Energy meters in District-Heating Substations for Heat Consumption Characterization and Prediction Using Machine-Learning Techniques

Mikel Lumbreras[a], Roberto Garay[b], Antonio Garrido[b]

[a] ENEDI Research Group, Dpto. de Ingeniería Minera, Metalúrgica y Ciencia de los Materiales, Escuela de Ingeniería de Bilbao, University of the Basque Country UPV/EHU, Rafael Moreno Pitxitxi 2, Bilbao 48013, Spain
[b] TECNALIA, Basque Research and Technology Alliance (BRTA), Bizkaia Science and Technology Park, Astondo bidea 700, Derio, Spain

Contact: mikel.lumbreras@ehu.eus

Abstract. The use of smart energy meters enables the monitoring of large quantity of data related to heat consumption patterns in buildings connected to DH networks. This information can be used to understand the interaction between building and the final users´ without accurate information about building characteristics and occupational rates. In this paper an intuitive and clarifier data-driven model is presented, which couples heat demand and weather variables. This model enables the disaggregation of Space-Heating & Domestic Hot water demand, characterization of the total heat demand and the forecasting for the next hours. Simulations for 53 building have been carried out, with satisfactory results for most of them, reaching R^2 values above 0.9 in some of them.

1. Introduction

Information and Communication Technologies (ICTs) are changing the way of working of many sectors within technology, revolutionizing the energy management. The computational capacity has rapidly increased in the last few decades, enabling a novel way of running simulations, named as data driven models based on a large amount of information. Regarding energy systems, the core technology that is currently use for collecting the necessary data is the smart energy meters. These devices enable the gathering of energy consumption, among other thermal variables such as temperature, with a specific time frequency, typically 15 min or 1 hour [1]. The drivers behind the employment of smart metering include energy load management, peak or demand reduction, fraud reduction, accurate billing and water conservation [2].

In the past decade, the interest of energy meters was focused on electricity smart grids where the introduction of these devices at household level enabled detailed recording of consumption. The study of the gathered data enables to the electricity operators to identify patterns of use at consumer level and this way optimizing electricity grid functioning and allowing to integrate flexibility techniques into the entire system. However, this study is focused on the possibilities that recent energy meters offer to District-Heating (DH) system and the heat flow management.

DH networks are responsible for covering the 13% of the total heat demand in the EU [3]. Although most of the DH networks are located in the northern countries due to their large load for Space-Heating (SH), [4] & [5] show the cost effectiveness of these heat supply system in other locations. Traditionally, DH systems have been based on a large production plant, typically a Combined Heat & Power (CHP) plants due to the advantages of producing electricity and heat at the same time. DH system, as well as other energy system need to evolve and adapt to new requirements, reaching the 4th Generation DH systems or 4GDH.

The so-called 4GDH introduce the design and operation improvements necessary to reach the de-carbonization objectives of the heat supply within the EU. The introduction of Renewable Energy Sources (RES) into the heat production mix is the core of the 4GDH. These new generation DH networks are also known as Ultra Low Temperature or ULT DH because the temperature of the supply line is supposed to be reduced up to 45ºC, when the current network supply heat typically above the 75ºC. The reduction of supply temperature enables the introduction of low-grade RES such as the solar thermal system, geothermal heat-pumps, waste heat streams form Data-centers etc. Moreover, a lower temperature gradient in the distribution pipelines reduce heat losses, optimizing the heat distribution system.

DH networks, as well as electricity grids, needs to evolve into active systems where the instant heat load that needs to be satisfied controls the heat production and vice versa. Gradual introduction of RES into the heat production mix also incorporates a degree of uncertainty, due to the changing conditions of weather variables. This makes even more important the information that can be achieved from the energy meters about the consumption patterns in the different building that conform the overall network.

This paper presents a data driven model based on machine-learning techniques, such as classification and regression, that allow the characterization of the total heat load in a building and enabling the heat load forecast for the next time-steps.

The rest of the paper is divided into 5 sections. In the next chapter, a literature review that summarizes the state of art of data-driven models in terms of DH networks is presented. Then, different methods used along the study are shown, followed by the analysis and results chapter. To conclude, the conclusions from the study are drawn, indicating as well, which are the possibilities for the future that can follow this study.

2. Literature Review

The field of application of data mining techniques is very extensive, however, this paper is focused on the application of identifying energy clusters and interesting distributions along the heat consumption that explain the building behavior.

Data-driven models focused on buildings´ energy consumption are widely applied to either profile energy consumption patterns (i.e. data-driven classification models) or estimate the building energy demand (i.e. prediction models) [6]. The most used data-mining techniques are clustering, classification and regression. Clustering and classification are usually applied to characterization problems, since they enable the classification or partition of the whole dataset into different groups or clusters with similar patterns. On the other hand, prediction problems are usually solved by regression solutions. Other algorithms such as the Neutral Networks (NN) are also used for prediction problems.

Regarding clustering processes, K-mean clustering is the most used algorithm which is based on calculating and optimizing the distance between K centroids that represents each the clusters. This way and based on the K-means clustering, [7] study the heating data from 139 single-family houses in Denmark concluding that as the same way than for electricity, heat consumption also shows different profiles for weekdays and weekends. Regarding supervised techniques, Decision-trees (DT) enable the data partition by a variable that is currently known.

There are several regression models, since the simplest unique-variable regression, up to the ones based on support vector machines. Regression models try to find the relation, linear or not, between the output (the variable that has to be predicted) and the rest of variables names as predictors. It is difficult to find studies about heat demand regression, but [8] investigates the importance of long term forecast of the electric and heat load in non-residential building through regression models.

In summary, there are several data-mining techniques available for the analysis of heat loads in DH context. This paper will be focused on supervised techniques and in detail, multivariate linear regression combined with DT.

3. Methodology

This study, as well as all the data-mining algorithms are based on a similar way of application. This methodology is shown in the following image.

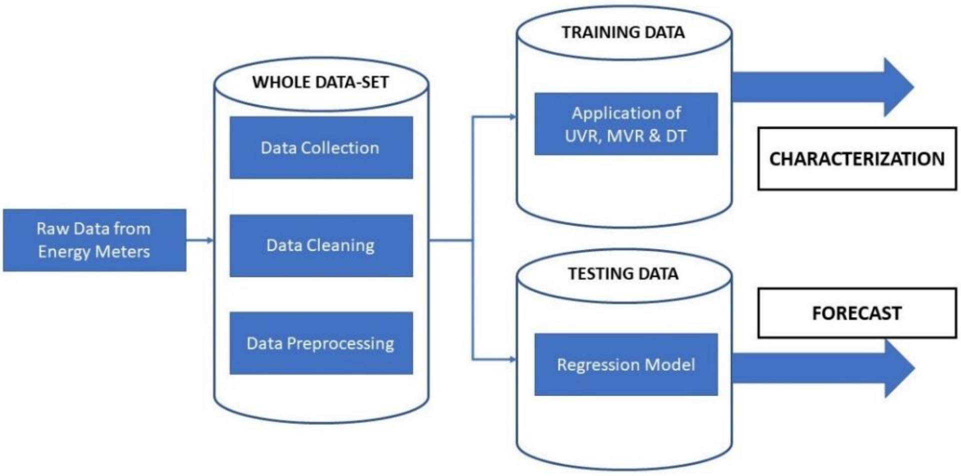

Figure 1. Methodology for data-mining application to a data-set.

The methodology applied in this study can be listed in the following steps:

- **Step 1:** The data received from the smart energy meters usually contains anomalous and missing values. The outliers have to be removed to ensure a high-quality data. This data can also be preprocessed and converted into a different time-step data. For example, if data is received with an hourly frequency, this can be converted into daily data if transient effects along the day are wanted to be removed.
- **Step 2:** The whole data set is divided into training and testing data. In this step, it is necessary to avoid diving data into two groups with different consumption patterns. For example, it has to be avoided that training data coincides with summer and testing data with winter or that only some days of the weeks are included etc. Training and testing data may be carefully chosen.
- **Step 3:** The training data is used for the application of the data-mining algorithms. In this step, the objective is to optimize the characterization of heat loads by applying the necessary modelling. This step can also be used to identify patterns that can improve the model.
- **Step 4:** The model trained with training data is used with the testing data. This way it is possible to know the accuracy of the model for prediction by comparing real data with the results obtained with the model.

All the calculations have been made in R [9], using adequate libraries for each of the applications.

3.1. Data Description

This chapter will describe the data used in this study. Two main sources are used: Heat load data from energy meters and weather data taken from an online query from the University of Tartu [10].

Data from energy meters in DH substations is kindly supplied by Fortum Tartu (Estonia), which is the operator of this large DH network. The received data corresponds to the hourly total heat consumption in 53 different buildings in the network. The buildings are represented by an ID completely independent from the real address, avoiding any type of identification problems. The energy meters are situated in the secondary side of the DH substation, gathering the total heat consumption (Space-heating (SH) + Domestic Hot water (DHW)) in the building. These readings are collected since the 1st of January 2019 and they are still metering and sending information. This way, for this study 8 months of hourly data (since January to October) is used, including this way summer and winter period.

It is known that heat load consumption, prominently, SH consumption is function of the weather variables of that moment. This is way, the models proposed in the following chapters rely on the relation between weather and heat demand variables. In the online query supplied by [10], data from 1999 to 2019 is available for all the climatic variables with a 5 minutes frequency. From all the variable available and considering the scope of this study, 4 climatic variables are used: Outdoor temperature [ºC], solar radiation [W/m2], wind speed [m/s] & the wind direction [-]. In the following image, correlation matrix between all the variable in the system is shown.

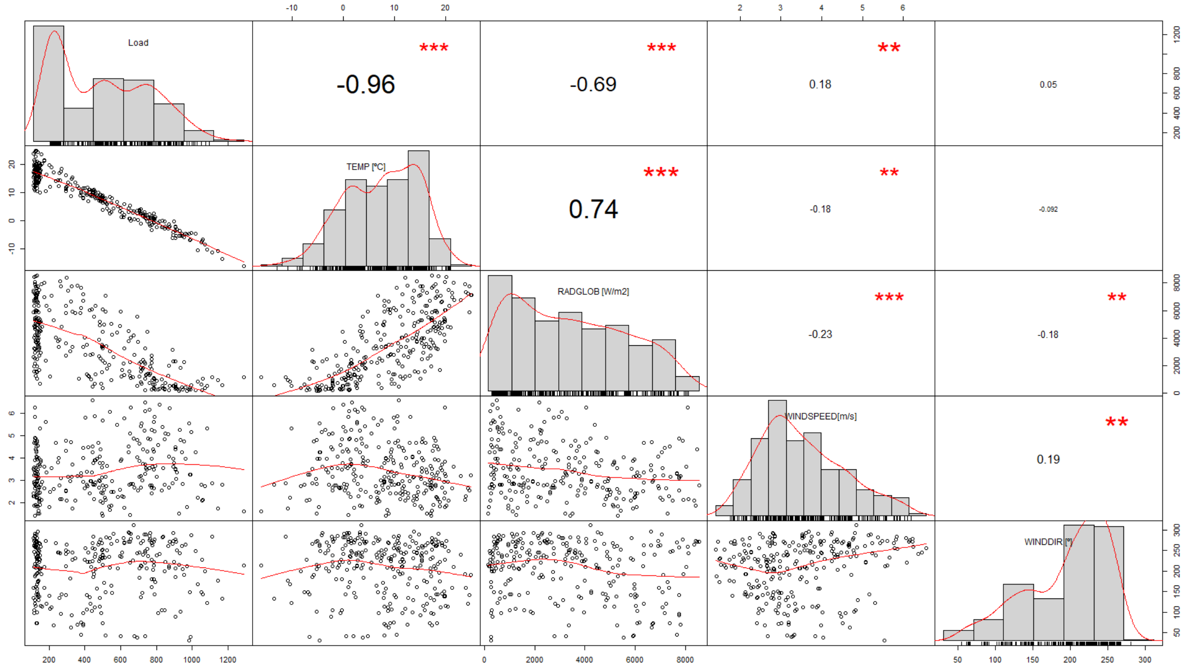

Figure 2. Daily correlation matrix between heat loads and weather variables

The whole data previously commented is coupled by the time variable. So, the whole data set is divided into training and testing data. To do so, 50/50 distribution is made, separating the data into even and odd days. Even days are used as training data and for characterization purposes, whereas odd days are used to test the model in prediction purposes.

3.2. T algorithm & Q Algorithm for building Characterization

The basis for both algorithms is the same and consists in a loop that divides data by a moving reference variable. The flowchart of this algorithm is shown in the Appendix.

As commented before, part of the demand is not dependent on weather variables. The part of the demand that corresponds with DHW is function of the user’s habits, independent from weather and it can be considered as a relatively constant value in daily and lower frequency. This is way, the T algorithm proposes to divide data by a determine temperature, TREF. This temperature coincides with the temperature that activates the SH demand. Above this temperature, it is supposed that there is only DHW consumption. This way, equation (1) is proposed.

$$\mathrm{T_{alg}} \quad \mathrm{Q} = \begin{cases} \alpha_1 + \alpha_2 \text{ x } \mathrm{T_{OUT}} + \alpha_3 \text{ x } \mathrm{G_T} + \alpha_4 \text{ x } \mathrm{W_S} + \alpha_5 \text{ x } \mathrm{W_D}, & \mathrm{T} < \mathrm{T_{REF}} \\ \alpha_0, & \mathrm{T} \geq \mathrm{T_{REF}} \end{cases} \qquad (1)$$

However, this model is not able to characterize part of the demand. This is way, a new algorithm is proposed, defined by equation (2).

$$Q_{alg} \quad Q = \begin{cases} \alpha_1 + \alpha_2 \text{ x } T_{OUT} + \alpha_3 \text{ x } G_T + \alpha_4 \text{ x } W_S + \alpha_5 \text{ x } W_D\,, & Q < Q_{REF} \\ \alpha_0, & Q \geq Q_{REF} \end{cases} \tag{2}$$

In Q-algorithm data is divided by Q_{REF}, turning into another output variable of the model. This way and observing Figure 3, data are divided by a horizontal equation. Below Q_{REF}, only DHW is consumed whereas above this constant, there is DJW & SH demand.

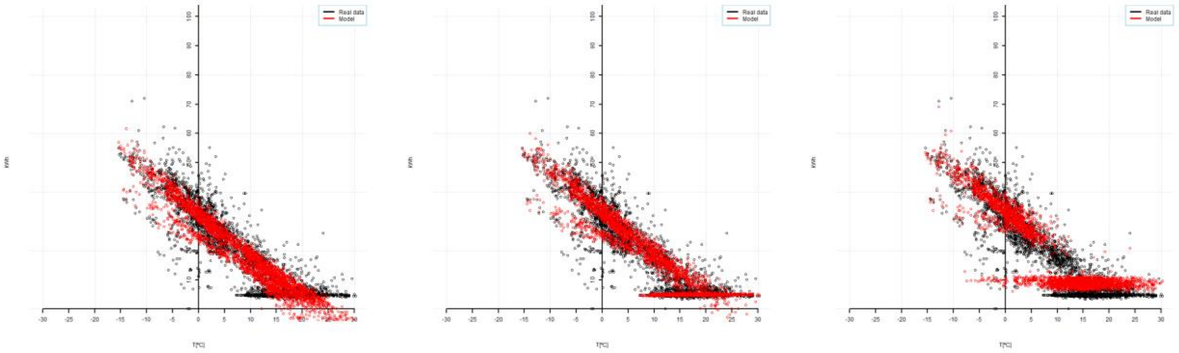

Figure 3. Q-algorithm for different Q_{REF} applied for one building

This Q-algorithm is applied for both, hourly and daily data.

Moreover, time clusters are recognized by data analysis. Different patterns for weekdays/weekends (in daily data) are observed and night setback is also observed in hourly data. This is why, 2 different DT are applied.

- **DT1:** The whole data is divided by the day of the week (Mon, Tue, Wed…)
- **DT2:** The whole data is divided by the hour of the day (0AM, 1AM, 2AM…)

This way, equation (2) is applied to each of the clusters. In daily characterization of the buildings, DT1 is only applied to the Q-algorithm, whereas hourly characterization requires the application of both, DT1 & DT2 decision trees.

3.3. *Heat Load Forecast/Prediction*

Once that all the coefficients have been achieved from the application of the Q-algorithm in the training data, these coefficients are now applied to the testing data. If the model simulates correctly the thermal behavior of the building, the testing data may validate the model accuracy. The prediction capacity also relies on the quality in the weather variables forecasting and outdoor temperature above all.

3.4. *Calculation of model accuracy*

The assessment of the proposed models for the characterization and prediction of the buildings´ heat load under study may be enforced to some common metrics.

R squared value is the statistical measurement of the correlation between the predictors and the regressed variable, with the following mathematical presentation

$$R^2 = 1 - \frac{SSE}{SSYY} \tag{3}$$

$$SSE = \sum_{i=1}^{N} (\, Y_i - X_i \,)^2 \tag{4}$$

$$SSYY = \sum_{i=1}^{N} \left(Y_i - mean(X)\right)^2 \qquad (5)$$

Where, Yi is the value got from the regression, Xi is the real value for the total heat consumption and N is the size of the testing data. This value

4. Analysis & Results

The analysis of the result got from the application of the Q-algorithm model relies on the analysis of the evolution of the R^2 in the different buildings connected to the DH network.

In Figure 4 characterization results in hourly frequency is shown, comparing real data (black points) with the data coming from the model (red points) against the outdoor temperature.

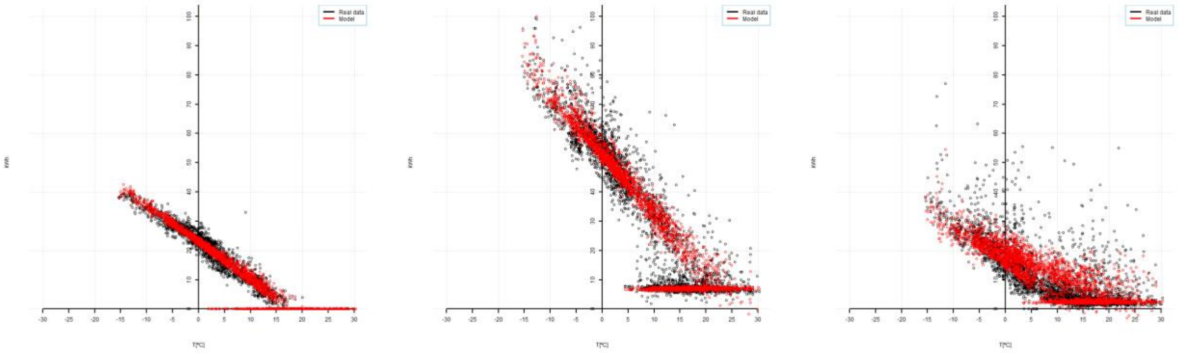

Figure 4. Characterization results for 3 different buildings in hourly frequency

Regarding characterization of the daily aggregated data, Figure 5 shows the characterization results for the same building than in Figure 4.

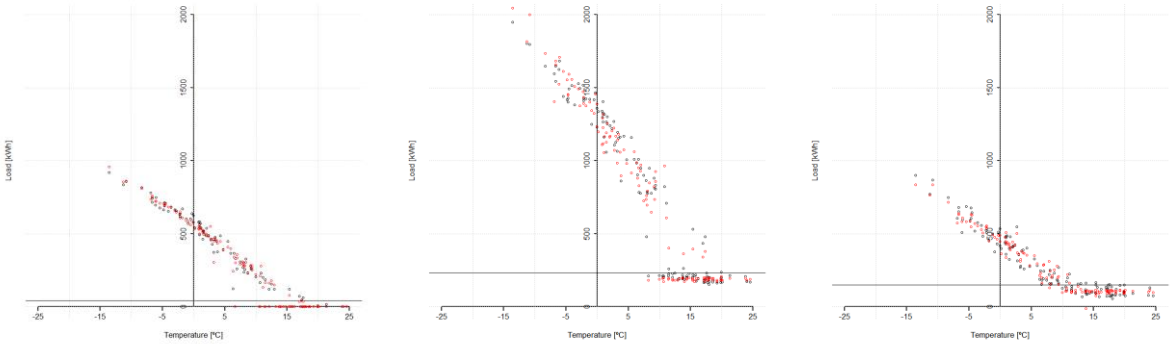

Figure 5. Characterization results for 3 building in daily frequency

Finally, when the model is applied to predict heat load, the results for R2 for daily and hourly data are shown in the following image. Each of the building is named with its respective ID.

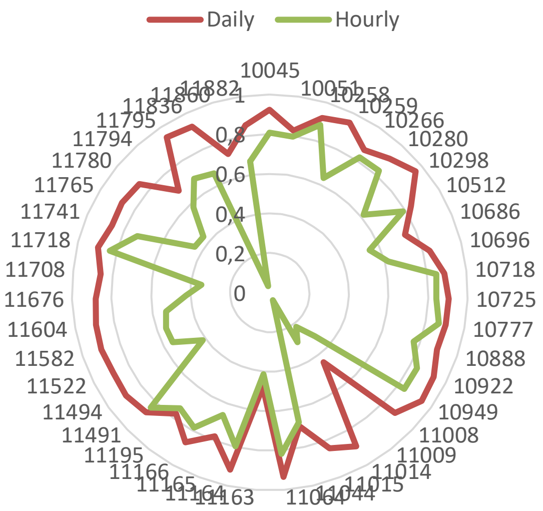

Figure 6. R2 values for daily and hourly prediction

5. Discussion

The main objective of this study has been to present a methodology with simple data mining technique that enable both, characterization and prediction of heat loads in buildings connected to a DH network. 53 different buildings have been analyzed, with different uses and characteristics, some of them with DHW consumption and some of them without it.

For this purpose, 2 different models are presented. Firstly T-algorithm that have been proved not to be very satisfying, but the Q-algorithm have been demonstrated to work quite well.

The model proposed in this paper is simple and clear, enabling the identification of both DHW & SH consumption.

The presented data-driven model could be applied for different purposes. On the one hand, it can be used for heat load prediction in DH context. When all the buildings in a district are characterized and using climatic forecasting, heat load that needs to be satisfied can be calculated and this why, produce only the required energy every moment, optimizing energy costs at production level. Other objectives, such as energetic analysis or pattern recognition of energy consumptions are identifies among others.

Observing daily data, different patterns of consumption have been identified for weekdays and weekends and in consequence, DT1 have been applied. In hourly data, two working modes have been identified, day and night setback. To model this difference in working modes and the transition between these two functioning modes, DT2 have been applied.

The training data have been used for the characterization purposes, whereas the testing data have been used for forecasting heat load. As the model have been calculated with the testing data, characterization results are more accurate than prediction.

Observing results for R^2 in prediction, daily data show better results. When aggregating data to daily frequency, transient effects caused by the different patterns of use along the day are avoided, resulting a in simpler model, easier to predict. In daily data, values of R2 from 0.44 up to 0.94 are obtained which can be considered very accurate results. However, in hourly frequency data, sue to the difficulty of predicting every single transient effect the R2 values are lower, from almost 0 to the 0.88. Even though the R2 is very low in some of the buildings, the results are not so inaccurate. When there is an outlier, and the prediction is far away from reality, the R2 value decreases drastically, even though the rest of the points are near the real data.

The consumption shape in each of the buildings is completely different and this is why, the error analysis should be particularly analyzed. In most of the cases the model slightly overestimates in summer period, what, from a DH operator perspective is not a critical problem, since the minimal energy requirement and in consequence, thermal comfort is ensured at final user level.

6. Conclusions & Future Work

Smart meters have increasingly been used for monitoring heating consumption. Data-driven models that enable the heat load forecasting in short time period will be very useful for DH operators,

For future work, the introduction of unsupervised machine learning techniques is proposed. Even though the complexity of the model will increase, and the intuitiveness of the model will decrease, effects and patterns that actually are not recognized may be identified.

References

[1] Liu X, Golab W, Ilyas IF. Benchmarking smart meter data analytics. In Proc of the 18th international conference on extending database technology: 2015. p. 385-96.

[2] Darby S. Smart metering: what potential for householder engagement? Building Research & Information 2010, 38(5), 442–457.

[3] Sven Werner, International review of district heating and cooling, Energy, Volume 137, 2017, Pages 617-631, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2017.04.045.

[4] Ehsanul Kabir, Pawan Kumar, Sandeep Kumar, Adedeji A. Adelodun, Ki-Hyun Kim, Solar energy: Potential and future prospects, Renewable and Sustainable Energy Reviews, Volume 82,Part1,2018, Pages 894-900, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2017.09.094.

[5] Henrik Lund, Renewable energy strategies for sustainable development, Energy, Volume 32, Issue 6, 2007, Pages 912-919, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2006.10.017.

[6] Yixuan Wei, Xingxing Zhang, Yong Shi, Liang Xia, Song Pan, Jinshun Wu, Mengjie Han, Xiaoyun Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renewable and Sustainable Energy Reviews, Volume 82, Part 1, 2018, Pages 1027-1047, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2017.09.108.

[7] Carolina Madeira R. do Carmo, Toke Haunstrup Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy and Buildings, Volume 125, 2016, Pages 171-180, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2016.04.079.

[8] K.B. Lindberg, S.J. Bakker, I. Sartori, Modelling electric and heat load profiles of non-residential buildings for use in long-term aggregate load forecasts, Utilities Policy, Volume 58, 2019, Pages 63-88, ISSN 0957-1787, https://doi.org/10.1016/j.jup.2019.03.004.

[9] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[10] University Tartu, Institute of Physics, Laboratory of Experimental Physics. http://meteo.physic.ut.ee/?lang=en (Visited in 2019)

Acknowledgments

The authors would like to thank to Fortum in Tartu who enabled the distribution of the smart meter data to them.

Research output presented in this paper has performed under project RELaTED (RELaTED, 2017). RELaTED project partners are: TECNALIA, Danish Technical Institute, Fortum Tartu, Beogradske Elektrane, Basque Government, Metro Therm, Nibe, Aventa, Industrias IMAR, Basque Energy Agency, Mazovia Energy Agency, Institute of Baltic Studies, FEDARENE, El Taller De Comunicación y CIA.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 768567.

Appendix A

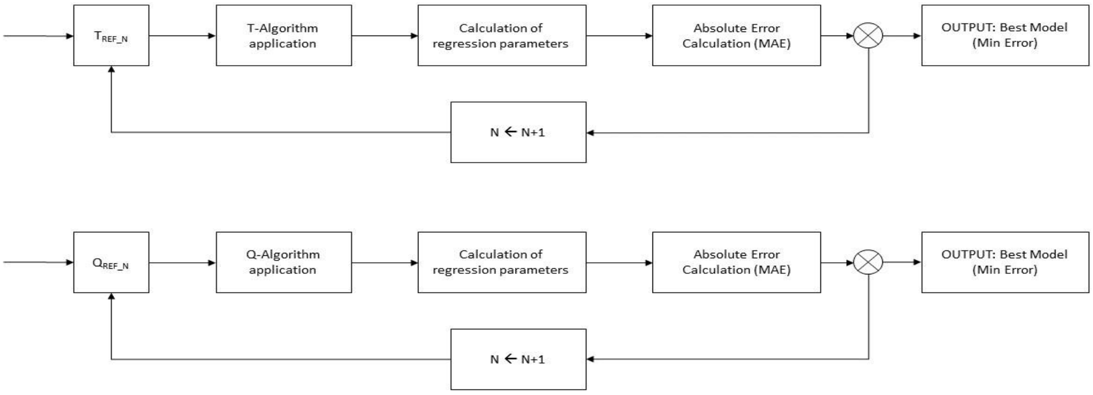
T_{REF_N}
T-Algorithm application
Calculation of regression parameters
Absolute Error Calculation (MAE)
OUTPUT: Best Model (Min Error)
N ← N+1
Q_{REF_N}
Q-Algorithm application
Calculation of regression parameters
Absolute Error Calculation (MAE)
OUTPUT: Best Model (Min Error)
N ← N+1

Figure 7. Algorithm scheme for T & Q algorithm