes considerable computational time and resources, the models discussed in this work appear to generate fast and reliable models, with low computational costs. These models could be embedded, for instance, in a second generation of smart meters where they could generate on-site forecasting of the consumption and/or production in the next hours, or even trade the excess energy with other smart meters. The paper is structured as follows: In section 2 an analysis of the Machine Learning and Soft Computing modelling approaches used in this research is offered. Then, section 3 presents the data sets used, that come from three buildings of the UPC (Technical University of Catalonia) with different profiles and locations, and the experiments performed. Next, section 4 describes the results obtained when applying the proposed methods to the data sets, as well as a discussion of the results encountered is performed. Finally, section 5 gives some conclusions, future work and new perspectives..

2. Machine learning, soft computing and statistical modelling techniques

In this research we compare the prediction accuracy of a Machine Learning methodology, two Soft Computing techniques and one traditional statistical method for the hourly energy forecasting in buildings: RF (Random Forest), Artificial NN (Neural Networks), FIR (Fuzzy Inductive Reasoning) and ARIMA (AutoRegressive Integrated Moving Average), respectively. A FSP (Feature Selection Process) is first applied to the historical consumption data in order to determine the features with the largest impact on the prediction accuracy of future consumption values. Afterwards, different experiments described in section 3, are designed. They combine the selected past consumptions with day and time information to be used as input variables for the three modelling approaches. Fig. 1, shows a scheme of the data flow and the components that take part in the building's consumption forecasting.

The FSP chosen in this research is a piece of the Fuzzy Inductive Reasoning methodology, which consists of using entropies instead of most typically used correlation approaches, such as CCA (canonical correspondance analysis) [14] or PCA (principal component analysis) [15], to find out the most important input variables for the modelling techniques. The entropy approach selects the set of variables that better explain, as a whole, a specific output (hence, called relevant). The FSP is described in the next section.

2.1. FSP (Feature Selection Process)

In this work the FSP of FIR methodology is used not only in FIR modelling but also as a pre-processing for NN and RF modelling. Therefore, a short insight of the FIR feature selection algorithm is given here.

In FIR methodology a model is composed of two parts, i.e. the model structure (or mask in FIR nomenclature) and the pattern rule base [16]. The process of obtaining a FIR model structure corresponds to an FSP. The model structure holds in the relevant features

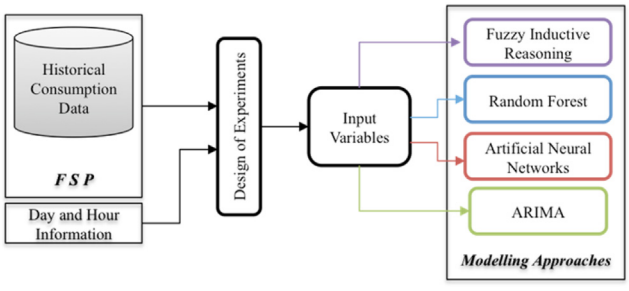

Fig. 1. Components involved in the building's consumption forecasting: FSP applied to historical consumption data, additional input variables such as day and time information, design of experiments and the three modelling approaches used in this work.

and is represented by a mask through which the causal relations (both spatial and temporal) between input and output variables 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.

Each negative element in the mask exhibits a causal relation with the output, i.e. it influences the output up to a certain degree. The single positive value denotes the 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, 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 Visual-FIR is used to obtain the best mask, i.e. the best FIR structure, for the system under study [19]. The procedure consists of 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. Then the best from these masks is 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 can also be found, where the mask has its output. Finally, elements of value 0 denote forbidden connections.

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 of 1 h is used, due to the data characteristics.

The optimal mask function of Visual-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 legal masks. Then it is proceed, by searching through all legal masks of complexity three, i.e. all masks with two inputs and the output, and finds the best of those. 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 [18].

Each of the possible masks 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).

$$Q_m = H_r \cdot O_r \tag{1}$$

The uncertainty reduction measure is defined in equation (2).

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

t	x				
	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

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).

t	x				
	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

$$H_r = 1 - H_m/H_{max} \tag{2}$$

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 an improved forecasting power. The masks with highest entropy reduction values generate forecasts with the smallest amounts of uncertainty. The highest possible entropy H_{max} is obtained when all probabilities are equal, and zero entropy is encountered for totally deterministic relationships. The overall entropy of the mask is then computed as described in equation (3).

$$H_m = -\sum_{\forall i} p(i)\cdot H_i \tag{3}$$

where $p(i)$ is the probability of that input state to occur and H_i is the Shannon entropy relative to the ith input state. The Shannon entropy relative to the ith input state is calculated from the equation (4).

$$H_i = \sum_{\forall o} p(o|i)\cdot log_2 p(o|i) \tag{4}$$

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, every state should be observed at least five times [17]. 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 Qm 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.

2.2. FIR (Fuzzy Inductive Reasoning)

The conceptualization of the FIR methodology arises of the GSPS (General System Problem Solving) 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 [16]. A FIR model is a qualitative non-parametric model based on fuzzy logic. Visual-FIR is a tool based on the FIR methodology (runs under Matlab environment), which offers a new perspective to the modelling and simulation of complex systems. Visual-FIR designs process blocks that allow the treatment of the model identification and prediction

phases of FIR methodology in a compact, efficient and user friendly manner [20]. The FIR model consists of its structure (relevant variables or selected features), which has been previously explained in section 2.1, and a pattern rule base (a set of input/output relations or history behaviour) that are defined as if-then rules. Once the best structure (mask) has been identified, it is used to obtain the pattern rule (called behaviour matrix) from the fuzzified training data set. Each pattern rule is obtained by reading out the class values through the 'holes' of the mask (the places where the mask has negative values), and place each class next to each other to compose the rule.

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 Fig. 2. This process is called qualitative simulation. The FIR inference engine is based on the KNN rule, commonly used in the pattern recognition field. The forecast of the output variable is obtained by means of the composition of the potential conclusion that results from firing the k rules whose antecedents have best matching with the actual state.

The mask is placed on top of the qualitative data matrix (fuzzified test set), in such a way that the output matches with 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 latter, to determine the future value of the output, which can then be copied back into the qualitative data matrix. The mask is then shifted further down one position 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, the ones with smallest distance, that 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. This is expressed by giving a distance weight to each neighbour, as shown in Fig. 2. The new output state values can be computed as a weighted sum of the output states of the previously observed five nearest neighbours.

The FIR methodology is, therefore, a modelling and simulation tool that is able to infer the model of the system under study very quickly and is a good option for real time forecasting. Moreover, it is able to deal with missing data as has been already proved in a large number of applications [16]. On the other hand, some of its weaknesses are that as long as the depth and complexity increase the computational cost increases too, and also the parameters to choose during the fuzzification phase (which can be mitigated using evolutionary algorithms to tune the parameters).

2.3. RF (Random Forest)

RF (Random Forest) is a set of CART (Classification and Regression Trees), which was first put forward by Breiman [21]. In RF, the training sample set for a base classifier is constructed by using the Bagging algorithm [22]. In traditional CART, each inner node is a subset of the initial data set and the root node contains all the initial data. RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. RFs for regression are formed by growing trees depending on a random vector such that the tree predictor takes on numerical values as opposed to class labels. The RF predictor is formed by taking the average over B of the trees. Fig. 3 shows a scheme of the random forest.

Assuming the following basic notations:

- Let the number of training cases be N and the number of variables be P
- Input data point → $v = (x_1, \ldots, x_P) \in \mathbb{R}^P$
- Output variable → c
- Let the number of trees be B

The schematic RF algorithm is the following:

1. For $b = 1$ to B
   a. Draw a bootstrap sample Z^* of size N from the training data

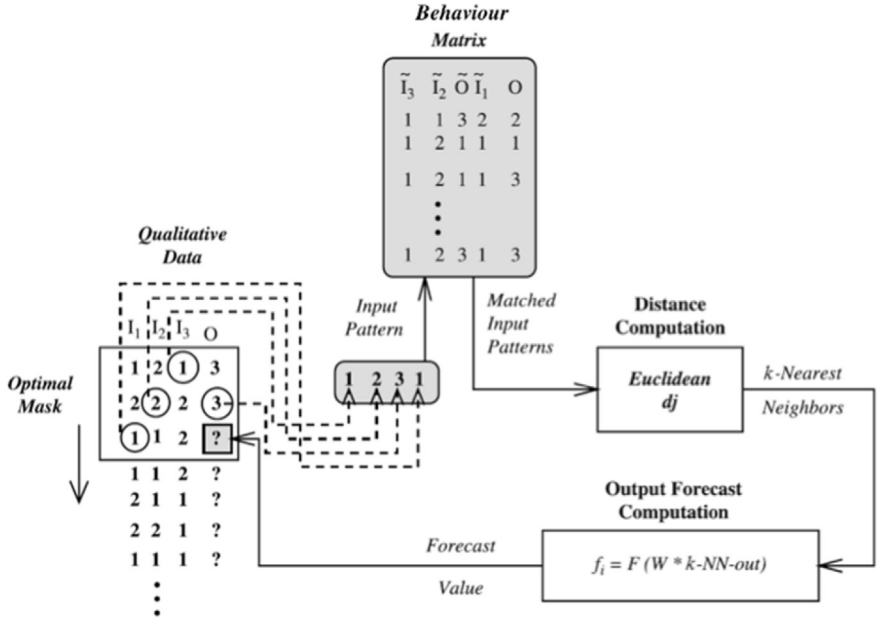

Fig. 2. Qualitative simulation process diagram (with an example containing three inputs and one output).

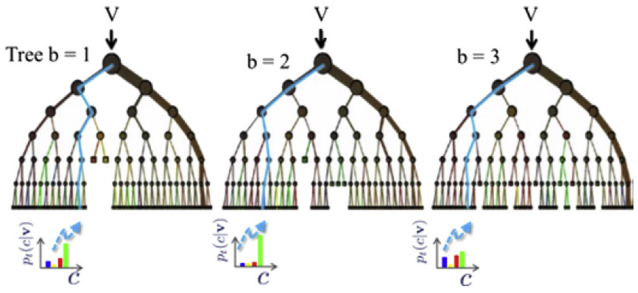

Fig. 3. Random Forest scheme containing three different trees.

b. Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the stopping criteria is reached:
   i. Select m variables at random from the P variables
   ii. Pick the best variable/split-point among the m
   iii. Split the node into two daughter nodes
2. Output the ensemble of Trees $\{T_b\}_1^B$: $p(c|v) = \frac{1}{B}\sum_1^B p_b(c|v)$

The size N of the bootstrap sample Z^* can go from a small size to the size of the whole data set. However, with large training data sets, assuming the same size can affect significantly to the computational cost. In addition, for big data problems such as the forecasting of consumption/production of all the buildings in a city, a good definition of the parameter N is mandatory. Moreover, there are different stopping criteria, two of the most commonly used are: 1) until the minimum node size n_{min} is reached, and 2) when a maximum tree depth is reached.

Although RF has been observed to overfit some datasets with noisy classification/regression tasks [23], it usually provides accurate results, generalizes well and learns fast. In addition, it is suitable to handle missing data and provides a tree structured method for regression [24].

2.4. Artificial NN (Neural Networks)

Neural networks are a very popular data mining and image processing tool. Their origin stems from the attempt to model the human thought process as an algorithm which can be efficiently run on a computer. Its origins date back to 1943, when neurophysiologist W. McCulloch and mathematician W. Pitts wrote a paper on how neurons might work [25], and they modelled a simple neural network using electrical circuits. Some years later, in 1958, F. Rosenblatt created the perceptron, an algorithm for pattern recognition based on a two-layer learning computer network using simple addition and subtraction [26].

Many time-series models are based on NN [27]. Despite the many desirable features of NNs, constructing a good network for a particular application is a non-trivial task. It involves choosing an appropriate architecture (the number of layers, the number of units in each layer, and the connections among units), selecting the transfer functions of the middle and output units, designing a training algorithm, choosing initial weights, and specifying the stopping rule.

It is widely accepted that a three-layer feed forward network with an identity transfer function in the output unit and logistic functions in the middle-layer units can approximate any continuous function arbitrarily well given sufficient amount of middle-layer units [28]. Thus, the network used in this research is a three layer feed forward network (Fig. 4). The inputs are connected to the output via a middle layer.

When working with univariate time series, the neurons of the input layer contain the present value and the relevant past values of the univariate time series. While the output is the value for the next time period, computed as described in equation (5).

$$S(t+1) = f(s(t), \ldots, s(t-n)) \quad (5)$$

where n is the number of past values of variable s, and f is a nonlinear function approximated by a multilayer feed FNN (forward neural network) [29].

Recurrent neural networks are able to obtain very good prediction performance, since their architecture allows that the connections between units form a directed cycle, which allows it to exhibit dynamic temporal behaviour. Unlike feed forward neural networks, recurrent networks can use their internal memory to process arbitrary sequences of inputs. Therefore, recurrent neural networks are very powerful, but they can be very complex and extremely slow compared to feed forward networks. As mentioned earlier, one of the main objectives of this research is to find powerful prediction methodologies with low computational costs, which could be embedded in a smart meter and generate on-site forecasting of the consumptions and/or productions. This is the reason why recurrent neural networks have been rejected in this work and a cooperative approach has been chosen instead, i.e. FSP and a feed forward neural network that uses, as input variables, the most relevant past consumptions values.

2.5. ARIMA (AutoRegressive Integrated Moving Average)

Traditional statistical time's series forecasting model has been incorporated to compare the forecasting results of the previous hybrid AI methodologies with traditional benchmark methods: ARIMA [30]. Three parts compose ARIMA: an AR (autoregression model), where there is a combination of past values; a MA (moving average component), which uses past forecast errors in a regression-like model; and an I (integration), referring to the reverse process of differencing to produce the forecast.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.

The forecasting equation is constructed as follows. First, let y denote the dth difference of Y, which means:

- If $d = 0$: $y_t = Y_t$
- If $d = 1$: $y_t = Y_t - Y_{t-1}$
- If $d = 2$: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

if we consider the time series stationary to simplify things, with $d = 0$, the equation can be written as:

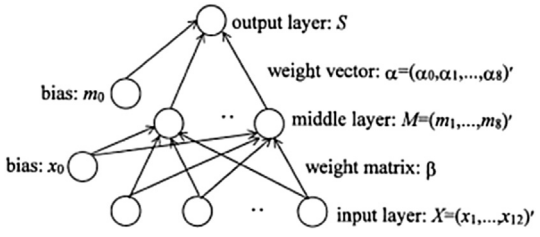

Fig. 4. Representation scheme of a three-layer feed forward neural network.

$$y_t = \underbrace{a_0 + a_1{*}y_{th1} + a_2{*}y_{th2} + \ldots + a_p{*}y_{thp}}_{\text{Autoregressive Part}} + \underbrace{b_0 + b_1{*}e_{th1} + b_2{*}e_{th2} + \ldots + b_q{*}e_{thq}}_{\text{Moving Average Part}} \quad (6)$$

where a_0, a_1 … a_p, are the nonseasonal autoregressive coefficients corresponding to a stable polynomial. Associated lags are y_{th1}, y_{th2} … to the degree of the nonseasonal autoregressive polynomial p. And b_0, b_1 … b_p are the nonseasonal moving average coefficients. Associated lags are e_{th1}, e_{th2} … to the degree of the nonseasonal moving average polynomial q. Note that e_{thi} is the past forecast error between *Real* y_{thi} − *Predicted* y_{thi}.

It is also possible to include seasonal components to the ARIMA model, but in this study we are not modelling the seasonality of the time series. On the contrary, we want to evaluate how robust and adaptive can the model be, across the changes according to season, week day and time of the day.

3. Methods

3.1. Data set

Data of three buildings of the UPC (Universitat Politècnica de Catalunya) was obtained for this study. These buildings belong to three different campuses and are located in different cities. In order to demonstrate the scalability of the models in any type of building, three functional zones with different profiles of usage and locations are used in the experiments. Thus, affecting different climatology (temperature, humidity, solar radiation, etc.), consumption patterns, schedules and working days: 1) the Library of the ETSEIAT[2] faculty in Terrassa of more than 500 m^2; 2) the bar of the ETSAV[3] faculty in Sant Cugat of 150 m^2; 3) a building with different classrooms at FIB[4] faculty in Barcelona of 630 m^2. The energy consumptions of these three buildings have been collected through a remote metering system every hour. Therefore there are 24 recordings per day and per location.

For all three zones, the data set comprise a whole year of electricity consumption, from 13/11/2011 to 12/11/2012, with 91% for training and 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. Table 3 shows the days chosen as test data for the electricity load forecasting and its distribution.

3.2. FIR FSP

In this study the modelling process consists in: 1) feature selection by means of the FIR methodology and 2) use these relevant features to derive a model for each proposed methodology, i.e. RF, NN and FIR, all of them trained by the consumption data of one year. This process is repeated for each location and each depth[5] studied. The data used for testing the models is removed from the initial data set and it is replaced by missing values. 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. In the near future we plan to study the modelling of each season separately versus the deseasonalization of the data.

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. And, 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 increases as well until the optimal values are reached. After that, increasing the number of variables may add noise to the system and end up with a result with lower quality. It has been empirically determined that more than four variables and depth higher than 72, do not increase significantly the quality of the FSP of FIR, while computational cost (in terms of time) does exponentially.

Therefore, in order to catch the most relevant previous consumptions in the electric load series, different depths are studied:

- previous 24 h;
- previous 72 h;
- 24 + 24: previous 24 h and the past 24 h of the previous week (48 past values in total that corresponds to a depth of the mask of 168).

Table 4 presents the results of the FSP performed by FIR for the different depths studied. As can be observed from Table 4, the features obtained for each depth analysed are almost the same for the Terrassa library and the C6 building. This makes sense since these two buildings have similar characteristics if compared to the Bar. The most relevant features encountered for Sant Cugat Bar differ from the other two buildings for depths 72 and 24 + 24. The FSP concludes that the consumptions at one, twelve, twenty-four and sixty-three hours before now and at one, eleven, twenty-four and one hundred fifty-eight hours before now are the ones that best represents the consumption for depth 72 and 24 + 24, respectively.

3.3. Outliers detection

Some data analysis methods are quite robust against outliers, i.e. observations that appears to deviate markedly from other observations in the sample, whereas other are extremely sensitive to outliers and might produce incoherent or nonsense results, just

[2] ETSEIAT: School of Industrial and Aeronautic Engineering of Terrassa.
[3] ETSAV: School of Architecture of the Vallès.
[4] FIB: Barcelona School of Informatics.
[5] Remember that the depth of the mask represents the temporal domain that can influence the output, and, therefore, the depth multiplied by the δt is the time in the past that is considered in the search space during the FSP. For instance if the depth is *72*, FIR FSP will search the most significant past values among the 72 previous consumptions values (hours).

Table 3
Distribution of the test data used for each model. In this case, each model has been trained using 7896 data points that correspond to the period 13/11/2011 to 12/11/2012 removing all the test days.

Autumn (13/11/2011 to 22/12/2011 + 23/9/2012 to 12/11/2012)	Winter (22/12/2011 to 20/3/2012)	Spring (21/3/2012 to 20/6/2012)	Summer (21/6/2012 to 22/9/2012)
23/11/2011 (Wednesday)	31/12/2011 (Saturday)	30/03/2012 (Friday)	30/06/2012(Saturday)
3 and 13/12/2011 (Saturday and Tuesday)	10, 20 and 30/01/2012 (Tuesday, Friday, Monday)	9, 19 and 29/04/2012 (Mon., Thur., Sunday)	10, 20 and 30/07/2012 (Thursday, Friday, Monday)
2, 12 and 22/10/2012 (Tuesday, Friday, Monday)	9, 19 and 29/02/2012 (Thurs., Sunday, Wednes.)	9, 19 and 29/05/2012 (Wednes., Satur., Tues.)	9, 19 and 29/08/2012 (Thurs., Sunday, Wednesday)
1 and 11/11/2012 (Tuesday, Friday)	10 and 20/03/2012 (Saturday, Tuesday)	8 and 18/06/2012 (Friday, Monday)	8 and 18/09/2012 (Saturday, Tuesday)

because of the presence of one or a few outliers. An outlier may indicate bad data due to a missing communication for instance between the smart meter and the central server or may be due to hardware errors. Therefore, it is useful to detect these observations and exclude them from the dataset.

There are different techniques for outlier detection. The box plot methodology [30–33], consists in depicting groups of numerical data through their quartiles. Box plots may also have lines which extend vertically from the boxes (whiskers), and they indicate variability outside the upper and lower quartiles. The problem with this technique is that the whisker value is selected according to standard deviation and coverage of the empirical distribution, which should be close to a normal distribution, and as it can be observed in the second row of Fig. 6, this is not our case. Similarly, in Ref. [34] the Grubb's test is proposed, however, the standard assumption is that the underlying distribution is normal as well.

Apart from the technique to be used for the outlier detection, it is important to know the nature of the data and the domain. In the electricity load, unexpected changes occur several times, for instance, in the morning when almost all appliances are turned off and residence wakes up, electricity load increases significantly. This unexpected change, although normal, can be detected as an outlier by some of the techniques.

A possible solution to this problem is to treat the abnormal load values as outliers and use corrective filters to pre-process the data and produce quality observations that can serve as input to the forecasting models. A simple powerful method is the running median [35], which consist of:

1) Computing an n-hour running median L_t^{med} of the original load series L_t;
2) Constructing filter bands $B_t = L_t^{med} \pm 3 \cdot SD \cdot (L_t - L_t^{med})$, where SD $(\cdot)$ is the standard deviation;
3) Identifying all observations outside the filter bands as outliers.

The running median is employed here as it is more robust to outliers than the commonly used moving average. It is advisable [35] to use both short (n = 5) – and long (n = 49) – term running medians, as the former have problems with adjacent outliers and the latter can detect only very large deviations from the standard range of the signal. As an example, the resulting filter bands (upper filter band – red and lower filter band – green) for the Terrassa Library is depicted in Fig. 5.

Unfortunately, automated corrective algorithms sometimes do not perform satisfactorily and human experts have to supervise the process.

Table 5 shows the number of hours selected in the running median, the outliers detected with the algorithm and the performance of the algorithm after human supervision. A total of 18 outliers have been detected and eliminated: 14 at the Building C6, 3 at the Terrassa Library and 1 at the Sant Cugat Bar.

The reason why for Terrassa Library the number of hours selected for the running median window is 5 instead of 49 is because the consumption data is more dynamic compared to Building C6 and Sant Cugat bar. This means that more unexpected changes are observed and in order to capture the shape of the time series it is needed to reduce the samples used. Meanwhile for Building C6 and Sant Cugat Bar the time series observed are more smoothed and therefore, a higher running median window captures better the nature of the data.

The outliers are treated as missing values and therefore, discarded form the training and test datasets. The reason why it has been preferred to leave them as missing values instead of being replaced by the KNN, is because (i) we may add wrong information in the time series and (ii) there are more than 8.000 data instances, thus the impact is not significant.

Table 4
Relevant hours of consumption obtained by FIR FSP (represented here as delays from current time). For instance, when a depth of 24 h is used, the FIR FSP obtains that the principal features are the consumptions at 1, 7, 13 and 24 h in the past.

Depth	Most important previous consumptions		
	Terrassa Library	*Sant Cugat Bar*	*Building C6*
24 previous hours	1,7,13,24	1,8,14,24	1,7,13,24
72 previous hours	1,24,31,61	1,12,24,63	1,24,30,61
24 + 24 previous hours	1,11,24,168	1,11,24,158	1,11,24,168

3.4. Statistical analysis

A statistical analysis of the output variable, i.e. electricity load consumptions, and input variables, i.e. past electricity load consumptions, hour of the day and is working day are done in order to select the best modelling approach. Table 6 summarizes the input and output variables in this study, introduces their symbols and indicates the type of variable.

Fig. 6 presents different graphs that help to understand and analyse the historical consumption data. In the first row the mean electricity load in each building is plotted, showing different shapes for each building. In Terrassa Library the consumption starts to increase significantly at 5 am and it maintains a constant consumption until 8 pm when starts to decrease. In case of C6 and Sant Cugat Bar the average load shape varies from Terrassa Library. They start to increase the consumption at 5:30 am and between 7 and 8:30 am approximately there is a slight decrease, probably because cleaning tasks start early in the morning and finish at 7, whereas lessons start on average at 9 am. Therefore, there are some minutes without significant consumption. After that, consumption increases in both buildings with peaks around 15 pm. The second, third and fourth rows of Fig. 6 present the empirical probability distributions (histograms) of *Y1, X1* and *X2* variables. Histogram of variables *X3.1, X3.2, X3.3* and *X3.4*, which are the four most significant past values

Fig. 5. One year of electricity load (blue) with 3 outliers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Parameters and performance of the running median algorithm for the outliers detection in each installation.

Installation	Number (n) of hours	Outliers detected with the algorithm	Performance	Remarks
Terrassa Library	5	3	100%	
Building C6	49	14	87.5%	Two outliers were not detected
Sant Cugat Bar	49	1	100%	

of the electricity load consumptions (see Table 4), are not shown due to the fact that they have almost identical representation than the histograms of variable *Y1*. As expected, the probability distribution of variable *X2* is the same for each hour, whilst for variable *X1* there are more working than no-working days. These plots demonstrate that none of the variables follows the normal distribution. In addition to that, it is interesting to mention that there is no clear pattern in the time series electricity consumption during the whole year, with a lot of peaks and unexpected changes. There is a clear decline of consumption in August, which is the vacation period in the UPC. February and June are exams periods with high density of students in the Terrassa Library; thus more devices are connected to the electricity grid.

The last two rows of Fig. 6 display the scatter plots for the input variables *X1* and *X2* with the normalised output variable *Y1*. The scatter plots of the past electricity load consumptions, i.e. *X3.1, X3.2, X3.3* and *X3.4*, with the normalised output variable Y1, when depth is 24 + 24 (see Table 4), are shown in Fig. 7. From the scatter plots of both figures it can be concluded that any functional relationship of the input variables and the output variable is not trivial.

The only exception is the relation between the input variable *X3.1* and the output variable *Y1* in the Building C6, which is close to be linear. Another interesting insight from these figures is that although all the scatter plots of the Building C6 follow similar patterns than the other two buildings, few values are zero in the *x-axis* and *y-axis*. This confirms the fact that this building is always consuming a considerable amount of electricity, because it houses some servers in the basement.

In conclusion we can reasonably expect classical learners, such as linear regression, to fail to find an accurate mapping of the input variables to the output variable. Therefore, these plots intuitively justify the need to experiment with non-linear learners such as RF, NN and FIR.

3.5. Evaluation criteria

There are many measures of forecast's accuracy in the literature [36]. We require a statistical quality measure, which is able to compare the different forecasting methods in buildings with different average loads.

The NMSE (Normalized Mean Squared Error), described in equation (7), is used as the error measure to evaluate the forecasted results.

$$NMSE = \frac{1}{N}\sum_{t=1}^{N}\left[\left(y_r(t) - y_f(t)\right)\right]^2 \Big/ var\left(y_{training}(t)\right) \tag{7}$$

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 data set.

Using the Mean Squared Error to evaluate the performance of the model can lead to a misinterpretation of the results. Values predicted in some buildings are in the order of 250 kWh, whereas, for instance, the Sant Cugat Bar is in the order of 20 kWh. Hence, we decided to use the MSE (mean squared error) divided by the variance of the training data set, in order to avoid this problem.

Table 6
Input and output variables involved in this study, their symbols and its type. I stands for input and O for output.

Symbol	Input (I) or output (O) variable	Type of variable
X1	Is Working Day (I)	Binary
X2	Hour of the Day (I)	Categorical {0,1,…,23}
X3.1, X3.2, X3.3, X3.4	Electricity Load Past Consumptions (I)	Continuous
Y1	Electricity Load (O)	Continuous

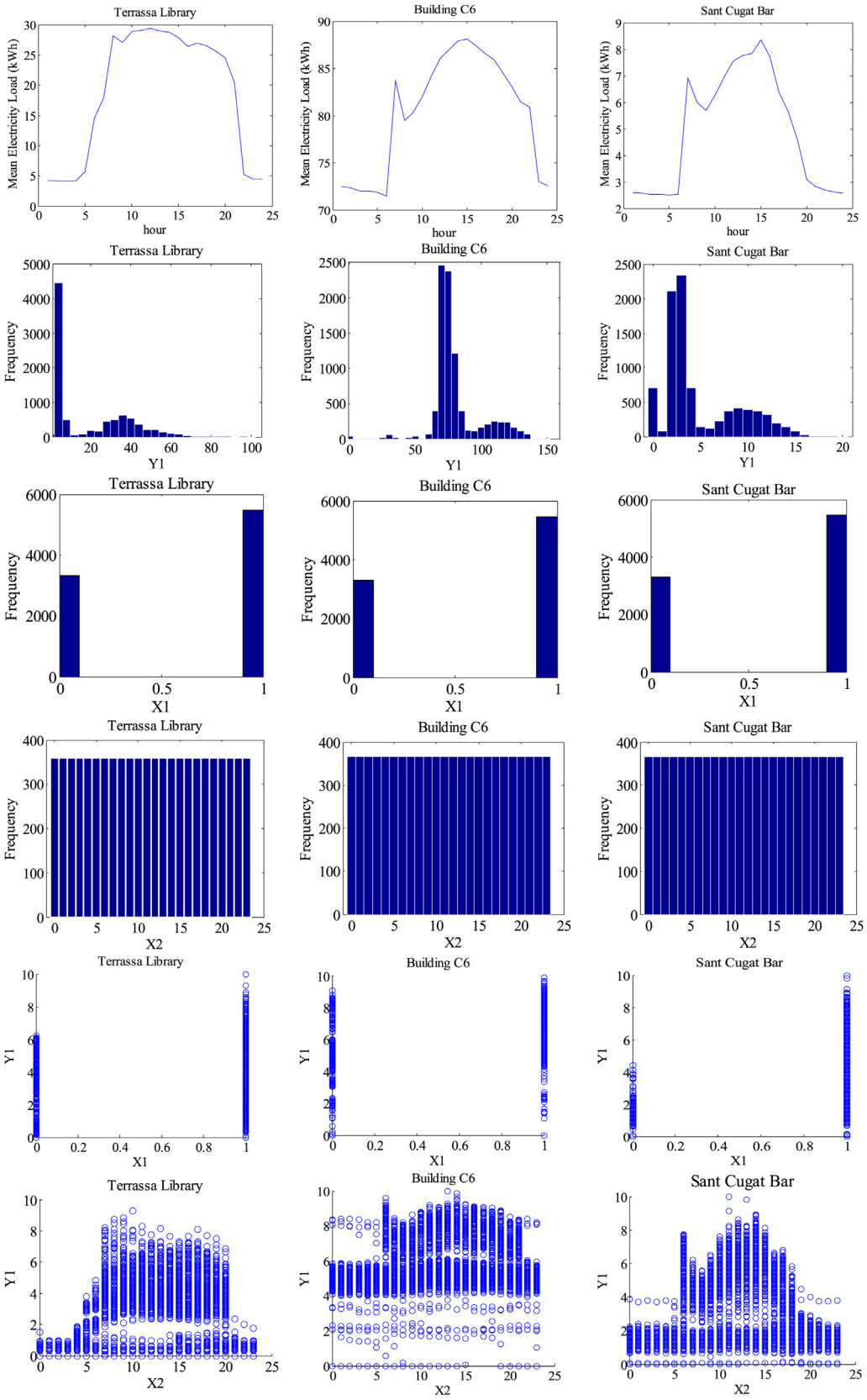

Fig. 6. Mean electricity load (first row), empirical probability distribution (second, third and fourth row) of variables Y1, X1, X2. The units of variable Y1 are kilowatt-hour (kWh). In variable X1 a 1 value represents a working day and a 0 value a non-working day. The units of X2 variable are hours of the day. Fifth and sixth rows present the scatter plots demonstrating visually the relationship between variables X1 and X2 with Y1, in Terrassa Library, Building C6 and Sant Cugat Bar.

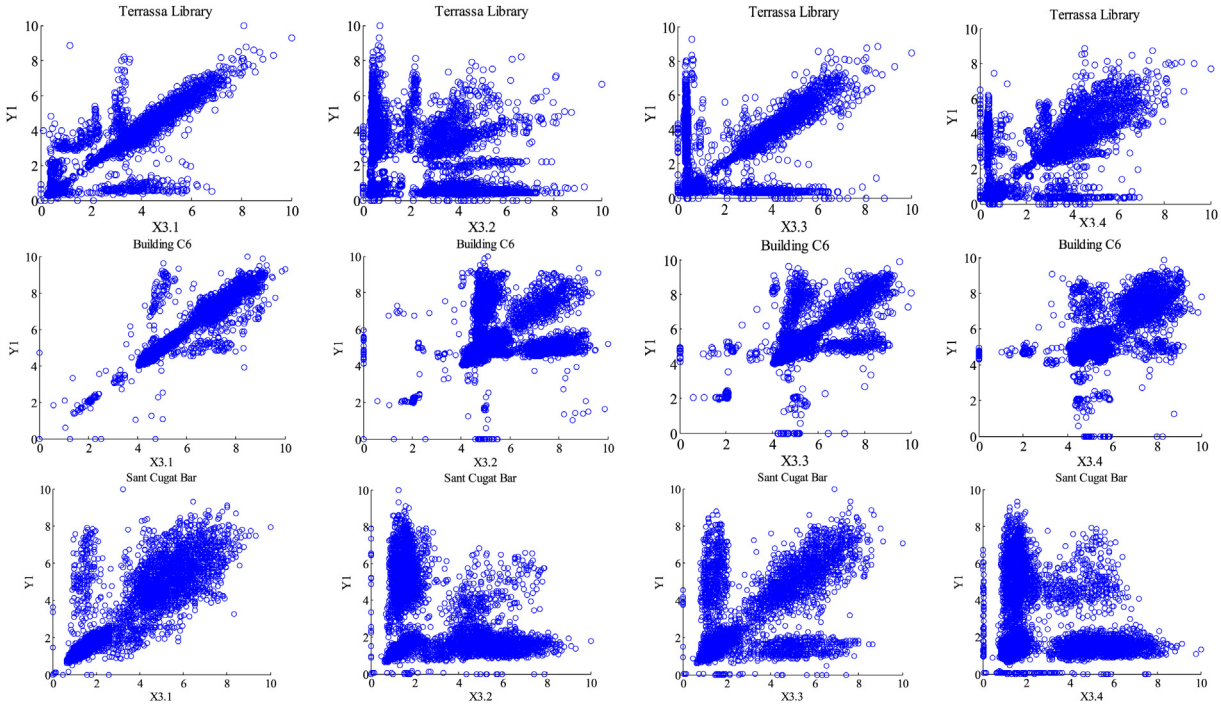

Fig. 7. Scatter plots demonstrating visually the relationship between the input past electricity load consumption variables, X3.1, X3.2, X3.3 and X3.4 with the output variable Y1, when depth is 24 + 24 (see Table 4).

The NMSE is not the only criterion and there are several other commonly used evaluation criteria. The present experiments also use the MAPE (Mean Absolute Percentage Error) to offer a forecasting performance from a multi-dimensional perspective. The reason for this choice is that MAPE can be used to compare the performance on different data sets, because it is a relative measure. However, it has to be highlighted that measures based on percentage errors have the disadvantage of being infinite or undefined if $Y_t = 0$ for any t in the period of interest. MAPE also has the disadvantage that puts a heavier penalty in negative errors than in positive errors [36].

The MAPE, described in equation (8), is used as a second error measure to evaluate the forecasted results.

$$MAPE = 100 * \frac{1}{N} \sum_{t=1}^{N} \left| \left(y_r(t) - y_f(t) \right) \Big/ y_r(t) \right| \quad (8)$$

3.6. Experiments

The work has been divided in three separated experiments described in Fig. 8. Each of the experiments is divided in two stages: the FSP and the model development. Since one of the aims of these experiments is to understand how the model's accuracy is affected by the insertion of new input variables, it has been decided to create three experiments with different number of input variables.

One for each FSP studied. The first experiment only takes into account the four most significant historical electricity load consumption values selected in the FIR FSP. The second experiment includes the past values of the electricity load consumption, besides a binary variable that specifies if it is or not a working day. Therefore, the second experiment has five input variables. Finally, the third experiment includes all five variables of the previous experiment plus the hours of the day variable. As it is shown in Fig. 8, the FSP has been applied to different depths: 24, 72 and 24 + 24.

Those experiments performed with ARIMA only consider the historical electricity load as it is justified in section 2.5. Therefore, ARIMA experiments only take into account 4 input variables. The goal to include this experiment is to give the reader a comparison with a standard statistical technique.

3.7. Model parameters

3.7.1. FIR

With respect to the FIR methodology, the parameters associated to the FSP are mostly addressed in previous sections. However, it is mandatory to set the complexity of the masks parameter, as described in section 2. In this research a complexity of 9 is used, but the mask with higher quality was always the mask that selects four past values of the electrical load consumption. Therefore, the four past consumptions values selected by FIR are taken into account in all the models developed in this work.

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. Is working day variable is binary and, therefore, it has been discretized into two classes.

3.7.2. Random forest

Breiman has proved that for both random forest's classification and regression, the generalization error converges to a limit when the number of trees becomes larger [21]. The number of trees for the Random Forest algorithm has been selected to be 20 trees. In Fig. 9, it is shown that growing around 20 trees is a good compromise with the out-of-bag regression error [24] which is almost stabilized. This parameter directly affects the computational cost of the algorithm. In Breiman's original implementation of random forest algorithm, each tree is trained on about 2/3 of the total training data. As the forest is built, each tree can be tested (similar to leave one out cross validation) on the samples not used

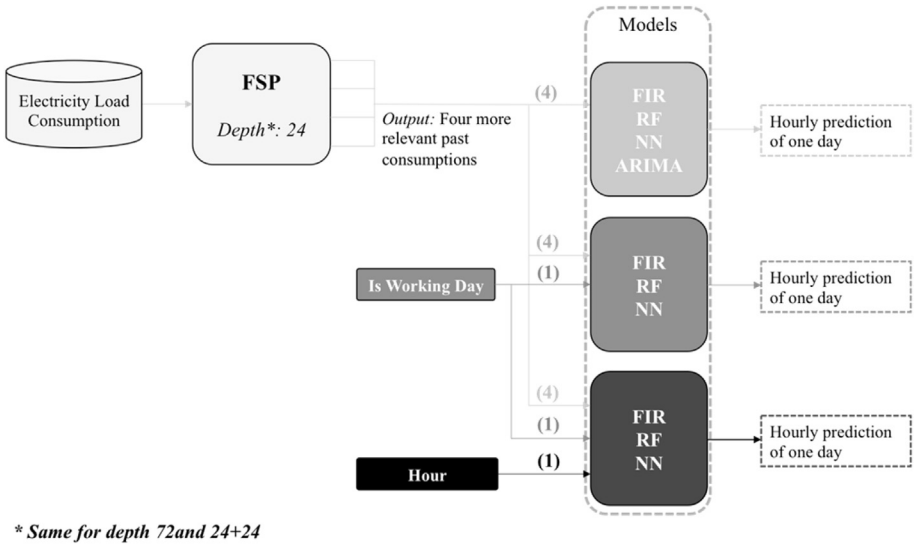

Fig. 8. Scheme of the experiment when depth is equal to 24. The same scheme applies to depths 72 and 24 + 24.

in building that tree. This is the out of bag error estimate − an internal error estimate of a random forest as it is being constructed.

There are different approaches regarding the stopping criteria. The most used is to stop when each node contains less than *n* data points. It has been decided to set the minimum number of data points per tree leaf in 20. Thus, splits will stop when for each tree leaf (node) there are equal or less than 20 input values.

Another important parameter to consider within the RF algorithm is the number of variables for each decision's split selected at random. If this argument is set to any value smaller than the total number of input variables the Breiman's RF algorithm is invoked [22]. In this research, after some experimentation it has been decided to set this parameter to 1/3 of the number of variables, i.e. 2. It has been proven that if all the variables are used in each decision's split, the prediction errors of test data are larger.

3.7.3. Artificial neural network

As described previously, the network used in this research is a three layer feed forward network. The inputs are connected to the output via a middle layer. Additionally, in this work the NN uses a back propagation training function that uses Jacobian derivatives, in concrete a Levenberg–Marquardt back propagation training function.

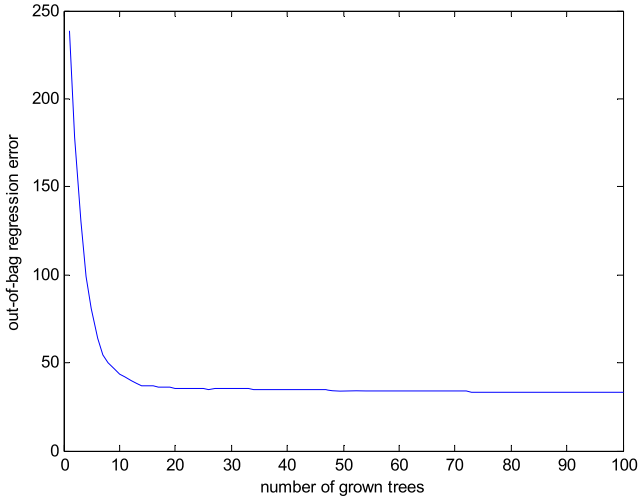

Fig. 9. Out-of-bag (OOB) error vs. number of trees grown in the Random Forest.

Regarding the number of hidden layers in the NN model, after several experiments, it has been concluded that no substantial improvement is achieved with more than one hidden layer, whilst time to build up the model increases significantly. Therefore, this parameter remains as one. Additionally, the suitable number of neurons in the hidden layer is selected by calculating the NMSE (see equation (6)) for each model from 1 to 30 neurons, and selecting the model and number of neurons with the best performance. The search has been defined up to 30 neurons, since with more neurons the cost of time increases significantly and the performance does not improve.

3.7.4. ARIMA

The parameters used for the experiments are ARIMA (4,0,1). Therefore, the resulting equation and parameters to determine are:

$$y_t = a_0 + a_1{*}y_{th1} + a_2{*}y_{th2} + a_3{*}y_{th3} + a_4{*}y_{th4} + b_0 + b_1{*}e_{th1}$$

where y_{th1} *to* y_{th4} are the four most important previous consumptions shown in Table 4, while et_{h1} is the difference between the previous real consumption $real_{t-1}$ and the predicted previous consumption y_{t-1}.

The coefficients a_0, a_1 … a_n *and* b_0, b_1 are computed by the ML (maximum likelihood estimator) [37] which is the most common used method for finding estimators in statistics.

4. Results

Table 7 shows the results obtained by each methodology for the three different depths studied: 24, 72 and 24 + 24, and for the three experiments with different number of input variables, i.e. only electricity load past consumptions (4 input variables); past consumptions plus is working day variable (5 input variables) and past consumptions plus is working day and hour of the day variables (6 input variables).

The main reason why high error values appear in some cells of Table 7 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.

An example of the prediction error distribution is shown in Fig. 10. The distribution is performed using the boxplot representation [30]. Each boxplot representation contains the prediction errors based on the experimental settings previously explained. In

Table 7
Average NMSE of the 35 test days representing an entire season year, in the Library of the ETSEIAT faculty in Terrassa, the Bar ETSAV in Sant Cugat and Building C6 in Barcelona, obtained by means of RF, NN and FIR.

	Terrassa Library											
	Depth: 24				**Depth: 72**				**Depth: 168 (24+24)**			
	RF	**NN**	**FIR**	**ARIMA**	**RF**	**NN**	**FIR**	**ARIMA**	**RF**	**NN**	**FIR**	**ARIMA**
	Input Variables: 4											
Average NMSE	0.828	1.283	1.071	0.792	0.826	1.096	0.786	0.690	0.556	0.674	0.396	0.595
Average MAPE (%)	225.634	265.339	173.128	191.691	204.172	203.783	163.085	187.010	154.873	190.382	55.691	140.376
	Input Variables: 5											
Average NMSE	0.601	0.380	0.362		0.551	0.478	0.278		0.582	0.521	0.367	
Average MAPE (%)	161.451	95.091	21.690		157.506	100.236	32.107		144.956	78.249	40.781	
	Input Variables: 6											
Average NMSE	0.442	0.590	0.404		0.417	1.442	0.166		0.443	0.495	0.103	
Average MAPE (%)	141.597	170.308	61.536		134.765	60.167	14.811		139.797	88.759	15.867	

	Sant Cugat Bar											
	Depth: 24				**Depth: 72**				**Depth: 168 (24+24)**			
	RF	**NN**	**FIR**	**ARIMA**	**RF**	**NN**	**FIR**	**ARIMA**	**RF**	**NN**	**FIR**	**ARIMA**
	Input Variables: 4											
Average NMSE	0.678	0.705	0.875	0.660	0.701	1.134	0.740	0.562	0.357	0.473	0.609	0.495
Average MAPE (%)	43.842	98.780	123.615	112.463	46.727	67.796	94.198	131.236	29.139	33.080	95.958	89.224
	Input Variables: 5											
Average NMSE	0.230	0.261	0.259		0.277	0.408	0.277		0.147	0.160	0.255	
Average MAPE (%)	20.470	22.320	80.556		21.167	22.997	46.626		20.000	90.492	65.934	
	Input Variables: 6											
Average NMSE	0.183	0.134	0.235		0.166	0.169	0.100		0.113	0.127	0.135	
Average MAPE (%)	16.539	24.436	15.278		19.159	31.987	13.950		17.339	36.509	16.013	

	Building C6											
	Depth: 24				**Depth: 72**				**Depth: 168 (24+24)**			
	RF	**NN**	**FIR**	**ARIMA**	**RF**	**NN**	**FIR**	**ARIMA**	**RF**	**NN**	**FIR**	**ARIMA**
	Input Variables: 4											
Average NMSE	0.852	1.003	0.923	0.778	0.791	0.684	0.797	0.800	0.727	1.287	0.738	0.754
Average MAPE (%)	26.161	26.175	10.373	21.186	25.731	25.344	9.789	19.841	24.786	30.653	8.946	20.064
	Input Variables: 5											
Average NMSE	0.526	0.661	0.524		0.518	0.805	0.404		0.535	0.747	0.576	
Average MAPE (%)	22.106	21.562	6.710		21.874	26.012	6.136		22.056	26.324	7.565	
	Input Variables: 6											
Average NMSE	0.462	1.230	0.402		0.434	1.313	0.183		0.454	0.941	0.801	
Average MAPE (%)	18.183	36.479	5.525		21.757	39.983	4.499		21.622	30.629	8.720	

this representation we do not include those experiments with 5 and 6 input variables, because ARIMA could not be compared, as it is using only the 4 most important past consumptions. Each boxplot contains the results from 35 test days with 3 different configurations (105 test days in total for each technique except ARIMA, which contains only 35). For example, the boxplot representation *NMSE in Library Terrassa with Depth 24* aggregates the results of all the experiments performed in Terrassa Library with 4, 5 and 6 input variables with depth 24, thus *35x3* prediction errors. Fig. 10 only includes the results with Terrassa Library. However, the results and insights, which are highlighted later on, are very similar to the other two buildings.

Fig. 11 shows the evolution of the prediction error versus the number of variables (left column) and versus the depth (right column) for each of the three buildings studied. Notice that the errors shown in the left hand column are the average of the errors obtained from the aggregation of the results with the same depth, i.e. 24, 72 and 24 + 24. And the errors of the right hand column are the average of the errors obtained from the aggregation of the results with the same number of variables, i.e. 4, 5 and 6.

Fig. 12 shows the real hourly electricity consumption of the buildings that have been used for these experiments in four different days. Depth and Number of Variables selected for each methodology are selected considering graphs of Fig. 11, when the minimum error is achieved. Table 8 summarise these parameters:

The autumn, spring and summer days correspond to the 23/11/2011, 09/05/2012 and 18/09/2012 respectively, which were normal working days, while the winter day corresponds to 19/12/2012, which belongs to a weekend day.

Several insights can be obtained when analysing the results achieved, in the experiments carried out in this research. They are summarized in the following points:

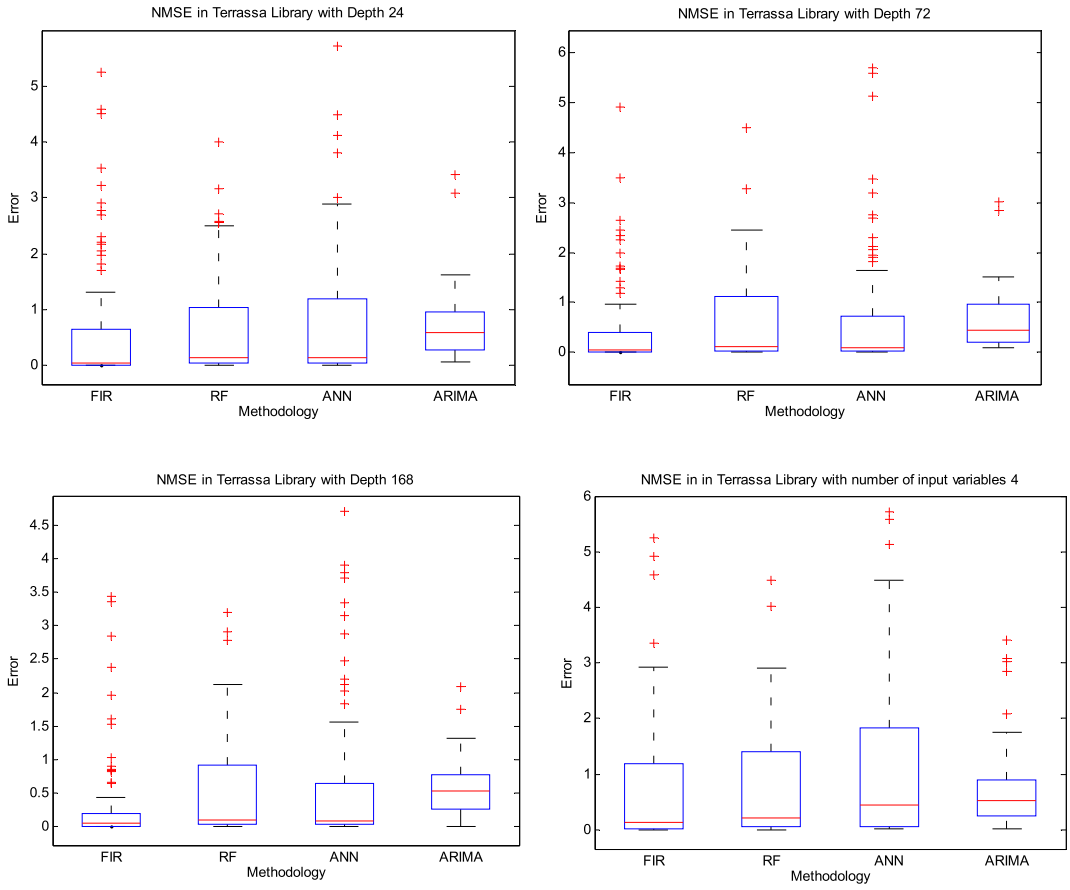

Fig. 10. NMSE distribution for each technique.

1) When the number of input variables is 4, i.e. only the past values of the electricity load consumption are used, ARIMA performs, on average, slightly better than FIR, NN and RF (left hand column of Fig. 11) considering the NMSE as evaluation criteria.
2) When the four methodologies can be compared considering the NMSE (see the three rows with *Average NMSE* of Table 7 where 4 input variables are used), ARIMA is the technique that obtains better performance, followed by RF, FIR and NN. ARIMA outperforms the other methodologies in 6 out of 9 cases, RF in 2 cases, while FIR and NN in only 1 case each.
3) When the four methodologies can be compared considering the MAPE (see the three rows with *Average MAPE* of Table 7 where 4 input variables are used), FIR is the technique that obtains better performance, followed by RF. FIR outperforms the other methodologies in 6 out of 9 cases, while RF in 3 cases.
4) When additional information is added, in general FIR has better performance than the other two methodologies (Notice that with 5 and 6 input variables ARIMA is not included in the comparison).
5) When only FIR, RF and NN are compared considering the NMSE (see the three rows with *Average NMSE* of Table 7 where 5 and 6 input variables are used), FIR is the technique that obtains better performance, followed by RF and NN. FIR outperforms the other methodologies in 11 out of 18 cases, RF in 5 cases, while NN in only 1 case.
6) When only FIR, RF and NN are compared considering the MAPE (see the three rows with *Average MAPE* of Table 7 where 5 and 6 input variables are used), FIR is the technique that obtains better performance, followed by RF. FIR outperforms the other methodologies in 15 out of 18 cases, while RF 3 cases.
7) The prediction errors of all the methodologies decrease considerably when variable is working day is added to the set of input variables that the model can use. As it becomes clear in the left hand plots of Fig. 10. However, when the number of input variables is increased to 6 (including the hour of the day variable), not all the errors decrease. NN errors only decrease in Sant Cugat Bar, while for FIR and RF decreases in all buildings, but less than when the number of input variables is increases up to 5. This suggests that the hourly consumption pattern is already captured by the past consumption variables. And thus, in some cases, the sixth variable could be removed or not taken into account.
8) Furthermore, the prediction error does not follow a clear pattern regarding the depth. However, it is observed that with the increase of the depth the errors tend to decrease. On average, the optimal depth value is 24 + 24 (168) for all buildings and methodologies. The more depth, the more visibility and better results (see the right hand side of Fig. 11). RF and ARIMA methodology remain quite stable with changes in the depth though.
9) The distribution of the prediction errors in Fig. 10 shows how the median of the prediction errors decreases when the depth increases. Moreover, FIR, RF and NN always have a lower median than ARIMA. However, this last one does not have as many error prediction out of Q3+1.5*RIC [30] as FIR, RF and NN. The main reason is that soft computing and machine learning techniques adapt better to consumption changes when they predict, achieving low prediction errors (high accuracies). This can be seen in the low median that they have. Nevertheless, when unexpected changes are predicted and they fail, the errors are very high. ARIMA is on

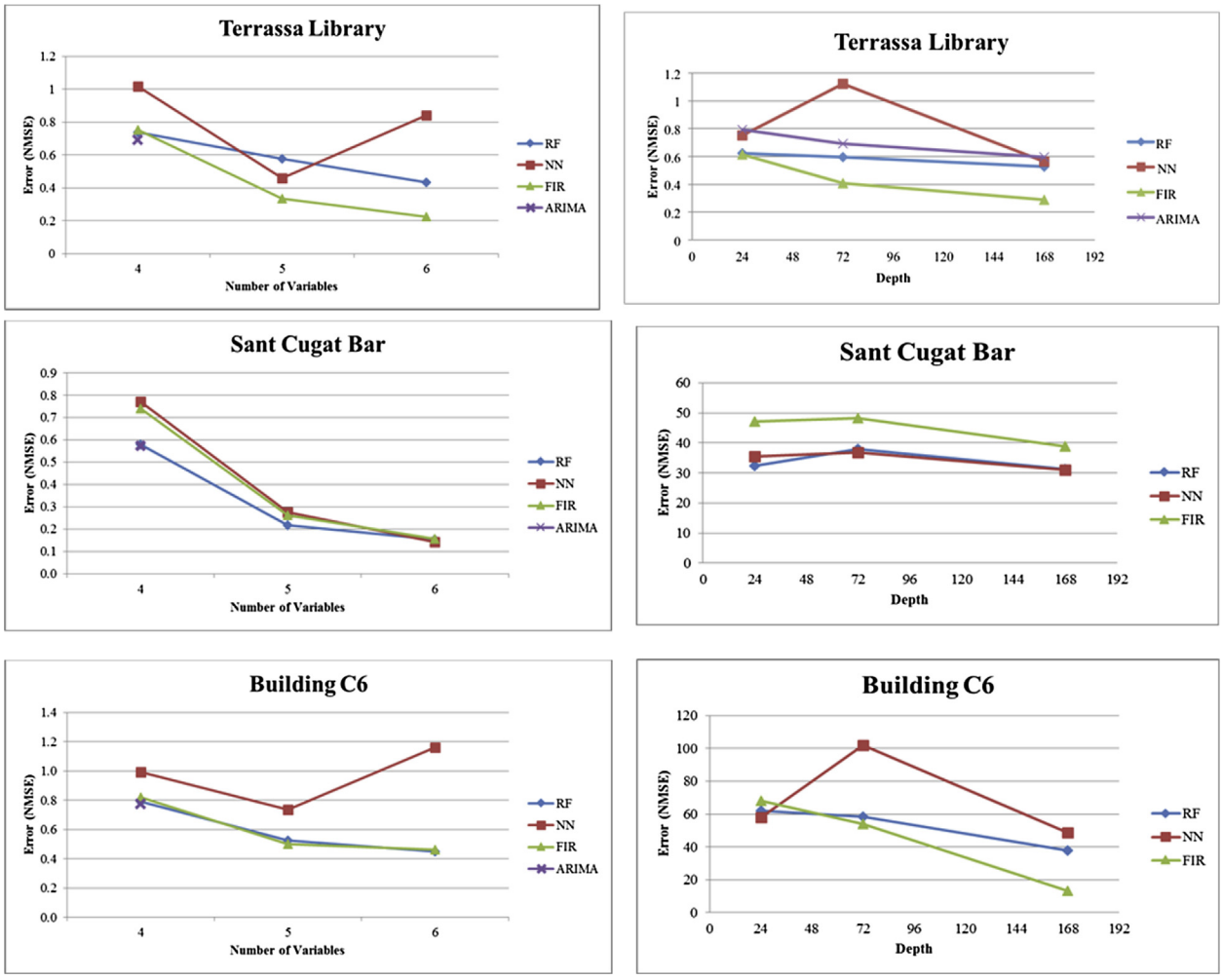

Fig. 11. The left column represents the evolution of the prediction error with respect to the number of input variables (average of all predictions with depths 24, 72 and 24 + 24) and the right column represents the evolution of the prediction error with respect to the depth (average of all predictions with 4, 5 and 6 variables).

average more conservative and its predictions are not very accurate, although prediction errors are not so high.

10) In general, FIR, RF and NN methodologies forecast quite well the electricity load consumption, as shown in Fig. 12. Its predictions follow the real shape of the consumption curve, except for the winter day that corresponds to Sunday when the electricity consumptions are very low compared to the average consumption. That is due to the fact that all the buildings belong to the academic domain and, therefore, they are closed on Sundays. Notice that in Spain the University libraries close on Sundays as a consequence of the crisis. In this case FIR is the methodology that better performs.
11) Unlike the soft computing and machine learning methodologies, ARIMA performs always a similar shape of the electricity consumption curve. This is due to the nature of the technique, which is built-up by multiple linear regressions.
12) The summer test days are shown in the last row of Fig. 12. All the methodologies obtain good prediction performance, except for a specific day (18/09/2012) in Sant Cugat bar where FIR is not able to obtain a good prediction (right hand plot of Summer Day row of Fig. 12). After analysing possible reasons of this low performance where all methodologies are using the same depth (24 + 24), we have arrived to the conclusion that FIR is giving more importance to the 24 past values one week before than RF and NN. The hourly consumption of this test day one week before is very low due to the fact that it was holydays at Sant Cugat campus.
13) The lower prediction errors for all building are obtained with FIR, as shown in Table 7.
14) There is not a clear technique that obtains better predictions than the others. On average, there are more occasions that FIR performs better than the other methodologies. Nevertheless, in some cases RF, ARIMA and NN are better than FIR, thus, this suggest that simple cooperative systems could fit with electricity load problems.
15) With respect to the computational cost, the four methodologies are very fast when training of the model is performed (less than 10 s for a training set of a year hourly data) as well as prediction (less than 0.5 s, achieved with an Intel Core2 Duo CPU E8400 with a RAM of 4 GB). On the contrary, the process to obtain the most important past consumptions within the FIR FSP increases exponentially with the depth. However this is an offline process.

The results outlined in the previous points make clear that the two soft computing and machine learning methodologies chosen in this research are able to obtain good results for the task at hand.

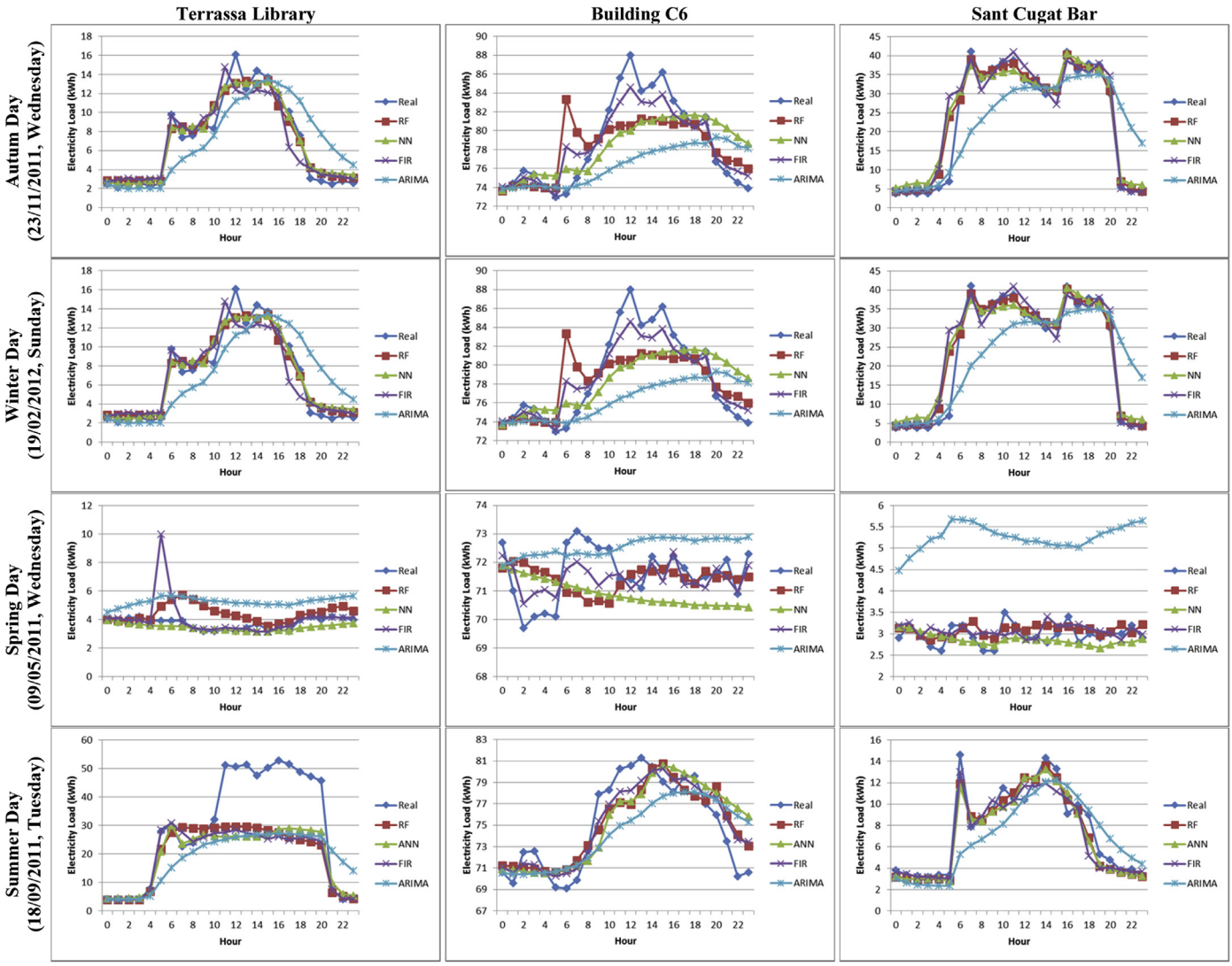

Fig. 12. Real energy consumption values and consumption predictions obtained by the RF, NN and FIR methodologies, in four different test days (in each season of the year). The y-axis represents the consumption in kWh and x-axis the time of the day in hours. The RF, NN and FIR models correspond to those that use 6 input variables and with depth of 24 + 24 h.

That is why we think that it could be interesting to work with voting strategies to deal with electricity load problems. On the other hand, it is also clear that in general terms the performance of FIR is higher. We think that this is due to the intrinsic characteristic of FIR that extracts knowledge from the past information and keeps it in the pattern rules, as part of the model. FIR models are synthesized rather than trained, since is a non-parametric methodology, which speeds up the modelling phase in comparison with other inductive modelling techniques.

It is important to highlight the aim of this work: to provide fast and reliable prediction techniques for accurate forecasting of the hourly consumption. There are many researches where different methodologies are tested but they are focused on a specific building and/or using several *input variables*. However, we provide three reliable methodologies performing accurate forecasting in buildings with different profiles and using available variables such as past values of the electricity load consumption, as well as the information of the hour of the day and if it is a working day or not.

Table 8
Optimal Depth and Number of Variables (NoV) for each building based on the results of the 35 test days.

Building	RF		NN		FIR		ARIMA	
	Depth	NoV	Depth	NoV	Depth	NoV	Depth	NoV
Terrassa Library	168	6	168	5	168	6	168	4
Sant Cugat Bar	168	6	168	6	168	6	168	4
Building C6	168	6	72	5	72	6	168	4

5. Conclusions

In this work different AI methodologies, i.e. RF (Random Forest), NN (Neural Networks) and FIR (Fuzzy Inductive Reasoning), are proposed to perform short-term electric load forecasting (24 h). These approaches could help inside the future Smart Grid framework providing accurate and quick predictions of electricity consumption in different type of buildings, allowing for a better distribution plan.

In order to demonstrate the scalability of the models in any type of building, three functional zones of the Technical University of Catalonia are used in the experiments. They all have different profiles of usage and locations, thus, affecting different climatology, consumption patterns, schedules and working days.

The designed experiments are divided into two stages: 1) an FSP based on Entropy, common to all three methodologies plus a fourth one, ARIMA, which helps to compare them with a traditional time series forecasting statistic technique, and 2) a FIR, RF, NN or ARIMA model training process. In the FSP stage, three depths are studied: 24 h, 72 h and 24 + 24 h. In the model training stage, different sets of input variables are studied. One of the aims of these experiments is to understand how the model's accuracy is affected by the insertion of new input variables without increasing the computational cost in terms of time. To do so, a study of the prediction errors versus the number of most important variables and depth of the past values in the FSP has been performed, pointing out that more than four variables for past consumptions and a depth higher than 80 past hours, would increase too much the time to perform the FSP.

Three experiments with different number of input variables for each FSP studied have been created. The first experiment only takes into account the four most significant historical electricity load consumption values selected in the FSP. The second experiment includes the past values of the electricity load consumption plus a binary one that corresponds to the variable that specifies if it is or not a working day. The third experiment includes all the previous input variables plus the hours of the day variable.

Based on this study, FIR is the methodology that performs a better forecast followed by the RF and the NN. On average, there are more occasions that FIR performs better than the other methodologies. However, in several cases RF is better than FIR, thus, this suggest that voting strategies could be good approaches to deal with electricity load problems. In general, AI methodologies adapt better to consumption changes when they perform the predictions, following the real shape of the curve, detecting better the peaks and achieving very low prediction errors. With regards to ARIMA, it is a more conservative methodology, which does not produce high errors but the accuracy is far from FIR.

It can also be concluded that the prediction errors of all the methodologies decrease considerably when is working day is added to the set of input variables. On average, the optimal depth value is 24 + 24 for all buildings and methodologies. The more depth, the more visibility and better results. Finally, as for the computational cost, all the three methodologies are very fast in order to obtain the model (less than 10 s for a training set of a year hourly data) and perform a prediction. On the contrary, the FSP increases exponentially with the depth and number of past values selected. However this is an offline process that could be performed for instance in the cloud.

Several future work arises from the results of this journal. For instance, how the performance of the predictions would be affected by the granularity of the data, i.e. gathering data every 1, 5 or 15 min, and which strategy should be follow in the FSP due to the curse of dimensionality. Or how to deal with those days identified in Fig. 10 that produce a high prediction error. Another future work will be to include a more robust prediction process in the FIR methodology, to deal with missing values in a more reliable way. Last but not least, these models could be embedded, for instance, in a second generation of smart meters where they could generate on-site forecasting of the consumption and/or production in the next hours, or even trade the excess energy with other smart meters. There are already European projects such as SCANERGY,[6] studying different energy exchange strategies and pointing out that smart meters with such functionalities would unlock the potential of these trading strategies.

[6] FP7 framework's Marie Curie Industry-Academia Partnerships and Pathways (IAPP).

References

[1] http://www.nist.gov/smartgrid/; [accessed July, 2014].
[2] http://greenenergyreporter.com/renewables/cleantech/china-to-invest-7-3-bln-in-smart-grid-projects-in-2010/; [retrieved December, 2011].
[3] Arenas-Martínez M, Herrero-Lopez S, Sanchez A, Williams J, Roth P, Hofmann P, et al. A comparative study of data storage and processing architectures for the smart grid. In: Proceedings of the first IEEE international conference on smart grid communications; 2010. p. 285–90.
[4] Catalina T, Virgone J, Blanco E. Development and validation of regression models to predict monthly heating demand for residential buildings. Energy Build 2007;40:1825–32.
[5] Aydinalp M, Ugursal VI, Fung AS. Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks. Appl Energy 2002;71:87–110.
[6] Tso Geoffrey KF, Yau Kelvin KW. Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. Energy 2007;32:1761–8.
[7] Khamis MFI, Baharudin Z, Hamid NH. Electricity forecasting for small scale power system using artificial neural network. In: Power engineering and optimization conference (PEOCO) 5th international; 2011. p. 54–9.
[8] Marvuglia A, Messineo A. Using recurrent artificial neural networks to forecast household electricity consumption. Energy Procedia 2012;14:45–55.
[9] Tranchita C, Torres A. Soft computing techniques for short term load forecasting. In: Power systems conference and exposition, vol. 1; 2004. p. 497–502.
[10] Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on machine learning (ICML2006); 2006. p. 161–8.
[11] Nesreen AK, Amir AF, Neamat G, Hisha EH. An empirical comparison of machine learning models for time series forecasting. Econ Rev 2010;29:594–621.
[12] http://www.forecasters.org/data/m3comp/m3comp.htm.
[13] Tso GKF, Yau KKW. Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks. Energy 2007;32:1761–8.
[14] Härdle W, Simar L. Canonical correlation analysis. Appl Multivar Stat Anal 2007:321–30.
[15] Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat 2010;2:433–59.
[16] Nebot A, Mugica F, Cellier F, Vallverdú M. Modeling and simulation of the central nervous system control with generic fuzzy models. Trans Soc Model Simul 2003;79(11):648–69.
[17] Law AM, Kelton WD. Simulation modeling and analysis. 2nd ed. New York: McGraw-Hill; 1991. p. 759.
[18] Acosta J, Nebot A, Villar P, Fuertes JM. Optimization of fuzzy partitions for inductive reasoning using genetic algorithms. Int J Syst Sci 2007;38(12):991–1011.
[19] Klir J, Elias D. Architecture of systems problem solving. 2nd ed. New York: Plenum Press; 2002.
[20] Escobet A, Nebot A, Cellier FE. Visual-FIR: a tool for model identification and prediction of dynamical complex systems. Simul Model Pract Theory 2008;16:76–92.
[21] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[22] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.
[23] Kleinberg E. An overtraining-resistant stochastic modeling method for pattern recognition. Ann Statistics 1996;24(6):2319–49.
[24] Li Y, Wang S, Ding X. Person-independent head pose estimation based on random forest regression. In: 2010 17th IEEE international conference on image processing (ICIP); 2010. p. 1521–4.
[25] McCulloch W, Pitts W. A logical calculus of ideas immanent in nervous activity. Bull Math Biophys 1943;5(4):115–33.
[26] Rosenblatt F. The perceptron: a probalistic model for information storage and organization in the brain. Psychol Rev 1958;65(6):386–408.
[27] Alon I, Qi M, Sadowski RJ. Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. J Retail Consum Serv 2001;8:147–56.
[28] White H. Connectionist nonparametric regression: multilayer feed forward networks can learn arbitrary mappings. Neural Netw 1990;3:535–49.
[29] Chow TWS, Cho SY. Neural networks and computing: learning algorithms and applications, vol. 7; 2007. p. 14–5.
[30] Box G, Jenkins G, Reinsel GC. Time series analysis: forecasting and control. 4th ed. 2008. p. 93–102.
[31] McGill R, Tukey JW, Larsen WA. Variations of box plots. Am Statistician 1978;32(1):12–6.
[32] Frigge M, Hoaglin DC, Iglewicz B. Some implementations of the boxplot. Am Statistician 1989;43(1):50–4. http://dx.doi.org/10.2307/2685173. JSTOR 2685173.
[33] R: box plot statistics. 2011 [R manual].
[34] Berthold M, Borgelt C, Höppner F. Data understanding – outlier detection. In: Guide to intelligent data analysis; 2010. p. 63–4. http://dx.doi.org/10.1007/978-1-84882-260-3.
[35] Weron R. Factors affecting load patterns. Case study: dealing with missing values and outliers. In: Modeling and forecasting electricity loads and prices. A statistical approach; 2006. p. 69–70.
[36] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast 2006;22:679–88.
[37] Box G, Jenkins G, Reinsel GC. In: Time series analysis: forecasting and control. 4th ed. 2008. p. 404–6.