Abnormal Electricity Consumption Detection from Incomplete Records in Power System

Yang Zhang, Pietro Colella,
Andrea Mazza, Ettore Bompard
Department of Energy
Politecnico di Torino
Torino, Italy
{yang.zhang, andrea.mazza, ettore.bompard}@polito.it

Emiliano Roggero, Giuliana Galofaro
Ireti SpA
Gruppo IREN
Torino, Italy
{emiliano.roggero,giuliana.galofaro}@ireti.it

***Abstract*—Due to the limited channel bandwidth or interference signals in the advance metering infrastructures, there are usually some missing or human-revised data among the electricity consumption records of civilian customers. In order to make full use of this kind of records, machine learning techniques are introduced in this paper for electricity consumption sensitivity analysis regarding to the weather features. With the missing and revised records filtered out, each customer would have an individual regression model between weather conditions and the power demand. The importance of variables in the regression model is regarded as the sensitivity to various weather features. Then the abnormal consumption patterns are detected with a typical outlier identification algorithm based on different weather sensitivities among all the customers. The methods used in this paper show good results to identify the abnormal consumption patterns effectively regardless the quality of the original data.**

***Index Terms*--Abnormal detection, electricity consumption, machine learning, local outlier factor, sensitivity analysis.**

I. INTRODUCTION

With the rapid development of information and communication technology, advanced metering infrastructures have been widely applied in power system at the customers' level. In US, the installation of smart meters has seen an increase since 2007. By the end of 2016, 72 million smart meters have been installed among over 55% of US houses [1]. The percentage of electricity customers in EU28+2 with smart meters are expected to reach 71% by 2023 [2]. In particular, the first generation of smart meters has started in Italy since 2001 and covered 95% of 36 million customers by the end of 2011 [3].

There have been already many researches focusing on the technical and optimal operations of distribution system [4][5]. While with the wide application of smart meters, the data analytics methods start to support the secure and economic operation of power systems. One of the practical applications with smart meters in power system is to detect the abnormal behaviors of customers, which is long of concern among utilities [6][7]. The hourly recorded energy consumption of building system has been analyzed in [8] with the classification and regression tree algorithm. The abnormal energy consumption is then detected by the generalized extreme studentized deviate algorithm. In the photovoltaic system, the fault degree is detected and evaluated with the local outlier factor (LOF) based algorithm in [9], which distinguishes the abnormal data with specific mathematical characteristics.

The recorded data regarding the customer consumption pattern may be not complete: in this paper, a possible solution for this situation is proposed, and the incomplete data are used for investigating the electricity consumption of individual customers. Since civilian customers' electricity consumption is normally affected by the weather conditions [10], a proper description of weather sensitivity among customers based on only the trusted part of records could be used for further analysis, including abnormal behavior detection. As one of the most widely used methods for sensitivity analysis, *regression model* is used to quantify the impact of model-input variables thanks to the fast computing and easy interpretation [11]. For example, a method of Bayesian Additive Regression Trees (BART) is introduced and utilized in [12] to capture the climate sensitivity among electric power consumption data. Furthermore, the abnormal electricity consumption patterns among civilian customers can be detected from the incomplete dataset.

The methodology proposed in this paper is summarized in Figure 1. and is composed of three main steps:

Data reading and pre-processing: the data refer to civilian consumption and weather data in year 2015. The issues faced by handling the data are presented in Section II and the corresponding methods for overcoming them will be presented in Section III.

Customers' modelling: this step focuses on the building of the regression model for revealing the relation between the weather conditions and the individual customer's consumption (Section III.A). The relatively importance of the different weather features considered in the model has been obtained through a sensitivity analysis shown in Secton III.B.

978-1-7281-3520-5/19/$31.00 ©2019 IEEE

Abnormal consumption detection: thanks to the model obtained in Step 2, it is possible to detect the customers whose sensitivities are diverging with respect the majority of the other customers as shown in Section IV.

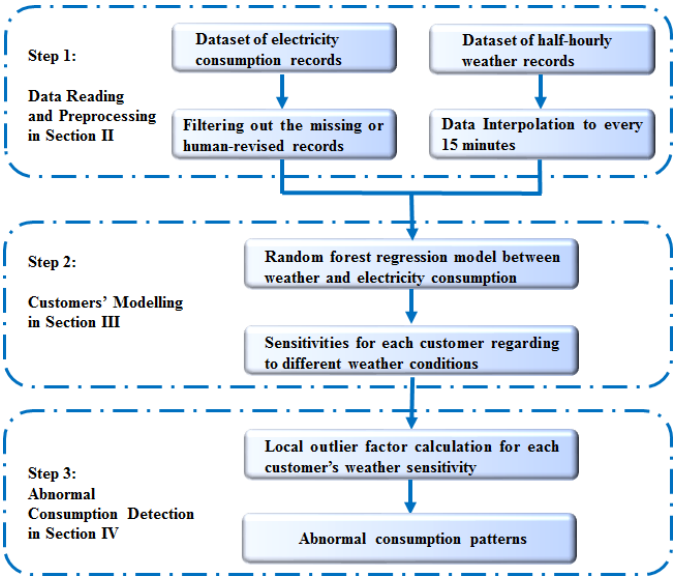


Figure 1. Scheme of abnormal electricity consumption patterns detection

II. DATA DESCRIPTION

In this section, the datasets used in this paper is briefly introduced, including the civilian customers' electricity consumption records collected from the local Distribution System Operator (DSO) in Turin (Italy) and the weather condition records during the same period.

A. *Electricity Data*

The records of electricity consumption in every 15 minutes for civilian customers are collected by the DSO. All the customers' information is anonymous with only a specific reference number for identification as in Figure 2.

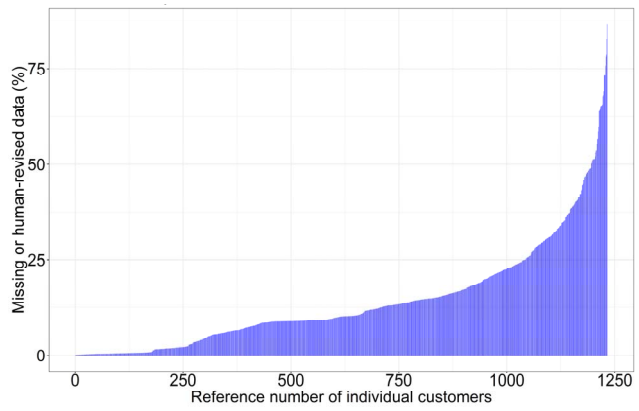


Figure 2. Quality of the civilian customers' electricity consumption records

As mentioned above, there are some missing and human-revised records in this dataset, which cannot be trusted in the regression model for sensitivity analysis. The percentage of untrusted records during the whole year for each customer is calculated and shown in Figure 2. . The civilian customers are sorted according to their percentage of untrusted records in an ascending order at the horizontal coordinate axis. Although all the civilian customers have untrusted records, over 95% of them have less than 50% missing or human-revised data. One typical electricity consumption curve of a customer during an entire week is presented in Figure 3. The missing part of the curve on Wednesday is due to the missing records in the original dataset. The red part is also from the dataset while with a special note as "revised". Since this part is far away from the typical consumption patterns, we could not use them for customers' behavior analysis.

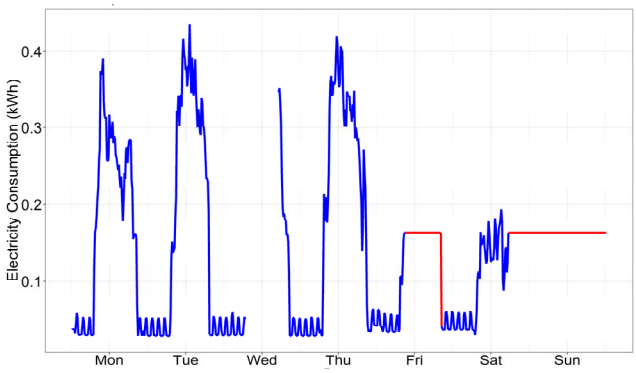
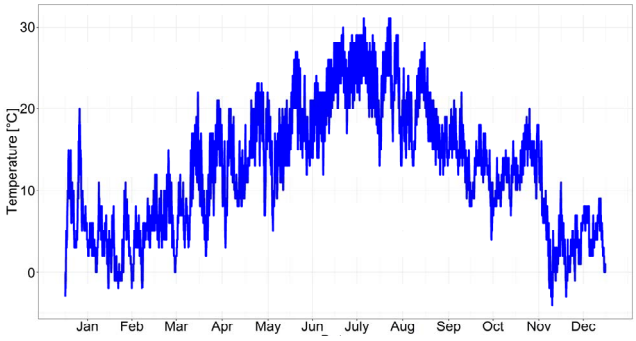


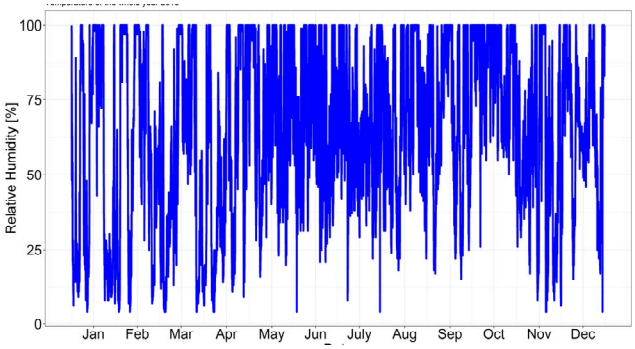
Figure 3. A typical incomplete electricity consumption curve for one week

B. *Weather Data*

The weather information covers all the year as shown in Figure 4.



(a) Temperature



(b) Relative humidity

Figure 4. A typical electricity consumption curve for one week

The weather data in this research include temperature, humidity, dew point, wind speed, atmospheric pressure and visibility at the sampling rate of half an hour. In coordination with the electricity consumption data, the linear interpolation is adopted in this work to fill the missing weather conditions in the middle of half an hour.

III. CUSTOMERS' SENSITIVITY WITH RESPECT TO THE WEATHER

Sensitivity analysis is an important tool to identify the key variables affecting the energy use from observational study [13]. The impact of climate factors on the energy analysis in buildings has already been conducted through the sensitivity analysis in some researches [10][14]. In this paper, a regression model-based sensitivity is applied in the datasets and reveals the underlying importance of different weather features to the use of electricity.

A. Regression model

Regression model-based method is a typical approach for the sensitivity analysis thanks to its clear meaning and easy interpretation. The critical process is to build an accurate regression model between the input variables and the output. In our case, each individual customer will be described with a specific regression model for the different electricity consumption characteristics. The six weather features (listed in Section II.B) at each quarter of an hour are the input variables of the model while the corresponding electricity consumption quantity during those 15 minutes is taken as the output. Since there is no need to build the regression model based on the sorted data according to sampling time, the missing and revised records at different periods in the dataset is no more a limitation for the application of the incomplete data. For a complete electricity consumption dataset, there are supposed to be 35040 records for each civilian customer. In data preparation, all the missing and human-revised records among the electricity consumption dataset are filtered out at first. In this work, the number of customers used is N_{cust} = 1150 , corresponding to the customers with more than 50% of the total records (i.e., ≥17520) valid.

Due to the complex impact of weather conditions on the use of electricity in civilian customers, it is difficult to describe the relations with traditional linear and non-linear regression models [13]. Random forest regression model, as a powerful and classic machine learning technique, is a promising solution to this kind of analysis.

Random forest is a machine learning technique stems from the concept of Classification and Regression Tree (CART) [15]. With the bootstrap resampling of data, several subsets of samples are used to train different CARTs, which is known as bagging regression trees. Compared to the other complex machine learning techniques like neural networks, the computation of CART decreases a lot by dichotomizing the input variables. The random forest algorithm in Figure 5. further introduces the randomness in subset of features used for training independent trees as weak classifiers or regressors. Finally, the results of de-correlated CARTs are collected as the final output with a significant improvement of accuracy.

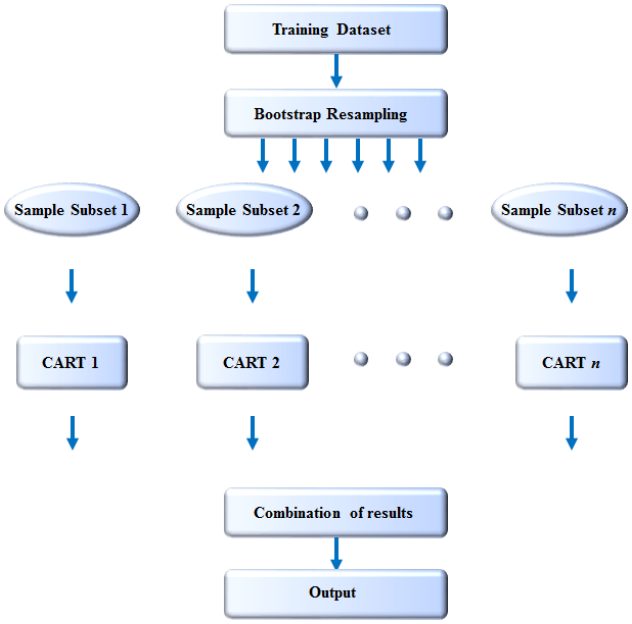


Figure 5. Typical structure of random forest

To construct the structure of random forest, the critical parameter needs to be determined is the number of CARTs. Typically, the more trees involved in the random forest, the more accurate the output would be compared to the customer's real consumption, while at a higher cost of time consuming.

In random forest algorithm, one part of the training data is reserved for each tree as the out-of-bag (OOB) samples in the bootstrap resampling step. The accuracy of regression model can be tested on OOB samples. One of the typical evaluation indices is the mean-square error (MSE) in (1):

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \quad (1)$$

where the m is size of OOB samples, y_i is the real value of the i-th sample and $\hat{y}_i$ is the output of regression model. The effect of the number of trees could be evaluated with the performance of random forest model as shown in Figure 6.

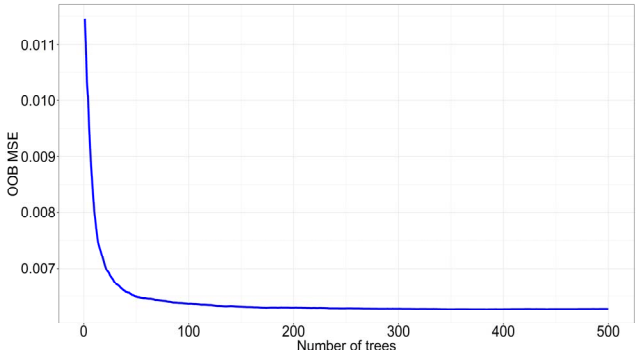


Figure 6. OOB MSE with different number of trees

In Figure 6. , a typical customer's electricity consumption records are analyzed by the random forest regression model with different number of trees. In this case, the performance of the regression model becomes stable after 100 trees.

B. *Sensitivity analysis*

In classic linear or nonlinear parameterized regression models, the impact of key input variables on the output are determined with the standardized regression coefficients [13]. However, as a non-parameterized model, there is no predetermined form in the random forest regression. Alternatively, the sensitivity of electricity consumptions to the input variables could be evaluated with the importance of variables in building the random forest regression model.

In the modeling of random forest for a specific customer, the variable's importance is also evaluated with the OOB samples by three steps:

1) Calculate the OOB MSE of the j-th CART in random forest, denoted as E_j;

2) Introduce noise randomly to the *i*-th variable, and then calculate the OOB MSE for the *j*-th CART, denoted as E_{ij}^*;

3) Calculate the importance of the *i*-th variable by considering all the *n* CARTs as in (2).

$$f_i = \sum_{j=1}^{n} \left(E_j - E_{ij}^* \right)^2 \tag{2}$$

If the accuracy (1) of random forest decreases a lot after introducing the noise into the *i*-th variable, the two OOB MSEs E_j and E_{ij}^* would be different to a large extent, which leads to a high value of f_i. This means that the *i*-th variable has a high impact on the performance of the regression model.

By repeating the evaluation steps over all the variables, the importance of weather features to a specific customer could be determined, which is also recognized as the sensitivity of electricity consumption to different variables in our case. As an example, for the customer analyzed in Figure 7. , the electricity consumption is more sensitive to the temperature, dew point and atmospheric pressure than the relative humidity.

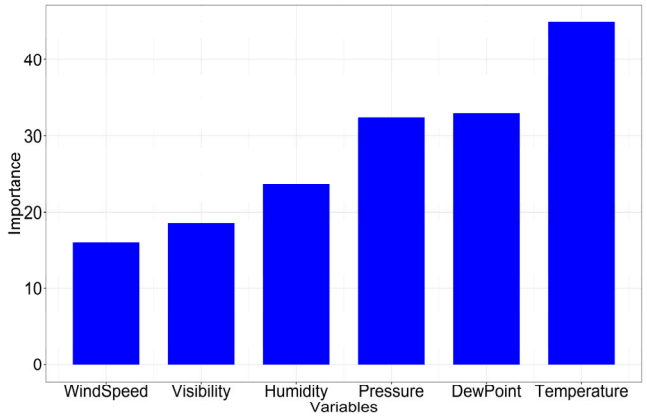


Figure 7. Variable importance of weather features to the electricity consumption

IV. Abnormal Electricity Consumption Detection

Even though the collected dataset is incomplete as demonstrated in Section II, it still contains valuable information and needs efforts to be explored. In our study, all the civilian customers could build their regression models from only the trusted part of the consumption records with the weather conditions at the same sampling frequency. Based on these regression models, the civilian customers' electricity consumption sensitivity to various weather conditions could be obtained as a complete dataset for describing their characteristics.

Different civilian customers may behave largely different to weather conditions. With the same variation of weather conditions, their consumptions may change accordingly to different degrees.

The sensitivity analysis in Section III quantifies such change in a relative value for each customer instead of the absolute value. In general, almost all the civilian customers should react more or less in regard to the environment. Therefore, the abnormal electricity consumption detection could be accomplished by locating the *outliers in the customers' sensitivities* to weather features.

Outlier detection is an important data mining method. Without the prior knowledge of normal customers' sensitivities, the distribution-based outlier detection is not much practical in this case. LOF, a density-based method proposed in [16], shows good properties by assigning to each customer a degree of being an outlier. To get the local outlier factor, the concepts for the density evaluation of object *p* is introduced as below.

In the dataset *X*, the distance between *o* and all the other objects *p* are calculated and sorted in an ascending order. The *k*-distance of object *o* is defined as the *k*-th distance between *o* and all the other objects, denoted as $dist_k(o)$. The *k*-distance neighborhood of *o* is the objects within the $dist_k(o)$. Then, the reach-distance of *p* is defined as $reach\text{-}dist_k(p,o)$ in (3).

$$reach-dist_k(p,o) = \max\left\{ dist_k(o), d(p,o) \right\} \tag{3}$$

where $d(p,o)$ refers to the distance between object *p* and *o*.

The local reachability density of *p* is defined as below in (4).

$$lrd_k(p) = \frac{k}{\sum_{o \in N_k(p)} reach-dist_k(p,o)} \tag{4}$$

Finally, the LOF index $LI_k(p)$ referring to the *k*-th distance of point *p* is defined in (5)

$$LI_k(p) = \frac{1}{k} \sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)} \tag{5}$$

The definition of $LI_k(p)$ is the ratio of the neighbors' average local reachability of point *p* to its own local reachability. If the object under test is not an outlier, this value should be very close to 1. On the contrary, for the outliers, such ratio is far from 1 due to the large distance to its

neighborhoods. The LOF index $LI_{50}(p)$, with p=1,…,N_{cust}, referring to the 50-distance (i.e., k = 50) of all the customers from our sensitivity dataset is calculated as shown in Figure 8.

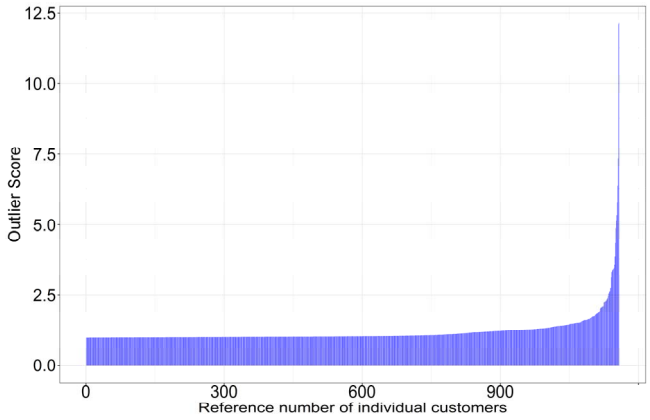


Figure 8. LOF of all the customers' weather sensitivities

As can be seen from Figure 8. , most of the customers p=1,…, N_{cust}, have a value of $LI_{50}(p)$ around 1. Only 44 customers are detected out with a $LI_{50}(p)$ larger than 2. Most of these selected customers have large period of the year with a pretty flat electricity consumption curve. It may be caused by the frozen warehouses or some apartments without any people living in for several months while with some appliances on. A further detailed investigation could be conducted on those customers for improving energy efficiencies or detecting non-technical losses.

V. CONCLUSIONS

This paper proposed a data-driven method for evaluating the weather impact on the behavior of civilian customers' electricity consumption. This work could be taken as the base for non-technical loss detection in distribution network. The first challenge in our work is to deal with the incomplete records in the dataset collected from the first-generation smart meters. Investigations on the consumption curves at the same time scale become difficult due to the randomly missing and human-revised data among different customers. Instead, we used an indirect method to study the sensitivities of individual customers only based on the trusted part of data.

As a non-parameterized model, random forest regression in this paper is used to describe the relations between electricity consumption and weather conditions. In this model, the sensitivities are evaluated by the importance of different features. All the sensitivities are believable only when the regression could accurately reveal the behavior of customers.

By building individual regression models for all the customers, their sensitivities could be obtained as a new dataset for abnormal behavior detection. Since the models are built in a customized way, the absolute quantity of power variation to different weather conditions will not affect much in the sensitivities. Neither increase nor decrease of the power demand will be regarded as the same reaction to the change of environment. However, only those customers who hardly have any reactions to the environment or changed their behavior largely are detected as the abnormal cases. Finally, limited number of abnormal customers are successfully detected out for an incomplete dataset. This information could be used by the power utilities for a further investigation.

REFERENCES

[1] The Edison Foundation, Electricity Company Smart Meter Deployments: Foundation for A Smart Grid, December 2017.

[2] Berg Insight, Smart Metering in Europe, 2018.

[3] USmartConsumer Project, European Smart Metering Landscape Report, 2016.

[4] P. M. Quevedo, J. Contreras, A, Mazza, et al, "Reliability assessment of microgrids with local and mobile generation, time-dependent profiles and intraday reconfiguration", *IEEE Trans. Indus. Appl*., vol. 54, no. 1, pp. 61-72, 2018.

[5] Y. Zhang, W. Z. Wu, C. H. Huang, "Steady-state analysis for isolated three-phase induction generator system with asymmetric loads", Dianli Zidonghua Shebei/Electric Power Automation Equipment, vol. 37, no. 2, pp. 171-175, 2017.

[6] Sook-Chin Yip, Wooi-Nee Tan, ChiaKwang Tan, et al, "An Anomaly Detection framework for Identifying Energy Theft and Defective Meters in Smart Grids", Electrical Power and Energy Systems, vol. 101, pp. 189-203, 2018.

[7] Tang Yijia and Gao Hang, "Anomaly Detection of Power Consumption based Waveform Feature Recognition", the 11th International Conference on Computer Science and Education, August 23-25, 2016, Japan.

[8] Imran Khan, Alfonso Capozzoli, Stefano P. Corgnati, et al, "Fault Detection Analysis of Building Energy Consumption using Data Mining Techniques", Energy Procedia, vol. 42, pp. 557-566, 2013.

[9] Hanxiang Ding, Kun Ding, Jingwei Zhang, et al, "Local Outlier Factor-based Fault Detection and Evaluation of Photovoltaic System", Solar Energy, vol. 164, pp. 139-148, 2018.

[10] Peter de Wilde, Wei Tan, "Predicting the performance of an office under climate change: a study of metrics, sensitivity and zonal resolution", Energy and Buildings 2010;42:1674–84.

[11] Gemma Manache and Charles Melching, "Identification of Reliable Regression and Correlation-based Sensitivity Measures for Importance Ranking of Water-Quality Model Parameters", Environmental Modeling and Software, vol. 23, pp. 549-562, 2008.

[12] Sayanti Mukherjee and Roshanak Nateghi, "Climate Sensitivity of End-use Electricity Consumption in the Built Environment: An Application to the State of Florida, United States", Energy, vol. 128, pp. 688-700, 2017.

[13] Wei Tian, "A Review of Sensitivity Analysis Methods in Building Energy Analysis", Renewable and Sustainable Energy Reviews, vol. 20, pp. 411-419, 2013.

[14] Wei Tian, Pieter de Wide, "Uncertainty and Sensitivity Analysis of Building Performance using Probabilistic Climate Projections: A UK Case Study", Automation in Construction, vol. 20, pp. 1096-1109, 2011.

[15] Leo Breiman, "Random forests", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[16] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, et al, "LOF: Identifying Density-based Local Outliers", the 2000 ACM SIGMOD international conference on Management of data, Dallas, Texas, USA — May 15 - 18, 2000.