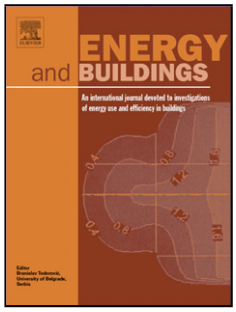

Title: Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes

Author: Joaquim Massana Carles Pous Llorenç Burgas
Joaquim Melendez Joan Colomer

PII: S0378-7788(16)30786-1
DOI: http://dx.doi.org/doi:10.1016/j.enbuild.2016.08.081
Reference: ENB 6980

To appear in: *ENB*

Received date: 9-5-2016
Revised date: 24-7-2016
Accepted date: 27-8-2016

Please cite this article as: Joaquim Massana, Carles Pous, Llorenç Burgas, Joaquim Melendez, Joan Colomer, Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes, <*![CDATA[Energy & Buildings]]*> (2016), http://dx.doi.org/10.1016/j.enbuild.2016.08.081

Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes

Joaquim Massana, Carles Pous, Llorenç Burgas, Joaquim Melendez, Joan Colomer

University of Girona, Campus Montilivi, P4 Building, Girona, E17071, Spain,
{joaquim.massana, carles.pous, llorenc.burgas, joaquim.melendez, joan.colomer}@udg.edu

Abstract

An accurate short-term load forecasting system allows an optimum daily operation of the power system and a suitable process of decision-making, such as with regard to control measures, resource planning or initial investment, to be achieved. In a previous work, the authors demonstrated that an SVR model to forecast the electric load in a non-residential building using only the temperature and occupancy of the building as attributes is the one that gives the best balance of accuracy and computational cost for the cases under study. Starting from this conclusion, a simple, low-computational requirements and economical hourly consumption prediction method, based on SVR model and only the calculated occupancy indicator as attribute, is proposed. The method, unlike the others, is able to perform hourly predictions months in advance using only the occupancy indicator.

Due to the relevance of the occupancy indicator in the model, this paper provides a complete study of the methods and data sources employed in the creation of the artificial occupancy attributes. Several occupancy indicators are defined, from the simplest one, using general information, to the most complex one, based on very detailed information. Then, a load forecasting performance discrimination between the artificial occupancy attributes is realized demonstrating that using the most complex indicator increases the workload and complexity while not improving the load prediction significantly. A real case study, applying the forecasting method to several non-residential buildings in the University of Girona, serve as a demonstration.

Keywords: load forecasting, support vector machines, sensor data,

Mediterranean climate, occupancy indicator.

1. Introduction

In order to build a fair and more sustainable society, new approaches and initiatives have appeared in all areas. Energy resources are limited, and there is the need to generate new technologies and legislation that allows to achieve a certain environmental balance. The Lisbon Treaty [30] and the Kyoto Protocol [10] are examples of legal initiatives that have the aim of reducing consumption and emissions. To reduce the consumption, it is necessary to improve the existing electricity grid making it more efficient and robust. The smart grid, in conjunction with decentralized power generation, could avoid many of the shortcomings of the classical electrical grid.

Thus, to increase the efficiency of the electricity grid, a balance of power generation is required such that there is no waste or lack of resources. Due to the apparition of micro-grids, there is a balance between the generation of power and the users' consumption. Given that buildings are responsible for a large part of the electricity consumption, having tools to predict their consumption is key in the adjustment process. Predicting the consumption of a city is different from predicting the consumption of a building, in that in the case of buildings there is much variability. Disaggregated environments are more difficult to predict. Thus, short-term load forecasting (STLF) methodology is used to reduce the building's consumption since it must deal with non-linearities and noise.

Recent research on energy efficiency in buildings include optimal decisions and an overall improvement in human behaviour, not just technology. The International Energy Agency's Energy in Buildings and Communities Programme (IEA-EBC) has recently completed a project related to strengthening the robust prediction of energy usage in buildings, with the goal of enabling the proper assessment of short and long-term energy measures, policies and technologies. The results of this project are collected in Annex 53 [15]. The analysis methods, developed models and results of Annex 53 were taken as the starting point for other several working areas. In particular, and due to the important effect of occupancy in energy prediction, the IEA-EBC is working on Annex 66 [16]. On this annex they are trying to define and simulate occupant behaviour in a consistent and standard way. Based

on these works, some new proposals have arisen, as is shown in [13]. The ontology represents energy-related occupant behaviour outlined as a DNAS (drivers, needs, actions and systems) framework, providing a systematic representation of energy-related occupant behaviour in buildings. Generally, researchers working on this topic follow a methodology that consists of monitoring, modelling and simulating, such as [13] and [14], as seen in Figure 1.

Figure 1: Technical framework used in occupancy behaviour.

These models are built after monitoring and collecting enough data about occupancy of the building. As stated in [14], this data is obtained from observational studies, occupant surveys and interviews, laboratory studies and unresolved issues in occupant monitoring, such as contextual factors. The occupancy models take into account the actions that occupant can do on the building, such as open the light, close the window, track or predict the occupant movements, and so on. It can be seen that the building must be sensorized to some extent to have this information available, a fact that is not always possible.

Although computing technology continues to develop, some forecasting models training on databases with dozens of attributes and millions of instances, may lead to high computational cost. Therefore, reducing the database is still necessary, always taking care to ensure that performance does not deteriorate. Most of the papers that propose the use of STLF methods in non-residential buildings often use weather data and, in some cases, occupancy information. Other works, such as [22], [17] and [19], conduct comparison studies using similar models and arrive to a different conclusion, selecting other model as a best approach. The type of building and the test and training conditions can greatly affect the results. So, it is important to study different type of models in order to choose the best option in each case. According to [24], a model predicting consumption with minimal instances using support vector regression (SVR) with temperature and occupancy attributes provides excellent results for our buildings under study.

Obtaining predictions of temperature, in order to know the temperature of a particular place, is normally possible, although acquiring information of future occupancy remains difficult. In [24], occupancy information collected from passive infra-red (PIR) sensors was used. However, this information is not available in advance. The non-residential buildings usually dispose of

work or scholar schedules, or other information about their occupancy. A technique designed to generate this information beforehand is needed. The goal is to obtain a model that is not dependent of any information unavailable months ago, such as previous consumption or temperature. This model can perform consumption predictions months ahead. Perhaps resulting accuracy level of the model may not be as good as the other works in this topic, but this is only a first step in this new direction.

The aim of the work is to test the load forecasting accuracy using several occupancy indicators. It is not centred on occupant behaviour modelling, but estimating the occupancy is necessary, as it is one of the main factors that contributes to the accuracy of the SFTL. Concerning the occupancy estimation, we deal with buildings that are poorly sensed. That means there is not information about occupant actions, such as open/close the window, switch on/off the lights or plug a device, even if the actions are taken. There is information about scholar and working schedules, classrooms dimensions, expert knowledge, etc. Furthermore, there is only one of the buildings under study having sensors to estimate the amount of occupants inside the building by means of PIR sensors. Due to this limitation, several occupancy indexes have been defined using the available information.

In the first part of the paper, artificial occupancy indicators for the buildings are generated using different techniques and information available in advance such as academic calendars and work schedules. Then, SVR model is trained to forecast the consumption of the respective buildings, using these indicators of occupancy. Subsequently, an analysis of the relationship between the forecasting performance and the workload based on occupancy indicators, is performed. The idea is to show that there is a balance point in the artificial occupancy indicators, between forecast accuracy and workload. From a certain point on, increasing the complexity of the indicator does not improve significantly the prediction.

The paper starts with related works and follows with back-ground material. Then, the dataset is explained. This is followed by a presentation of the methodology, where the several occupancy indicators are defined, and the test process explained. Next, the results are presented and the method is discussed. Finally, conclusions are shown.

2. Related works

There have been a large number of papers on the topic of STLF with regard to residential and non-residential buildings. The non-residential buildings are basically malls, schools, universities, hospitals and offices. Assuming that the use of information concerning the occupancy of buildings is key for improving prediction, the present state of the art focuses on the following topic: STLF in non-residential buildings based on occupancy data.

In the present state of the art, the advantages and disadvantages of the several methods associated with using the building's occupancy information in a prediction model are evaluated. The methods can be grouped into 4 blocks, as seen in Table 1.

Method	**Sources**	**Works**
Calendar	Day types, months, etc.	[2] [17] [28] [29] [37] [5] [31] [7]
Schedule	Work, student or use schedules.	[20] [6] [21]
Sensors	Motion, CO_2, noise, light, etc.	[8] [18] [24] [27] [33] [23] [26]
Expert knowledge	Surveys, interviews or inspections.	[25] [35] [19]

Table 1: Occupancy related methods.

In the first block, there are eight works that use calendar information. The first [2], is the case of a campus in Los Angeles that uses temperature and occupancy information, based on calendar data such as day of the week and holidays, with a regression tree model. In the paper [17], based on synthetic data and a non-residential building located in Athens (Greece), using meteorological data including temperature, solar flux, relative humidity and wind speed and the profile of the days of the week, the consumption is predicted using an ANN model. The work [28] and [29], in the campus of the University of Deusto (Spain), use weather data such as relative humidity, precipitation, temperature, wind speed and wind direction in conjunction with the use of types of day comprising Saturdays, working and non-working days using AR, ANN and SVR principally. The work [37], with regard to an office building in Hong Kong, uses weather data including temperature, solar radiation and relative humidity and also takes into account if it is a weekday

or a weekend using an ANN. In [5], an ANN is trained to predict the consumption of a commercial office building in Iowa (USA) using weather data such as precipitation provability, rain indicator, outdoor dry-bulb temperature, outdoor relative humidity, wind speed and sky condition in conjunction with the use of day type indicator. The work [31] proposes a non-linear autoregressive model with exogenous inputs to forecast the load in a college campus in Texas (USA) employing weather variables including temperature, relative humidity and calendar information such as hour, day of the week or month. The paper [7] presents an ANN based on indoor and outdoor temperature and relative humidity and occupancy data including day type in a supermarket in UK.

In the second block, where schedules are used, the work [20] predicts the consumption of the university library in Zhejiang (China) using temperature data and an index of occupancy based on the opening schedule of each of the rooms of the library using a fuzzy inference system. The paper [6], a commercial building in Iowa (USA), uses SVR and ANN based on weather data such as outdoor air dry bulb temperature, outdoor air relative humidity, outdoor air flow rate, diffuse solar radiation rate, direct solar radiation rate, zone air temperature, zone air relative humidity and zone thermostat cooling set point temperature and occupancy data including schedules of building equipment, building light and HVAC operation. The work [21] proposes a model predictive control to forecast the consumption in a simulated commercial building (Energy plus) using meteorological data such as outdoor air temperature, indoor temperature and solar radiation and an equipment schedule ratio.

In the third block, there are works that employ occupancy information through the collection of sensor data. The work [8], in the case of the Research Centre in Rome (Italy), involves meteorological data such as temperature and solar radiation, and creates occupancy indicator counting the number of people who check-in using a card, and then models using an autoregressive integrated moving average, ANN and Naive Bayes. The paper [18], deals with an office building in Hong Kong involving weather data such as outdoor temperature, relative humidity, rainfall, wind speed and global solar radiation and an occupancy attribute created using the hourly total power consumption of the primary air unit, and uses an ANN to create the model. The work [24], in the University of Girona, uses temperature and occupancy data collected with PIR sensors, using MLR, ANN and SVR models. In [27], an ANN is used in conjunction with sensor data such as parking and

building occupancy in the campus of the university of Lisbon (Portugal). In [33], the consumption of an office building in Sweden is forecasted based on weather data such as indoor temperature, outdoor temperature, daylight level, solar radiation and wind speed and PIR sensor data using an MLR. The case [23], the electrical load of an sports hall in Finland is predicted using autoregressive models based on meteorological data comprising indoor and outdoor air temperatures and sensor data such as CO_2 measurements. The work [26] presents an autoregressive integrated moving average model that uses outdoor temperature and sensor data such as contact closure, PIR, CO_2 and network activity sensors to predict the consumption in an office building in Ontario (Canada).

In the fourth block, there are works that use expert knowledge such as inspections or surveys to collect information related to occupancy. The case [25], the Administration building of the University of Sao Paulo (Brazil), uses weather data and an attribute related to occupancy, generated performing expert inspections in order to describe the use and the features or the internal loads such as lighting and computers, with an ANN model. In [35], a fast-food restaurant in Cyprus, an autoregressive model based on representative indicators per energy end use of the building such as lighting, kitchen, and refrigerators is used to predict the consumption. The paper [19] proposes an ANN to predict the consumption of 19 subway stations in Hong Kong using outdoor temperature and relative humidity and expert information such as area of concourse, area of platform, shops area, plant room area, staff accommodation area and weekly amount of passengers.

All these methods provide proper results but present shortcomings. There is a demand for methods based on data available in advance which has the ability to perform hourly long-term predictions, predict with few attributes which means low computational cost and do not require continuous sensor data which is economic. In short, there is a need for simple, economic and fast systems that predict the load accurately. The main shortcomings of these methods are as follows:

- The works that employ meteorological data including temperature and solar radiation need weather forecasts which are not always available. In addition, weather forecasts can only be obtained for few days ahead and contain uncertainties. A method without weather data is needed.

- In the case of methods that use occupancy sensor data, there is no data

available in advance, so there is no data to predict. Artificial occupancy data is needed.

- Most of these methods are able to forecast consumption just a few days ahead, but cannot do so months in advance. Therefore a method to predict consumption several months in advance is needed.

- The expert knowledge is not always available and contains uncertainties. A repeatable and objective method is needed.

Therefore, all the occupancy methods are employed to artificially create an index of occupancy using previously available data such as calendar, old PIR sensor data, school schedules and other information. These different artificial indicators of occupancy are then tested in order to know which gives the best consumption forecasting results. From the simplest to the more sophisticated method, an explanation of the generation process, an analysis of the load forecasting performance and a contrast of the workload of each one needs to be performed.

In short, on the basis of the existing literature, all paradigmatic data sources and different techniques are used to generate several occupancy indicators. Then, a compendium of STLF performances and the pertinent workloads for each occupancy indicator is provided. The presented methods solve the previously commented shortcomings.

3. Background

Taking into account that a large amount of instances is available, covering a broad range of weather and building conditions, three paradigmatic black box models such as an MLR [9] model, an ANN [38] model and an SVR model were tested in [24]. The results showed that the SVR model provides the most accurate prediction for this kind of data and models. In that case, a grid search algorithm was used to adjust the training parameters of the models, in this case an evolutionary algorithm is used to adjust them. This section gives a brief explanation of the SVR model and the parameter optimization method.

3.1. Support vector regression

The support vector machine [36] model consists of separate classes, that are not linearly separable, transforming them using kernel functions and

moving them to a high-dimensional feature space where the data is classified through a hyperplane. On the other hand, the SVR performs a linear regression on this new high-dimensional feature space. The SVR function is seen on Equation 1:

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) k(x_i, x) + b \tag{1}$$

Where:

α_i and α_i^* are Lagrangian multipliers.
$k(x_i, x)$ is a kernel function.
b is a computed parameter.

There are several kernel functions [3] with different features, proper for each case. However, the most common are linear, radial basis function and polynomial kernel. There is no clear rule about which is better.

3.2. Parameter optimization

There are two main reasons for using a parameter optimization method. The first one is because all the occupancy indicators in the experiments must have equal conditions. The second is because the manually search of the suitable training parameters is a slow process and the grid search is computationally expensive.

The evolutionary computation approach executes a sub-process a multiple number of times to find the optimal values for the specified parameters. The evolutionary strategies, based on the theory of Rechenberg created in 1970 [32], help us to solve an optimization problem without falling into local optimum and premature closure.

Evolutionary search [4] is based on a parental and offspring candidate solution. These solutions, called individuals, are subject to random changes and selection of best solutions iteratively. Based on the principle of biological evolution, the concepts of recombination, mutation and selection are used to solve the problem. First, a recombination selects x parents and combines their parts to create new solutions. Then, the mutation adds random changes to the preliminary solutions. Finally, n individuals are selected and constitute the parental population of the following cycle. Until the termination condition is not achieved, the process continues.

4. Dataset

In this study the experiments are performed using data from four buildings (PI, PII, PIV and Faculty of Science) located at the University of Girona. The buildings are composed mainly by classrooms, offices and laboratories. The buildings PI and PII are used in all the experiments and the buildings PIV and Faculty of Science are only used for contrasting purposes. Regarding HVAC, the heating systems consist of gas boilers and fancoils. The cooling systems are composed of compression refrigeration systems and fancoils.

Building PI, built in 1983, has 6 floors and a volume of 26150 m^3. The frontage has an area of 3791 m^2, of which 610 m^2 are glass surface. Building PII has 6 floors, a volume of 25560 m^3 and was built in 1992. The frontage has an area of 2326 m^2, of which 1351 m^2 are glass surface. Building PIV has 3 floors, a volume of 12000 m^3 and was built in 2003. The frontage has an area of 1836 m^2, of which 630 m^2 is glass. The Faculty of Science building has 5 floors, a volume of 34810 m^3 and was built in 1997. The frontage has an area of 4903 m^2, of which 1233 m^2 is glass.

The set-point temperature is manually adjusted in the summer to 26°C, and in the winter to 20°C. The HVAC control system detects the temperature of the offices and classrooms, and modifies the fan speed to achieve the set point temperature. The profile of these buildings in relation to the HVAC is similar, in that the four buildings have systems where most of the consumption is produced with gas boilers.

As previously stated, temperature and occupancy are the main attributes used in the non-residential buildings forecasting. The data used in this work is as follows:

- Electric load data: electrical load data of the buildings PI, PII and PIV and Faculty of Science is collected using a power meter (PM810 Power Logics of Schneider) linked to the campus infrastructure monitoring system.
- Temperature data: temperature data using a sensor (Vaisala) from the Department of Physics outside the buildings.
- Calendar data: information about working and non-working days, holidays, exams, etc.
- School schedule: the hourly schedule of each classroom.

- Working schedule: the work schedule of the teachers and employees.
- Classroom size: the number of student places for each classroom.
- Classroom devices: the list of electrical devices and their features with regard to each classroom.
- Expert knowledge information: information about the building occupancy based on interviews with experts with experience.
- Occupancy sensor data: the data of occupancy collected in PIV using PIR sensors from previous work.

Based on this information, several artificial occupancy attributes with different levels of complexity are artificially created. The main objective is to analyse which one provides the best forecast. In the search of the proper occupancy indicator, there is probably a balance point between workload and forecasting performance.

The number of data instances of PI is 27375, covering a total of 38 months, from 1^{st} September, 2011 to 15th October, 2014. The total of instances of PII is 16589, covering a total of 24 months, from 21^{st} November, 2012 to 15th October, 2014. The number of instances of PIV is 16590, covering a total of 24 months, from 23rd November, 2012 to 15th October, 2014. The number of instances of Faculty of Science is 27366, covering a total of 38 months, from 1^{st} September, 2011 to 14th October, 2014.

The patterns of consumption and temperature for a summer (August 5th, 2013) and a winter (February 18th, 2013) week for the PI and PII buildings are shown in the Figure 2:

Figure 2: Temperature and consumption data for summer and winter weeks.

5. Methodology

This section contains the description of the artificial occupancy indicators and the forecasting method used.

5.1. Occupancy indicators

The main target is to create an artificial occupancy indicator to determine the occupancy in advance. Using some available information, there is the option of creating occupancy indicators to predict consumption some months ahead.

In this section, 43 occupancy indicators, with different levels of complexity, are created in order to find the best one. There are 7 methods used to create the indicators, ranging from low to high complexity. These 7 methods comprehend the main used techniques in the literature and some new lines are proposed, trying to cover all the possible data sources. The different indicators are tested in several experiments and finally a method is selected. The occupancy indicator is an attribute that varies from 0% to 100%. In summary, 43 short-term load forecasting models are trained and tested using only one occupancy indicator set as attribute each time with the aim of finding which performs better.

The first number of the indicator is referred to as the set, while the other ones are the data sources. For example, the indicator 4.32 is generated with the method of set 4 on the basis of indicators 2.3 and 3.2. The indicators, organized in sets, are as follows:

1. Indicator set 1.

    - Binary occupancy. The simplest indicator. If the university is open, there is a 100% occupancy. If the university is closed, there is 0% occupancy.

2. Indicator set 2 (2.1 to 2.3). Daily profile. These 3 indicators are based on daily profiles. There are 7 different daily profiles: school day, non-school day, examination day, school-leaving examination day, August day, holiday and weekend day and, finally, Easter and Christmas holiday. Each daily profile has its own level of occupancy. Each one of the 3 indicators of this set is created using only one of the 3 data sources. The data sources to describe each daily profile are:

    - Expert knowledge. Based on the experience of the employees of the university, a level of occupancy of the building for each day type is created.

- Sensor data. Based on the PIR sensor data collected for the previous work in the PIV building, a level of occupancy for each day type is created. The average of the level of occupancy for several days of each type of day is used.

- Teacher scheduling. Based on the schedules of certain employees of the university, a level of occupancy for each day type is created.

3. Indicator set 3 (3.1 to 3.3). Hourly profile. These 3 indicators are based on hourly profiles. There are 24 different hourly profiles. Each hour has its own level of occupancy. As in the previous case, each one of the 3 indicators of this set is created using only one of the 3 data sources. The data sources to describe each hourly profile are:

   - Expert knowledge. Based on the experience of the employees of the university, a level of occupancy of the building for each hour of the day is created.

   - Sensor data. Based on the PIR sensor data collected for the previous work in the PIV building, a level of occupancy for each hour type is created. The average of the level of occupancy for several hours of each type of hour is used.

   - Teacher scheduling. Based on the schedules of certain employees of the university, a level of occupancy for each hour type is created.

4. Indicator set 4 (4.1.1 to 4.3.3). Aggregation function profile. These 9 occupancy indicators are created by aggregating the indicators of sets 2 and 3. The main idea is to merge the hourly information with that of the days. Up to 5 aggregation functions are tested in order to discover which provides the best results. Then, the aggregation function which provides the best performance in terms of forecasting, is selected. The aggregation functions are the following ones:

   Aggregation function A is presented in the Equation 2:

$$I_A = \frac{I_2 + I_3}{k} \tag{2}$$

   Aggregation function B is presented in the Equation 3:

$$I_B = \frac{I_2 \times I_3}{k} \tag{3}$$

Aggregation function C is presented in the Equation 4:

$$I_C = \frac{\sqrt{I_2^2 + I_3^2}}{k} \tag{4}$$

Aggregation function D is presented in the Equation 5:

$$I_D = \frac{(I_2 + I_3)^2}{k} \tag{5}$$

Aggregation function E is presented in the Equation 6:

$$I_E = \frac{I_2 \times I_3}{k \times (I_2 + I_3)} \tag{6}$$

Where:

I_2 and I_3 are the aggregated indicators of sets 2 and 3.
k is the value to scale the output to the proper range, from 0 to 100.

5. Indicator set 5 (5.1.1 to 5.3.3). Summation of classes. These 9 indicators are based on the data of the previous indicator. The school, examination and school-leaving examination days instances are substituted for new values of occupancy. These new values are calculated taking into account the summation of the active classrooms for each hour. Therefore, the hour with more active classes is the hour with the maximum level of occupancy. Then, an adjustment is needed to equilibrate the instances of the previous indicator (holidays, non-school days and night hours) and the instances of the summation of classes.

   The occupancy of the building for a determined active hour is shown in Equation 7:

$$Oh_i = \frac{\sum_{i=1}^{m} Ac_i}{Mac} \times Eaf \times 100 \tag{7}$$

   Where:
   Oh_i is the level of occupancy of one building for a determined hour.

Ac_i is the number of active classrooms for a determined hour.
Mac is the maximum number of active classrooms.
Eaf is the adjustment factor. Varies from 0 to 1.

6. Indicator set 6 (6.1.1 to 6.3.3). Summation of weighted classes. On the basis of the previous indicator, a weighting that considers the electrical devices used in the classroom is added to the method. Each classroom is analysed and then the total electrical power of the devices is calculated. The weighting considers all types of rooms, from the laboratories which contain big electric motors, to theory classrooms which only have lights.

   The occupancy of the building for a determined active hour is shown in Equation 8:

$$Oh_i = \frac{\sum_{i=1}^{m} Ac_i}{Mac} \times \frac{\sum_{i=1}^{m} Edp_i}{Mep} \times Eaf \times 100 \tag{8}$$

   Where:
   Oh_i is the level of occupancy of one building for a determined hour.
   Ac_i is the number of active classrooms for a determined hour.
   Mac is the maximum number of active classrooms.
   Edp_i is the summation of the power of the electric devices for a determined classroom.
   Mep is the power of the classroom with more electric power.
   Eaf is the adjustment factor. Varies from 0 to 1.

7. Indicator set 7. (7.1.1 to 7.3.3). Summation of weighted classes with events. Using the data of indicator set 6, some variations in the occupancy are added at the beginning and at the end of certain events. The events are: the summer, the Christmas holidays, the examination period, the Easter week holidays, local festivities and university parties. In these events the occupancy is very slightly reduced.

In Figure 3 the several occupancy indicators for a week during school term are shown. The figure shows that the complexity of the profiles increases between the occupancy indicators.

Figure 3: Example of occupancy indicators of sets 4, 5 and 6.

5.2. Procedure block diagram

The proposed methodology consists of several blocks as shown in Figure 4. In the first block, the missing values are filtered. Then, the instances are normalized. In the following block, the outliers are filtered. In the next block, a feature selection is performed. Then, the data is split with 1/3 of the data to testing and 2/3 to training. In the case of PI, the training data goes from September 1^{st}, 2011 to September 13th, 2013 and the test data goes from September 13th, 2013 to October 15th, 2014. In the case of PII, the training data goes from November 23rd, 2012 to February 23rd, 2014 and the test data goes from February 23rd, 2013 to October 15th, 2014. At that point, an instance selection (20%) is performed with the training data, and an evolutionary search of the suitable training parameters is performed over the selected model. Finally, the validation of the model is done using test data.

Figure 4: Block diagram of the process.

5.2.1. Missing values filter

Due to mistakes in sensor readings, there is always a small amount of lost values. The percentage of missing values needs to be minimized as much as possible. There are several methods used to filter the missing values such as filling or deleting. In this case, the method that provides best performance in terms of forecasting, is the deletion of the instances with missing values.

In the case of PI, the instances with missing values are 691 out of a total of 27375, which represents 2.5%. For PII, the instances with missing values are 592 out of a total of 16589, which represents 3.6%. In the case of PIV, the instances with missing values are 579 out of a total of 16590, which represents 3.5%. In the case of the Science Faculty the instances with missing values are 683 out of a total of 27366, which represents 2.5%.

5.2.2. Normalization

Normalization is needed to work with different scales and units. The use of the same data scale improves the forecasting. The normalization range used is from 0 to 1.

5.2.3. Outliers filter

By filtering the outliers the performance of the model is increased. The outliers need to be detected and can then be deleted or filled. In the filtering,

the more restrictive the process, the greater the amount of data lost. In the present case, the method used to detect outliers is the local outlier factor, that consists of calculating the anomaly score according to the local outlier factor algorithm proposed by Breunig [11]. The instances with high scores are then removed.

In the case of PI, the instances with outliers are 227 out of a total of 26684, which represents 0.85%. For PII, the instances with outliers are 119 out of a total of 15997, which represents 0.74%. In the case of PIV, the instances with outliers are 118 out of a total of 16011, which represents 0.74%. In the case of the Science Faculty the instances with outliers are 227 out of a total of 26683, which represents 0.85%.

5.2.4. Feature selection

In order to remove irrelevant and duplicate data, the redundant and non-correlated attributes are removed. Reducing the size of the database, the computational cost of the training process is reduced. The feature selection consists in two blocks. In the first block the correlation with the class of each attribute is calculated and the features with low correlation are removed. In the second block the correlation between attributes is calculated and the attributes with high correlation with other attribute are removed.

That block is only for the experiments in which calendar nominal attributes are used, not for regular experiments, where only occupancy and temperature are used.

5.2.5. Instance selection

In order to reduce the computational cost of the training process, the number of instances is reduced. A random sub-sample of about 20% of the training data is selected. Some previous validations demonstrate that samples about this percentage reduce the computational time while maintain the forecasting performance levels.

5.2.6. Evolutionary search

The evolutionary search [4] is used to search the training parameters of the model. The objective of this is to deliver the same opportunities to each experiment in which all the models are trained using the same scenario. Each occupancy indicator has equal possibilities of providing the best forecasting results.

The main parameters of the evolutionary search are: maximum generations that specifies the number of generations after which the algorithm should be terminated; population size that stipulates the population size; mutation type that determines the type of the mutation operator; tournament fraction that specifies the fraction of the current population which should be used as tournament members and crossover probability that stipulates the probability of an individual being selected. The parameters of the evolutionary search method are given in the Table 2:

Parameter	**Value**
Max generations	35
Population size	5
Mutation type	Gaussian
Tournament fraction	0.25
Crossover probability	0.9

Table 2: Parameters of the evolutionary search.

5.2.7. Support vector machine

The main training parameters of SVR [34] are the C parameter, the type of kernel and the kernel parameters [3]. The tested kernels and their parameters, are linear (C), Polynomial (C and degree) and Radial Basis function (C and gamma). The C parameter is the complexity constant and adjusts the misclassification tolerance. If C is too large there is an over-fitting, but if it is too small there is an over-generalization. The polynomial kernel is defined by $k(x, y) = (x \times y + 1)^d$ where d is the degree of polynomial. The radial kernel is defined by $k(x, y) = exp(-g||x - y||^2)$ where g is gamma.

The optimization of the training parameters of SVR for each experiment, is performed using the evolutionary search method. A range for each parameter is defined before undertaking each experiment. Then, when the training process is finished, the proper parameters are found.

5.2.8. Validation

In the validation process, the model generated with training data (65%), is used to calculate the class attribute of the test data (35%). This data is then validated with a MAPE (mean absolute percentage error) indicator in front of the real values. The MAPE performance indicator is chosen due to its popularity in the forecasting field. The data is chronologically selected, so that the first period of time is used to predict the last period of time.

6. Experimental Results

The experiments have been realized using Rapid Miner [12] and a computer with an Intel Core i7-4500U processor and 8 GB of DDR3 RAM. In the next section the indicator used to measure the performance is described. Then, several scenarios and the results obtained are described.

6.1. Error indicator

Among the different methods used to calculate the quality of the model, mean absolute percentage error (MAPE) is the most common indicator found in the forecasting literature. The MAPE performance indicator, showed in Equation 9, does not depend on the magnitude of the unit of measurement, and is used to compare models. The smaller the MAPE, the more accurate is the model.

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_{m(i)} - y_{p(i)}}{y_{m(i)}} \right| \times 100 \tag{9}$$

Where:

N is the number of observations.
y_m is the measured output.
y_p is the predicted output.

6.2. Quality factor

The quality factor is a parameter calculated performing the weighted average of the MAPE and the workload, and then scaling it between 0 and 100, as seen in the Equation 10. The workload is created calculating the hours invested in the generation of each indicator set and then translating it in a 0 to 100 range.

$$Quality factor = \frac{MAPE + \times 0.1 * WL}{2 \times MQF} \times 100 \tag{10}$$

Where:

$MAPE$ is the mean absolute percentage error in a 0 to 100 range.
WL is the workload.
MM is the maximum value of MAPE.
MQF is the maximum value of quality factor.

6.3. Analytical results

In this section the results of several scenarios are analysed, and then the results of the best model are plotted. In each scenario a comparison is performed with the aim of resolving doubts and reaching a conclusion. The main purpose of the experiments is to discover which method generates the best occupancy indicator, a trade-off between workload and forecasting accuracy. The performance of the model (MAPE) is the main output in each experiment.

In the first scenario, the average performance of each indicator set is calculated. In the second scenario, the influence of the temperature attribute is analysed. In the third scenario, the prediction accuracy of each data driven-model is studied. In the fourth scenario, the performance of each data source is examined. In the fifth scenario, the several aggregation functions are assessed. In the sixth scenario, the performances of the SVR kernel functions are contrasted. In the seventh scenario, the proposed model is compared to a model based on several calendar nominal attributes. In the eighth scenario, the presented model is compared to an autoregressive model. In the ninth scenario, the performance of the model for several buildings is presented. In the tenth scenario, the forecasting accuracy for each sensor data treatment is evaluated. In the eleventh scenario, the workload for each indicator set is provided.

The several experiments, unless otherwise indicated, have been performed with PI and PII data, using occupancy and temperature attributes, with an SVR model, linear kernel, aggregation function E, 65% of the training data, 35% of the test data and 20% of the training sub-sample.

6.3.1. Scenario 1. Performance according to the indicator set

The first experiment is performed with temperature and each of the artificial occupancy attributes. For each building, a total of 43 of training and test processes are realized, one for each occupancy indicator. Then, the MAPE average is calculated for each occupancy indicator set. The main idea is to discover which is the best occupancy indicator set, that is the method that provides the best predictive accuracy, as seen in Table 3.

Indicator set	PI	PII
	Avg. MAPE (%)	Avg. MAPE (%)
Set 1	49.81	48.34
Set 2	24.34	22.90
Set 3	30.08	23.91
Set 4	18.11	18.05
Set 5	17.05	18.01
Set 6	17.03	17.96
Set 7	16.99	17.93

Table 3: Performance for indicator set.

The Table 3 indicates that most sophisticated indicator set, set 7, presents the best results. However, the improvement between sets 4, 5, 6 and 7 is very slight. So, there is a relationship between the complexity of the indicator and the MAPE, but it is not linear.

6.3.2. Scenario 2. Performance according to the temperature attribute

The objective of the second experiment is to determine if the temperature attribute improves the forecasting quality. The second experiment is implemented first with the temperature and then without the temperature attribute. The same steps as in the first experiment are undertaken, and the outcome is shown in Table 4.

Indicator set	PI		PII	
	Without Temp.	Temp.	Without Temp.	Temp.
	Avg. MAPE (%)		Avg. MAPE (%)	
Set 1	49.16	49.81	52.40	48.34
Set 2	23.21	24.34	27.01	22.90
Set 3	30.52	30.08	25.27	23.91
Set 4	17.13	18.11	18.67	18.05
Set 5	16.29	17.05	18.65	18.01
Set 6	16.28	17.03	18.53	17.96
Set 7	16.25	16.99	18.51	17.93

Table 4: Performance for temperature attribute.

As seen in Table 4, the outcomes do not show an improvement based on the use of the temperature attribute. The forecasting performance variation is minimal and in opposite directions.

6.3.3. Scenario 3. Performance according to the model

The third experiment has the aim of evaluating which is the most appropriate model. An MLR [9] and an ANN [38] models are compared with the SVR model. The adjustable training parameters are as follows: for the MLR the ridge factor and the feature selection method, and for the MLP the learning rate, the momentum and the training cycles. The experiment is conducted for indicator set 4, as seen in Table 5.

Model	**Avg. MAPE (%)**	
	PI	**PII**
MLR	23.69	21.13
MLP	21.31	34.96
SVR	18.11	18.05

Table 5: Performance for model (set 4).

According to Table 5, the SVR model provides the most suitable results, but there are no clear evidences as to whether the MLR or the MLP model is better. As seen in [24], the SVR model is the correct model for dealing with load consumption forecasting in non-residential buildings.

6.3.4. Scenario 4. Performance according to the data source

Experiment 4 consists of assessing the several data sources in order to know which is the most suitable: expert knowledge, sensor data or teacher schedule. The results for indicator set 4 are depicted in Table 6.

Data source	**Avg. MAPE (%)**	
	PI	**PII**
Expert knowledge	17.67	17.91
Sensor data	17.63	18.35
Teacher schedule	19.55	19.17

Table 6: Performance for data source (set 4).

The conclusion is that expert knowledge and sensor data sources enhance the forecasting accuracy of the teachers' schedule data source. The use of the sensor data source is preferable because is an impartial and repeatable method compared with the expert knowledge data source. The fact that the

sensor data were collected in PIV indicates that there is room for improvement. Therefore, the differences in the results between the three data sources are slight, as seen in Table 6.

6.3.5. Scenario 5. Performance according to the aggregation function

Experiment 5 analyses the effect of the aggregation function as described in 5.1. In the generation of the occupancy indicator 4, a process of aggregation between indicators 2 and 3, is carried out. To find out how to perform the aggregation process properly, five aggregation functions are tested. The results for indicator 4.2.2 are shown in Table 7.

Aggregation function	**Avg. MAPE (%)**	
	PI	**PII**
A	27.22	27.14
B	14.60	18.01
C	28.01	29.47
D	19.36	20.62
E	14.42	16.99

Table 7: Performance for aggregation function (indicator 4.2.2).

As shown in Table 7, the aggregation function is absolutely crucial. The aggregation functions B and E far exceed the results of the rest. The multiplicative aggregation functions improve the forecasting performance of the additive ones. The aggregation function E slightly exceeds performance method B.

6.3.6. Scenario 6. Performance according to the kernel

In experiment 6, the performance of each SVR kernel is tested. Linear, radial basis function (RBF) and polynomial kernels are analysed with the aim of comparing their forecasting accuracy. The parameters for the RBF kernel are C and gamma, for the polynomial kernel are C and the polynomial degree and for the linear kernel is C. The outcomes for indicator set 4 are listed in Table 8.

Kernel	Avg. MAPE (%)	
	PI	**PII**
Linear	18.11	18.05
Polynomial	25.47	22.55
RBF	20.59	19.45

Table 8: Performance for kernel type (set 4).

The experiments presented in Table 8, show that the linear kernel is the most efficient of all three kernels. In addition, the linear kernel involves a lower computational cost than the RBF and the polynomial.

6.3.7. Scenario 7. Performance for nominal attributes

The principal purpose of the experiment 7 is to prove that the utilization of one single occupancy attribute in the load forecasting model is more appropriate than the use of several calendar nominal attributes, such as year, month, week day, holiday, type of day and hour of the day.

In the experiment, all the nominal attributes are converted into numeric class to deal with the SVR. A comparison is performed between the nominal attributes model and the 4.2.2 occupancy indicator model. The results are shown in Table 9.

Attribute type	Avg. MAPE (%)	
	PI	**PII**
Calendar nominal attributes	26.67	21.63
Indicator 4.2.2	14.42	16.99

Table 9: Performance for attribute type (indicator 4.2.2).

As seen in Table 9, the presented model provides better prediction results than the model with the calendar nominal attributes. The presented model outperforms the models with a large set of attributes if the data used in the occupancy attribute creation is processed correctly. Significant differences can also be seen in terms of computational time, in favour of one single attribute model.

6.3.8. Scenario 8. Performance for auto-regression model

In the experiment 8 the main issue is to show that the presented model increases the forecasting accuracy compared with the auto-regressive models

[1]. So, a 24-hour ahead ARMA model with exogenous variables including temperature, is contrasted with the SVR model with the 4.2.2 occupancy indicator. The results are presented in Table 10.

Model	**Avg. MAPE (%)**	
	PI	**PII**
ARMA-X	26.84	19.87
SVR (indicator 4.2.2)	14.42	16.99

Table 10: Performance for model (indicator 4.2.2).

This experiment contrasts with an ARMA-X model where the past values of consumption and temperature are used as attributes, with the presented model. Usually, the auto-regressive models provide suitable results in this field. However, the presented results show that the ARMA model does not improve the occupancy indicator SVR model, as seen in Table 10. Moreover, due to the amount of attributes used in the ARMA-X, the difference in computational time, in favour of the single occupancy attribute, is remarkable.

6.3.9. Scenario 9. Performance according to the building

Experiment 9 is done to test the method in other buildings of the university. PI and PII buildings are compared to PIV and the Faculty of Science buildings. The results of the comparison for indicator set 4 are presented in Table 11.

Building	**Avg. MAPE (%)**
PI	18.11
PII	18.05
PIV	16.35
Science	18.75

Table 11: Performance for building (set 4).

The results of PIV and the Faculty of Science buildings are similar to the results of the buildings used in the experiments. The PI, PII and Faculty of Science buildings have the same profile in terms of offices, laboratories and classrooms. However, the profile of PIV is different as it consists mainly of offices. Due to the fact that the sensor data collection was realized in PIV, the prediction accuracy for PIV is higher.

6.3.10. Scenario 10. Performance according to the sensor data treatment

The experiment 10 analyses which is the most suitable method for processing the sensor data. The performance comparison is between the presented model, the 4.2.2 indicator, and an occupancy indicator generated by calculating for each of the 7 day profiles the 24 hourly occupancy levels, based on the average of the available sensor data, for an each hour of each day profile.

Sensor data treatment	**Avg. MAPE (%)**	
	PI	**PII**
Hour per day	14,76	17,19
Aggregation function (indicator 4.2.2)	14.42	16.99

Table 12: Performance for sensor data treatment (indicator 4.2.2).

The results show that the presented method is slightly better than the hour per day method, as Table 12 shows.

6.3.11. Scenario 11. Workload according to the indicator

Experiment 11 clarifies the workload product of the creation of each occupancy indicator.

Indicator set	**Workload units**
1	10
2	20
3	20
4	25
5	80
6	90
7	100

Table 13: Workload for each indicator set.

The aim of the experiment 11 is to show the great difference in the generation of indicators 1, 2, 3 and 4, that involve a small amount of work, and indicators 5, 6 and 7, the production of which requires more labour hours, as shown in Table 13.

6.4. Graphic results

In the present section some of the previous experiments are plotted. The following figures show the output of the model based on the indicators, with the best ratio between accuracy in terms of forecasting and the workload needed to produce it, appearing with regard to indicator 4.2.2.

6.4.1. Chart 1. MAPE vs. hour

In the Figure 5, a chart of MAPE vs. hour for both buildings is shown.

Figure 5: MAPE vs. hour.

Figure 5 shows that the prediction in the class hours presents good results. There are 4 hourly zones where the prediction is of poor quality. These time-slots are at the beginning and the end of the school day, at lunch time and during some night time hours.

6.4.2. Chart 2. MAPE vs. day of the week

In Figure 6, a MAPE vs. day plot for both buildings is presented, where the numbers 1 to 7 represent Monday to Sunday respectively.

Figure 6: MAPE vs. day of the week.

As seen in Figure 6, the forecasting of the midweek days is suitable. The prediction performance decays principally at the beginning and at the end of the working week and on Saturdays.

6.4.3. Chart 3. MAPE vs. day type

In Figure 7, a MAPE vs. day type chart for both buildings is plotted.

Figure 7: MAPE vs. day type.

Among the several profiles of days: school (1), exam (3), school-leaving examination (4), holiday and weekend (6) days are well predicted by the model. At a lower level of prediction performance there are: non-school (2), Easter week and Christmas (7) and August days (5).

6.4.4. Chart 4. MAPE vs. month

In Figure 8, a MAPE vs. month chart for both buildings is shown.

Figure 8: MAPE vs. month.

Overall, both models offer a poor level of prediction in April and August. Equally, the following months are better in terms of prediction: January, February, March, June, September, October, November and December.

6.4.5. Chart 5. Quality factor vs. indicator set

In Figure 9, a plot of the quality factor vs. the occupancy indicator set is presented.

Figure 9: Quality factor vs. occupancy indicator set.

The smaller the quality factor, the better it is. Figure 9 shows that indicator set 4 presents a balance between prediction accuracy and workload.

7. Discussion

Taking into account the experiments, there is a non-linear relationship between occupancy indicator complexity and forecasting accuracy. The computational cost is not the main issue in this work because most of the experiments use few attributes. In this work, the workload to generate the artificial attributes is a major concern. The more sophisticated indicators (5, 6 and 7) predict better than the simple ones (1, 2 and 3) but require a large amount of workload. There is a balance point situated on indicator set 4, where forecasting precision and workload are suitable, as the quality factor indicates in Figure 9. In addition, the occupancy indicators created using expert knowledge and sensor data sources provide a superior prediction than the teacher schedule ones, although the method based on collected sensor data is more expensive, delivers more impartiality and repeatability. In relation to the aggregation functions, it is shown that multiplicative aggregation functions such as aggregation function E, are much better than the additive ones. In addition, the experiments show that the temperature attribute, in this work, is not necessary, so it does not improve the load forecasting. This is due to the partial disaggregation of the HVAC system from the electric

consumption, since it is composed by gas boilers and fancoils, and a portion of the energy consumption is not electricity. Furthermore, the proposed sensor data treatment, based on the aggregation functions, enhances the hour per day treatment.

As in [24], SVR model outperforms the other tested models (MLR, ARMA-X and MLP), however the computational cost slightly increases due the low number of attributes, though it is entirely acceptable. Moreover, among the several tested SVR kernels, the linear kernel not only provides more accurate predictions, but also involves a short computational training time. Furthermore, the utilization of a single attribute of occupation in comparison with several calendar nominal attributes such as hour, day of the week, day type and month has resulted in a more compact and precise model. Additionally, the method has proved to work satisfactorily with other university buildings such as the PIV and Faculty of Science ones, as shown in Table 11.

Analysing the charts, can be seen that the worst consumption prediction periods are in the non-well defined human conduct intervals, and in the high variability intervals. In Figure 5, referring to MAPE vs. hour chart, the load forecast is less efficient in the nocturnal hours, at the beginning and the end of the school-day and at lunch time. In the nocturnal hours, this is due to the uncertainties generated by the cleaning and security services. At lunch time and at the beginning and the end of the school-day it is due to the variability in the individual behaviours of the users.

Comparing Figure 6, which refers to the MAPE vs. day chart, the prediction performance decreases principally with regard to the beginning and the end of the working week and Saturdays. Mondays and Fridays contain a large variability, especially Fridays, because there are no classes in the afternoon, but some teacher's offices are occupied. Saturdays are complicated with regard to prediction due to random activities in the university installations, which is not the case on Sundays.

In relation to Figure 7, referring to MAPE vs. day type chart, the accuracy of the model is reduced in non-school, Easter week, Christmas and August days. The profile of days that are adequately predicted are uniform days. For example, in school, examination and school-leaving examination days, the university is open and there are students. In the same way, in holidays and weekend days the university is closed and there are neither students nor teachers. On the other hand, non-school, Easter week and Christmas days are not accurately predicted. Given the dispersion of human behaviours, there are no students in the building, but some employees tend

to work during these periods. In relation to August, the HVAC system is not running, so the consumption pattern is slightly different, and there are some employees who work with non-defined schedules in some laboratories.

As shown in Figure 8, referring to the MAPE vs. month chart, the prediction presents the lowest forecasting levels during April and August. The month of April mainly contains Easter week, therefore is hard to predict, as explained previously. The month of August has been explained previously. In general, months with classes, where consumption patterns are mainly defined by the students' behaviour, the low-dispersion human behaviour periods, are the months that present the highest accuracy level with regard to prediction.

Among the improvements for future work, there is no doubt that enhancing the descriptive level of human behaviour in terms of the worst-defined time periods would improve forecasting accuracy. Furthermore, the chosen model that appears in the charts is based on the sensor data (indicator 4.2.2). For this reason the largest deviations in terms of prediction are located in specific hourly or daily periods. If another data source has been used, the forecasting divergences would be located in other intervals. So, the data sources could be analysed to know which are better for each time-slot, and then apply them selectively or mixed in order to achieve optimal performance. Finally, a revision and an improvement in terms of the occupancy levels of special days including Easter week, Christmas and non-school days, is necessary. Perhaps, the model could be improved by incrementing the number of captured special days in the sensors' database. Besides, although the results are not poor, the sensor data source uses general data from the PIV building. Performing a short data collection procedure in the other buildings to obtain specific data could improve the accuracy of the predictions. It is important to note that by improving the adjustment between indicator sets 3 and 4, some additional prediction accuracy could be obtained.

8. Conclusions

One of the most prevailing needs in terms of utilities is to adjust electricity generation to consumption. For this reason, consumption forecasting is a well understood domain. Also, 40% of electricity consumption is in the building sector. In a previous paper [24], the authors presented an STLF model for non-residential buildings for the University of Girona. The main results obtained showed that using occupancy and temperature as attributes, and as a model the SVR model provides the best load forecasting. However,

that model used continuous occupancy sensor data, unavailable in advance. In fact, the main purpose of the paper was to determine the appropriate attributes and models. Now, a fully operational STLF model for the non-residential buildings of the University of Girona is presented.

This paper aims to dispose of the occupancy data in advance. Therefore, several artificial occupancy attributes from different data sources have been created. Then, to find which is the best artificial occupancy indicator, several methods and data sources including sensor data, expert knowledge, class schedules and school calendar, are tested and analysed through the SVR model. Furthermore, this information is compared in terms of the workload resulting from the creation process associated with each occupancy attribute, searching for the most balanced occupancy indicator between performance and workload. Finally, some experiments are conducted to compare the proposed model to other classic models and attributes.

Although the prediction accuracy is lower with respect to previous work [24], the main objective of the presented work is to generate a model based only on artificial attributes, tracing a new path towards artificial occupancy attributes generation methods. The results show that the model which has the best ratio between forecasting precision and workload is an SVR model with a linear kernel trained only with one occupancy attribute generated from the aggregation of the hourly and daily profiles, based on sensor data. So, the SVR model provides the best results in comparison with other data-driven models (ARMA-X, MLR or ANN). Moreover, taking into account the partial disaggregation of the HVAC system, the model does not depend on temperature, converting it in a more compact and simple model and reducing the computational cost. Unlike the other models, this new model can perform hourly consumption predictions months in advance, using only occupancy data. In addition, the proposed method could interpolate the new consumption levels if new classrooms would be constructed, which differs from other works.

In summary, an STLF method for non-residential buildings is provided. This simple and compact model predicts the hourly consumption, months in advance, and is based only on occupancy. Other methods are based on auto-regression or on the need for previously unavailable exogenous variables, and thus require weather forecasts or consumption data to perform the prediction, making a long-term hourly forecast impossible. Moreover, this paper explains the methods for the generation of these occupancy indicators. Every occupancy attribute is assessed in order to determine which method and

data source provide the best results in terms of prediction. In future work, departing from the presented methods, some indicator adjustments and revisions of the data sources will be performed in order to improve the forecasting precision of the method.

Acknowledgments

This research project has been partially funded through BR-UdG Scholarship of the University of Girona granted to Joaquim Massana Raurich. Work developed with the support of the research group SITES awarded with distinction by the Generalitat de Catalunya (SGR 2014-2016) and the MESC project funded by the Spanish MINECO (Ref. DPI2013-47450-C2-1-R).

[1] H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

[2] S. Aman, Y. Simmhan, and V. K. Prasanna. Improving energy use forecast for campus micro-grids using indirect indicators. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 389–397. IEEE, 2011.

[3] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.

[4] T. Bäck. Evolutionary algorithms in theory and practice. In *Evolutionary algorithms in theory and practice*. Oxford university press, 1996.

[5] Y. Chae, R. Horesh, Y. Hwang, and Y. Lee. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy and Buildings*, 111:184–194, 2016.

[6] C. Cui, T. Wu, M. Hu, J. D. Weir, and X. Li. Short-term building energy model recommendation system: A meta-learning approach. *Applied Energy*, 172:251–263, 2016.

[7] D. Datta, S. A. Tassou, and D. Marriott. Application of neural networks for the prediction of the energy consumption in a supermarket. In *Proceedings of the International Conference CLIMA*, pages 98–107, 2000.

[8] M. De Felice and X. Yao. Neural networks ensembles for short-term load forecasting. In *Computational Intelligence Applications In Smart Grid (CIASG), 2011 IEEE Symposium on*, pages 1–8. IEEE, 2011.

[9] N. R. Draper, H. Smith, and E. Pownell. Applied regression analysis. In *Applied regression analysis*, volume 3. Wiley New York, 1966.

[10] M. Grubb, C. Vrolijk, and D. Brack. The Kyoto Protocol. A guide and assessment. *The Kyoto Protocol. A guide and assessment*, 1999.

[11] Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9):1641–1650, 2003.

[12] M. Hofmann and R. Klinkenberg. *RapidMiner: Data mining use cases and business analytics applications*. CRC Press, 2013.

[13] T. Hong, S. D'Oca, W. J. N. Turner, and S. C. Taylor-Lange. An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to the DNAs framework. *Building and Environment*, 92:764–777, 2015.

[14] T. Hong, H. Sun, Y. Chen, S. C. Taylor-Lange, and D. Yan. An occupant behavior modeling tool for co-simulation. *Energy and Buildings*, 117:272–281, 2016.

[15] IEA-EBC. Annex 53, Total energy use in buildings: analysis & evaluation ,methods. Technical report, International Energy Agencys Energy in Buildings and Communities Programme, 2013.

[16] IEA-EBC. Annex 66, Definition and simulation of occupant behavior in buildings. Technical report, International Energy Agencys Energy in Buildings and Communities Programme, 2015.

[17] S. Karatasou, M. Santamouris, and V. Geros. Modeling and predicting building's energy use with artificial neural networks: Methods and results. *Energy and Buildings*, 38(8):949–958, 2006.

[18] S. Kwok and E. Lee. A study of the importance of occupancy to building cooling load in prediction by intelligent approach. *Energy Conversion and Management*, 52(7):2555–2564, 2011.

[19] P. C. M. Leung and E. W. M. Lee. Estimation of electrical power consumption in subway station design by intelligent approach. *Applied Energy*, 101:634–643, 2013.

[20] K. Li, H. Su, and J. Chu. Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study. *Energy and Buildings*, 43(10):2893–2899, 2011.

[21] X. Li and J. Wen. Building energy consumption on-line forecasting using physics based system identification. *Energy and Buildings*, 82:1–12, 2014.

[22] X. Li, J. Wen, and E.W. Bai. Developing a whole building cooling energy forecasting model for on-line operation optimization using proactive system identification. *Applied Energy*, 164:69–88, 2016.

[23] X. Lü, T. Lu, C. J. Kibert, and M. Viljanen. Modeling and forecasting energy consumption for heterogeneous buildings using a physical–statistical approach. *Applied Energy*, 144:261–275, 2015.

[24] J. Massana, C. Pous, L. Burgas, J. Melendez, and J. Colomer. Short-term load forecasting in a non-residential building contrasting models and attributes. *Energy and Buildings*, 92:322–330, apr 2015.

[25] A. Neto and F. Fiorelli. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and buildings*, 40(12):2169–2176, 2008.

[26] G. R. Newsham and B. J. Birt. Building-level occupancy data to improve ARIMA-based electricity use forecasts. In *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, pages 13–18. ACM, 2010.

[27] J. A. Oliveira-Lima, R. Morais, J. F. Martins, A. Florea, and C. Lima. Load forecast on intelligent buildings based on temporary occupancy monitoring. *Energy and Buildings*, 2016.

[28] Y. K. Penya, C.E Borges, D. Agote, and I. Fernández. Short-term load forecasting in air-conditioned non-residential Buildings. In *Industrial Electronics (ISIE), 2011 IEEE International Symposium on*, pages 1359–1364. IEEE, 2011.

[29] Y. K. Penya, C.E. Borges, and I. Fernández. Short-term load forecasting in non-residential buildings. In *AFRICON, 2011*, pages 1–6. IEEE, 2011.

[30] J. C. Piris. *The Lisbon Treaty: A Legal and Political Analysis*. Cambridge University Press, 2010.

[31] K. Powell, A. Sriprasad, W. Cole, and T. Edgar. Heating, cooling, and electrical load forecasting for a large-scale district energy system. *Energy*, 74:877–885, 2014.

[32] I. Rechenberg. Evolution Strategy: Optimization of Technical systems by means of biological evolution. *Fromman-Holzboog, Stuttgart*, 104, 1973.

[33] C. Sandels, J. Widén, L. Nordström, and E. Andersson. Day-ahead predictions of electricity consumption in a Swedish office building from weather, occupancy, and temporal data. *Energy and Buildings*, 108:279–290, 2015.

[34] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[35] E. Spiliotis, A. Raptis, Z. N. Legaki, and V. Assimakopoulos. Forecasting electrical consumption of commercial buildings using energy performance indicators. *International Journal of Decision Support Systems*, 1(2):164–182, 2015.

[36] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[37] S. L. Wong, K. K. W. Wan, and T.N. T. Lam. Artificial neural networks for energy analysis of office buildings with daylighting. *Applied Energy*, 87(2):551–557, 2010.

[38] B. Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.