Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 994 – 999

Procedia Computer Science

International Workshop on Big Data and Data Mining Challenges on IoT and Pervasive Systems (BigD2M 2016)

# Towards Energy Efficiency Smart Buildings Models based on Intelligent Data Analytics

Aurora González-Vidal, Victoria Moreno-Cano, Fernando Terroso-Sáenz, Antonio F. Skarmeta

*Computer Science Faculty. University of Murcia, Spain*

{*aurora.gonzalez2, mvmoreno, fterroso, skarmeta*}*@um.es*

**Abstract**

This work presents how to proceed during the processing of all available data coming from smart buildings to generate models that predict their energy consumption. For this, we propose a methodology that includes the application of different intelligent data analysis techniques and algorithms that have already been applied successfully in related scenarios, and the selection of the best one depending on the value of the selected metric used for the evaluation. This result depends on the specific characteristics of the target building and the available data. Among the techniques applied to a reference building, Bayesian Regularized Neural Networks and Random Forest are selected because they provide the most accurate predictive results.



Peer-review under responsibility of the Conference Program Chairs

*Keywords:* pervasive computing, smart buildings, energy efficiency, intelligence data analysis techniques

## 1. Introduction

With the advent of the new advances and techniques on Information and Communications Technologies (ICT), every place, everything and everyone can be embraced by embedded technologies allowing connection and communication between them in a non-intrusive and efficient way. This is the technological basis promoted by the so popular *Internet of Things (IoT)* [1].

The high volume of data that can be generated nowadays in urban environments, coming from different data sources, provides a great scenario to achieve intelligent and efficient management systems of energy consumption based on IoT. *Big data analytics* [3] helps us to leverage the huge amounts of data provided by IoT-based ecosystems in order to reveal insights that help explain, expose and predict knowledge from them. [2]

* Aurora González; Tel.: +34-868-88-7866.
*E-mail address:* aurora.gonzalez2@um.es


Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.04.213

Specifically, in the field of smart buildings - which are a key piece of smart cities - it is increasingly common to apply intelligent algorithms to generate behavioural building models for solving problems like energy efficiency and comfort provisioning[4,5].

In this paper, a first approach to model the energy consumption of smart buildings is proposed considering a context that provides a reduced set of data to generate the model and the use of intelligent techniques to identify patterns that could help in the modeling of the smart building status related to its energy consumption. After selecting a set of recommended techniques, we propose a general procedure to analyze the performance once they have been applied to specific buildings, and the criteria to consider for selecting the optimal one according to the achieved results.

The structure of this paper is: Section 2 reviews the main techniques proposed in literature in order to model the smart building energy consumption. Section 3 shows how to proceed during the data processing to generate accurate building energy consumption models. Section 4 presents the reference building and its available data, describes the application of different techniques as well as the way to select the optimal one. Finally, Section 5 gives some conclusions and an outlook of our future work in this area.

## 2. Intelligent Data Analysis Techniques for Buildings Modeling

Intelligent data analysis techniques have been wisely introduced to model several building systems scenarios.

Artificial neural networks (ANN) are able to learn the key information patterns within a multidimensional domain. These have been applied in the field of solar energy, for modeling and design of a solar steam generating plant, for the estimation of heating-loads of buildings, etc. Also in heating, ventilating and air-conditioning systems, solar radiation, modeling and control of power-generation systems, load-forecasting and refrigeration[6]. The ANN used is Multi Layer Perceptron (MLP). Also, Bayesian Regularized Neural Network (BRNN) method has been used in the prediction of a series of building energy loads from an environmental input set[7] and the Random Forest model has been applied in order to predict energy consumption in residential buildings[8].

Likewise, Support Vector Machines (SVM) are proposed - and evaluated- to predict both the total short-term electricity load and the short-term loads of individual building service systems (air conditioning, lighting, power, and other equipment) in buildings that have electricity sub-metering systems installed[9].

Another common technique for non-linear regression proposed in literature to be applied are Gaussian Processes with Radial Basis Function Kernel (RBF)[10]. It has already been used to forecast electrical load[11] or to estimate the number of occupants in a room according to data related to the room status: motion detection, $CO_2$ reading, sound level, ambient light and door state sensing[12].

## 3. General Data Modeling Process

The five techniques that are introduced in the previous section as reference are used in order to train the energy consumption prediction of buildings, looking for the optimal configuration of their hyperparameters. For this purpose, we use the R[13] package named CARET[14]. This package is a set of functions that attempts to streamline the process for creating predictive models. The five techniques implemented in R enable us to adjust their tuning parameters. Table 1 shows the different values taken for each technique's hyperparameters. For example, we train the MLP model using the size values from 33 to 40 and select the one that reaches better results according to the evaluation metric.

Using each of these techniques, different building models are generated following the next steps (shown explicitly in Fig. 1):

1. Cleaning and transformation: selecting predictive variables, deleting energy consumption outliers that cannot be related to outliers in the rest of the variables, transforming categorical into numerical variables, and dividing the set of data into train (75%) and test (25%).
2. Standarization: transform the variables to have zero mean and unit standard deviation.
3. A common technique applied to data is the transformation of the data space using the so called Principal Components Analysis (PCA)[15]. PCA is a widely used technique for reducing dimensionality, identifying the directions in which the variance of the observations is accumulated.
4. Validation method: 10-fold cross validation and 5 repetitions over the training data set.

Table 1. Features of the evaluated algorithms

| Technique | Function in R | Tuning parameter | Values for tuning |
|---|---|---|---|
| Multi-Layer Perceptron (MLP) | mlp | size | {33, 34, 35, 36, 37, 38, 40} |
| Support Vector Machines with Radial Basis Function Kernel (SVM) | svmRadialCost | cost | {1, 2, 3, 4, 5, 8, 10} |
| Gaussian Process with Radial Basis Function Kernel (Gauss) | gaussprRadial | sigma | {0.01, 0.05, 0.1, 0.5} |
| Bayesian Regularized Neural Networks (BRNN) | brnn | neurons | {2, 3, 4, 5, 10} |
| Random forest (RF) | rf | mtry | {2, 3, 4, 5, 6, 7} |

5. Evaluation metric: RMSE (Root-Mean-Square Error) and R-Squared. The formula yields the values in the same units as the output of the estimators -KWh in this case- so the results can be interpreted easily. The coefficient of variation (CV) of RMSE, that indicates the uncertainty in the model, is the reference metric.

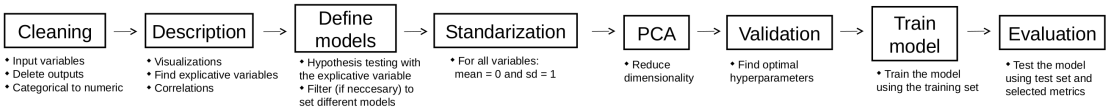


Fig. 1. Explicit data modeling process

## 4. Application in the Reference Building

The reference building in which the proposed procedure has been carried out to generate accurate building models is the Technological Transfer Centre (TTC) of the University of Murcia[*]. This building is used by technological companies and some research groups that collaborate with companies developing industrial scientific projects.

The building has a wide deployment of sensors and devices integrated in a home automation system which is working to improve indoor comfort at the same time that energy is saved. The home automation system installed in this reference scenario is called City explorer, which is composed by programmable logic controllers (PCL) and an SCADA system. On the one hand, the PLC is able to monitor the sensor status and regulate the infrastructures connected to City explorer. On the other hand, the SCADA system collects data and intercommunicates the PLCs with the actuators of the building. City explorer has been designed and developed at the University of Murcia, and is currently a commercial product provided by Odin Solutions S.L.[**].

### *4.1. The available data*

The available data for this building can only represent the very minimal situations in order to apply any algorithm to obtain some predictions onto its global energy consumption. These data are the environmental outdoor observations and the total energy consumption of the building from 1st December, 2014 to 18th February, 2016 in intervals of 8 hours. In total, 952 observations.

When energy is measured by commercial power meters, usually many variables are provided but *Active Energy* is our target. Active Energy is the active power (KW) consumed per time unit and it depends on the interval of time

[*] www.um.es/otri/?opc=cttfuentealamo
[**] www.odins.es/en

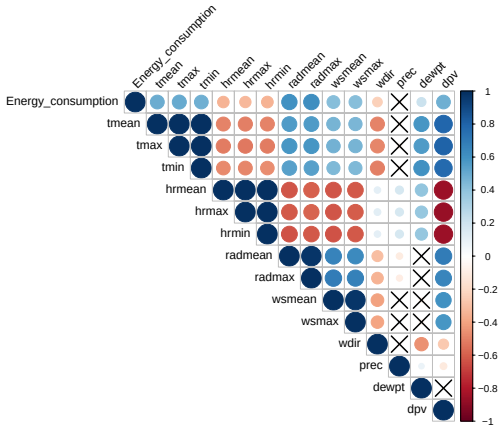


Fig. 2. Correlation heatmap between consumption and outdoor environmental conditions

because it accumulates its value. Hence, in order to have an accurate and meaningful measure of energy consumption (KWh), the intervals of time between observations have to be equal.

Measures are considered in intervals of 8 hours, and the origin of the consumption (HVAC, lighting or other electrical equipment) is unknown. Outdoor environmental measures are acquired from external sources. In this case, the IMIDA (The Research Institute of Agriculture and Food Development of Murcia) has provided us with an hourly historical set of data including the following variables: temperature (mean, min and max) (°C), humidity (mean, min and max) (%), radiation (mean and max) ($w/m^2$), wind speed (mean and max) ($m/s^2$), wind direction (mean) (degrees), precipitation (mm), dew point (°C) and vapour pressure deficit (kPa).

### *4.2. Correlations between consumption and outdoor environmental conditions*

Fig. 2 shows the pairwise correlations between all variables involved in the problem. Focusing on the firs row, we see that energy consumption correlates significantly ($\alpha = 0.95$) and positively (blue circle) with temperature, radiation, wind speed variables, vapour pressure deficit and dew point, and negatively (red circle) with wind direction and humidity variables. This means that we can use safely these variables as inputs of the energy consumption model of our reference building, because they all have clear impact in the energy consumption except precipitations (crossed out because the are not significant).

### *4.3. Occupation of the building*

Having as a goal to generate the most basic case of study and taking into account that occupation information it is not usually available in buildings - it requires an exhaustive sensor deployment - we have displayed an outline based on basic and logic usability estimations of the building:

- Moment 1: holidays, weekends and nights (22:00 PM- 06:00 AM)
- Moment 2: regular mornings (06:00 AM - 14:00 PM)
- Moment 3: regular afternoons (14:00 PM - 22:00 PM)

Looking at Fig. 3, it is possible to appreciate differences between moments but in order to have statistical support to define those groups a Kruskall Wallis H[16] was performed in order to check if there are differences in energy consumption between them. The test reveals that, indeed, there is a significant difference between groups (H(2) = 547.7, p-value < 0.01). An analysis of the differences by pairs performing the post-hoc Wilcoxon test[16], determines that it is

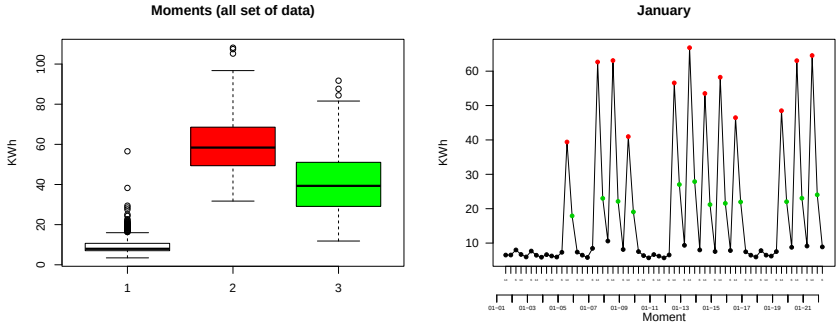


Fig. 3. Boxplot of the energy consumption by moments considering all data (left); and, the time series of the energy consumption by moments during January (right)

possible to divide data in those moments. This reasoning leads us to suggest three different models corresponding to the just mentioned partitions.

- Model 1. Range of energy consumption = [3.578, 14.1] KWh, mean of energy consumption = 7.904 KWh.
- Model 2. Range of energy consumption = [26.01, 86.19] KWh, mean of energy consumption = 54.27 KWh.
- Model 3. Range of energy consumption = [6.357, 53.290] KWh, mean of energy consumption = 31.48 KWh.

### *4.4. Results*

Every model gathers the energy consumption during 8 hours (as described in subsection 4.3), so we have 8 different observations for each environmental input (one each hour). Also, we create two new variables for every attribute by taking its mean and median. Just to clarify the considered inputs, for situation 1 and, for example, temperature, we will have 11 attributes: temperature at 6 AM, at 5 AM, ... at 22 PM, mean of temperature (from 6AM to 22PM) and median of temperature.

After training the models using several combinations of inputs we achieve the best results using day of the week, month, season, mean temperature and mean humidity with the Random Forest (RF) algorithm for situation 1 (mtry = 4, RMSE = 1 KWh) and situation 3 (mtry = 2, RMSE = 3.87 KWh) and Bayesian Regularized Neural Networks (BRNN) for situation 2 (number of neurons = 2, RMSE = 7.08 KWh) representing all these values between a 12.09% and a 12.86% of error (CVRMSE).

## 5. Conclusion and Future Work

In this paper, we have established a basic and successful procedure that can be used at the initial stages of the analysis of energy consumption in smart buildings. This process has been carried out in a reference building from which we have generated different energy consumption models. Among the techniques analyzed, BRNN and RF provided the most accurate results (mean errors within [1, 7.08] KWh). This procedure will be enriched progressively with the addition of more data sources. The immediate step is the usage and validation of the model trying to predict energy consumption for future days using environmental outdoor predictions. Future work is designing a strategy of control based on this model to save energy in buildings.

## Acknowledgements

This work has been partially funded by MINECO TIN2014-52099-R project (grant BES-2015-071956) and ERDF funds, by the European Commission through the H2020-ENTROPY-649849 EU Project, and the Spanish Seneca Foundation by means of the PD program (grant 19782/PD/15).

Table 2. Results obtained for each moment (see technique's acronym in Table 1)

| Moment | Technique | Best Parameter | RMSE (KWh) | CV RMSE (%) | $R^2$ |
|---|---|---|---|---|---|
| 1 | Gauss | $\sigma = 0.1$ | 1.1 | 13.43 | 0.57 |
| 1 | MLP | size = 34 | 1.1 | 13.46 | 0.55 |
| 1 | SVM | cost = 4 | 1.09 | 13.26 | 0.58 |
| 1 | BRNN | neurons = 3 | 1.1 | 13.47 | 0.55 |
| 1 | RF | mtry = 4 | 1 | 12.18 | 0.65 |
| 2 | Gauss | $\sigma = 0.1$ | 7.76 | 14.1 | 0.67 |
| 2 | MLP | size = 37 | 1.56 | 15 | 0.68 |
| 2 | SVM | cost = 1 | 4.26 | 13.4 | 0.71 |
| 2 | BRNN | neurons = 2 | 7.08 | 12.86 | 0.75 |
| 2 | RF | mtry = 2 | 7.48 | 13.59 | 0.72 |
| 3 | Gauss | $\sigma = 0.1$ | 4.5 | 14.07 | 0.67 |
| 3 | MLP | size = 37 | 4.81 | 15.03 | 0.69 |
| 3 | SVM | cost = 1 | 4.20 | 13.14 | 0.73 |
| 3 | BRNN | neurons = 5 | 4.31 | 13.45 | 0.73 |
| 3 | RF | mtry = 2 | 3.87 | 12.09 | 0.76 |

## References


1. L. Atzori, A. Iera, G. Morabito, The internet of things: A survey, Computer networks 54 (15) (2010) 2787–2805.
2. T. A. R. (auth.), Data Analytics: Models and Algorithms for Intelligent Data Analysis, 1st Edition, Vieweg+Teubner Verlag, 2012.
3. C. L. Stimmel, Big Data Analytics Strategies for the Smart Grid, CRC Press, 2014.
4. L. G. Swan, V. I. Ugursal, Modeling of end-use energy consumption in the residential sector: A review of modeling techniques, Renewable and sustainable energy reviews 13 (8) (2009) 1819–1835.
5. G. K. Tso, K. K. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (9) (2007) 1761–1768.
6. S. A. Kalogirou, Applications of artificial neural-networks for energy systems, Applied Energy 67 (1) (2000) 17–35.
7. D. MacKay, Bayesian non-linear modeling for the 1993 energy prediction competition, Maximum Entropy and Bayesian Methods (1993) 221–234.
8. F. Wahid, D.-H. Kim, Prediction methodology of energy consumption based on random forest classifier in korean residential apartments.
9. Y. Fu, Z. Li, H. Zhang, P. Xu, Using support vector machine to predict next day electricity load of public buildings with sub-metering devices, Procedia Engineering 121 (2015) 1016–1022.
10. C. K. Williams, D. Barber, Bayesian classification with gaussian processes, Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (12) (1998) 1342–1351.
11. D. J. Leith, M. Heidl, J. V. Ringwood, Gaussian process prior models for electrical load forecasting, Probabilistic Methods Applied to Power Systems (2004) 112–117.
12. S. Mamidi, Y.-H. Chang, R. Maheswaran, Improving building energy efficiency with a network of sensing, learning and prediction agents, in: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 45–52.
13. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2015). URL `http://www.R-project.org/`
14. M. Kuhn, Building predictive models in R using the caret package, Journal of Statistical Software 28 (5) (2008) 1–26.
15. H. Abdi, L. J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics 2 (4) (2010) 433–459.
16. J. M. Andy Field, Z. F. Niblett, Discovering Statistics Using R, 1st Edition, Sage Publications Ltd, 2012.