Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 2071–2079

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Model-based Outlier Detection in District Heating Systems

Roberto Garay-Martinez[a]* , Muhammad Talha Siddique[a] , Juan Manuel Lopez-Garde[a]

[a]Institute of Technology, Faculty of Engineering, University of Deusto, Avda. Universidades, 24, Bilbao 48007, Spain

Abstract

Data-driven methods are increasingly popular for building energy performance assessment. For these to be useful, it is required that good quality data is created through the filtering of outliers and imputation of missing information. In the field of energy use in buildings, there is a clear sensitivity of heat load to outdoor climate, which needs to be considered when identifying outliers and developing imputation methods.

We propose to use a well-known changepoint model to define the sensitivity of the data to climate, further segmented by time of the week. Then we use the residuals of the model to identify outliers, where those observation with residuals substantially out of the normality expectations are identified as outliers. Then missing data is repaired by means of linear imputation techniques, considering the patterns for same times of the week in the dataset.

As a result of the full process, we were able to identify 5% of outliers, which resulted in the improvement of model metrics in the range of 20% mean absolute error (MAE) and slightly better R2 values.

Keywords: Building Energy Performance; Data-driven model; Outlier Detection;

1. Introduction

Energy use in buildings contributes significantly, up to 40%, to the overall energy consumption in the European Union (EU) [1]. Given this substantial share, enhancing energy efficiency in buildings stands out as a pivotal objective in the EU's strategy to decarbonize the economy [2], [3]. This emphasis underscores the critical role that buildings play in achieving sustainability goals by curbing energy demand and reducing greenhouse gas emissions.

* Corresponding author. Tel.: +0-349-4413-9000 -Ext 2045
E-mail address: roberto.garay@deusto.es

1877-0509
10.1016/j.procs.2024.09.646

Nomenclature

4GDH	4th Generation District Heating
DH	District Heating
EU	European Union
IQR	Interquartile range
MAE	Mean absolute error
RMSE	Root mean square error
ST	Solar thermal
ULT	Ultra Low Temperature
WH	Waste heat

District heating (DH) networks are crucial components of urban infrastructure in cold countries where heating demand is relatively high. These DH networks presently fulfill approximately 13% of the EU's total thermal energy demand [4]. These networks have evolved significantly, with a notable shift towards lower supply temperatures. Initially, DH systems operated at temperatures exceeding 80°C; however, the introduction of the 4th Generation District Heating (4GDH) or Ultra Low Temperature (ULT) DH networks has led to heat supply at temperatures around 45°C [5], [6]. This advancement has facilitated greater integration of low-grade energy sources like solar thermal (ST) systems [7] or waste heat (WH) streams [8], [9], [10] into the heat network.

The growing significance of renewable energy sources in 4GDH systems introduces greater variability in the heat generation profile of heat production facilities. Consequently, there is a need to implement energy generation flexibility techniques to align heat production with demand in the network. Achieving this goal necessitates the use of accurate characterization methods for heat loads. These methods ensure that available energy sources are managed effectively, taking into account external variables such as weather conditions [11].

For the purpose of heat load characterization, heat meters are becoming increasingly prevalent in buildings, enabling the measurement of thermal energy consumption for each consumer connected to the heat network [12], [13]. Modern devices offer the capability to gather energy and operational data on an hourly or sub-hourly basis, and they can communicate continuously with the DH utility. Remote access to this data has led to the development of various energy management systems for heat production in DH networks [14], [15], which rely on frequent readings from smart heat meters at the consumer level. These systems typically include short-term forecasting capabilities, usually spanning a few hours or days.

Apart from forecasting, heat meters can also aid in heat load assessment and operating DH systems. They help optimize system design, sizing, and capacity planning by monitoring heat consumption and identifying inefficiencies. In DH systems operation, heat meters ensure billing accuracy, detect heat losses, meet regulatory requirements, and help detect anomalies for timely maintenance. The benefits of data from heat meters are granularity, ability to capture seasonal variation and real-time insights.

The presence of outliers in actual measurement data records raises the need for their detailed analysis before deciding how to handle them. These observations that deviate significantly from the general pattern or expected trend in the data set [16] may be due to a variety of reasons, such as measurement errors, data entry errors, unusual events or simply natural variability in the data. Outliers can distort the interpretation of the relationships between variables, disproportionately influencing the model parameters, and leading to over-fitting or under-generalization. By removing outliers, the model can better capture the underlying patterns in the data and make more accurate predictions. Overall, the detection of outliers prior to modelling is essential to ensure the reliability, accuracy and interpretability of the resulting models and analyses, as it allows for the identification and mitigation of potential sources of bias or error. [17] notes the large amount of effort involved in dealing with missing data and outliers while [18] examines their influence on the identification of linear and non-linear systems and discusses the problems of outlier detection and data cleaning

The most common way to detect outliers is based on identifying observations that are significantly different from the majority of the data points. This is done by calculating the percentiles of the data and marking as outliers those observations that fall above or below certain threshold percentiles. Thus, values below the 5th percentile and above

the 95th percentile are sometimes marked as outliers. Another widespread method based on box plots [19, 20, 21] is to treat as outliers all values that are further than 1.5 times the interquartile range above Q3 or below Q1

However, when data are related to external variables, careful consideration must be given to how to incorporate these variables into the analysis, to consider their effects on the primary variables of interest and to interpret the results appropriately. In this case, heat loads have a strong dependence on meteorology, so observed extreme values should not necessarily be treated as outliers. External variables - in our case temperature and radiation - must be incorporated into the models to adjust the outlier detection thresholds, and thus better distinguish between genuine outliers and observations that are consistent with such external factors.

To improve model predictions, a second stage of outlier removal can be added [22]. Model-based approaches to outlier detection consist of fitting a statistical model to the data and then identifying observations that deviate significantly from the expected behavior of the model [23]. These methods assume that normal data points fit the underlying model, while outliers show patterns or characteristics that are not captured by the model. Model-based outlier detection methods vary in complexity from simple parametric models to more sophisticated machine learning techniques. In general, model-based approaches to outlier detection can be tailored to specific data characteristics and analysis requirements, making them widely applicable in a variety of domains and contexts.

Energy signature models, despite being one of the simplest types of black-box models, can yield successful results for monthly or seasonal data. These models, which are widely used in data-driven approaches, express heating energy use as a function of weather variables. The purpose of energy signature models is to characterize energy usage with the help of historical data.

These are typically formulated as linear models with regards to heating/cooling degree-days [24] or changepoint models, with regards to ambient temperature and solar irradiation [25,26,27]. The concept of changepoints is based on the idea that a dataset can be divided into distinct segments, each with its own homogeneous properties. Detecting these changepoints allows for a better understanding of the underlying processes and can lead to more accurate predictions.

An evolution of this concept are time-of-the-week segmented changepoint models [28] that capture shifts in the data pattern that occur at different hours throughout the week. The model segments the time series by hour & day and allows for changepoints within each segment, representing sudden changes in the pattern, such as the start of a new working day affecting energy consumption.

In this paper, we develop a novel model-based approach for the outlier detection on heat load data for buildings. It is based on a 3-parameter changepoint model, segmented for every hour of the week, and incorporating the influence of solar radiation, as per the work in [27]. Based on the current literature, our approach is that this model is suitable to characterize both the physics and the user behavior in buildings. Accordingly, we assume that the residuals of this model (given no structural bias) are a good indicator of the goodness of the data. We analyze the residual in a statistical way and define outliers as those pieces of the data where the residuals of the model are significantly out of bounds.

Authors believe that this model-based outlier detection is at the same time simple, extendable (to more complex changepoint model) and robust to apply in real-life cases. And to authors' best knowledge, no such methods have been reported in the state of the art.

2. Methodology

The methodology employed in this paper aims to understand and manage outliers in a dataset representing heat load patterns of a building. The dataset consists of hourly values of heat load, ambient temperature, and solar radiation over a year. The methodology involves several key steps (figure 1):

- **Data Inspection**: Initial examination of the dataset to understand its structure and identify any anomalies or missing values. Various plots are used to explore the relationships between variables, such as power consumption, temperature, and solar irradiation, over time.
- **Changepoint Model and Outlier Detection**: A changepoint model is created and segmented for each hour of the week. This model is trained to predict the expected heat load based on temperature and solar irradiation using a genetic algorithm. Observations that deviate significantly from the model's predictions are identified as outliers.

- **Data Repair**: Outliers are corrected by filling in missing values using interpolation techniques. This step helps improve the overall quality of the dataset and ensures that the subsequent analysis is based on reliable data.
- **Recompute Changepoint Model**: After repairing the dataset, the changepoint model is recomputed to incorporate the corrected values. This step ensures that the model is updated with the latest data and can provide more accurate predictions. These steps are applied iteratively to each hour of the week to fit a separate changepoint model for each hour.
- **Model Statistics**: Finally, various model statistics, such as R-squared, root mean squared error (RMSE), and mean absolute error (MAE), are calculated to evaluate the performance of the changepoint model and the impact of outlier detection and repair on the dataset.

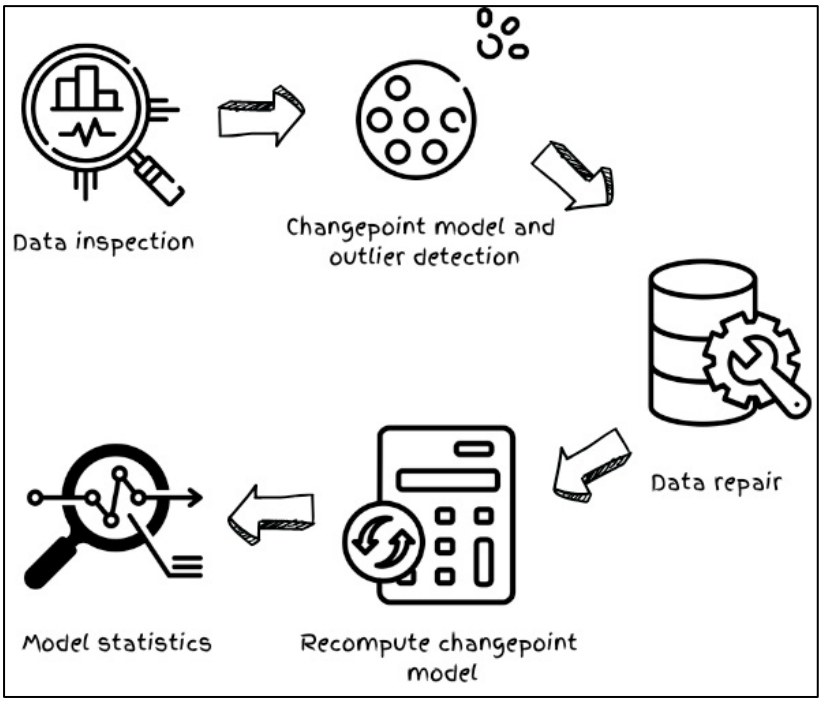

Fig. 1 Methodology

3. Dataset

The dataset used in this study involves a 1-year dataset for a multi-rise building connected to the District Heating Network in Tartu, Estonia. The dataset contains hourly values of heat load, ambient temperature, and solar radiation over a horizontal plane. The data sources for this dataset are GREN Eesti for heat load and the University of Tartu for weather data. Greater details on this dataset are available at [29,27,32].

The dataset considered for this study belongs to only one building and is structured with several key features:

- Date and Time: The dataset includes timestamps for each hourly observation, allowing for time-series analysis.
- Holiday Indicator: There is a variable indicating whether a given timestamp corresponds to a holiday.
- Temperature: Ambient temperature data is included, which is crucial for understanding its impact on heat load patterns.
- Solar Irradiation: The dataset also includes solar irradiation data, which is important as it affects the heat load of the building.
- Power Consumption: The primary variable of interest is the power consumption, which is measured in kWh

The dataset is used to analyze the heat load patterns of the building, exploring relationships between variables such as power consumption (figure 2), temperature (figure 3a), and solar irradiation over time (figure 3b).

Visualizations and statistical analyses are employed to gain insights into these relationships, helping to understand the building's energy consumption patterns.

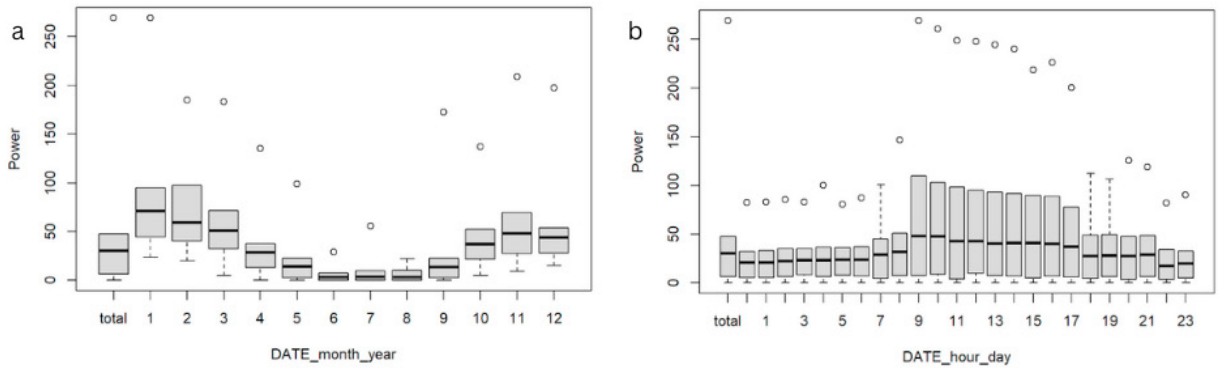

Fig. 2 (a) Monthly power vs time (b) Hourly power vs time

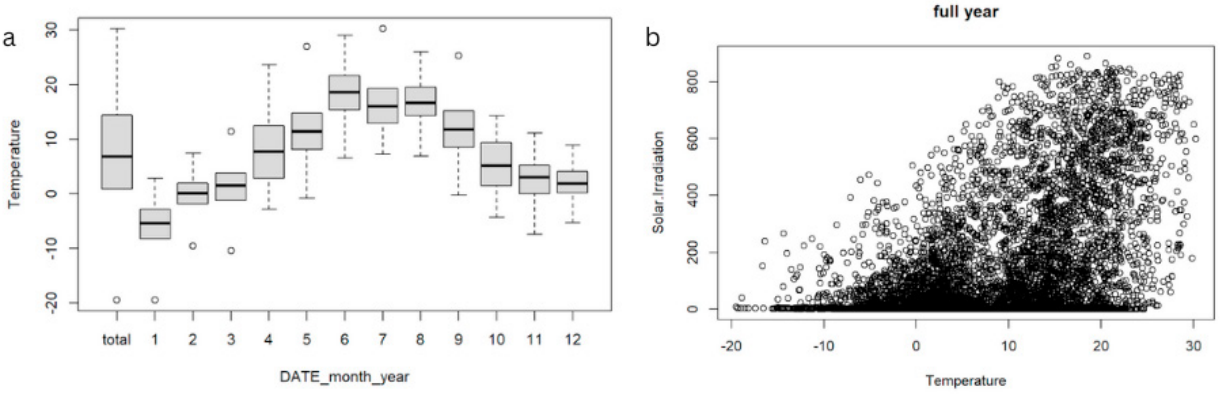

Fig. 3 (a) Temperature vs time (b) Solar irradiation vs temperature

Analyzing the plots, it is clear that the heat load is influenced by outdoor temperature and solar irradiation, with temperature being the primary factor, given its partial correlation with solar irradiation.

During hot summer months, there is a minimum heat load. In cold winter months, there is a linear relationship between heat load and temperature. Spring and autumn represent intermediate periods in terms of heat load.

Daily variations in heat load are influenced by various building-specific factors and schedules. These include typical office hours (around 9 am to 5 pm), night setbacks (reducing heating during late evening hours, around 10-11 pm), and weekly occupancy patterns.

4. Results

The changepoint model trained using the above-mentioned dataset predicts the heat load based on temperature and solar irradiation using a genetic algorithm [31]. As seen in the figure 4, Two primary patterns have been identified, each with its specific characteristics for every hour.

To detect outliers, the quartiles of each variable are computed as described in [24]. The interquartile range (IQR) is used as a criterion, where observations exceeding the third quartile plus 1.5 times the IQR or falling below the

first quartile minus 1.5 times the IQR are flagged as potential outliers. This approach for outlier identification has been widely employed in diverse research studies [20].

Upon implementation of the data analysis process, it is observed that approximately five percent of the dataset is identified as outliers. These outliers, once detected, are subsequently removed to ensure the integrity and accuracy of the data. Following the removal of outliers, the dataset undergoes a data repair process using interpolation techniques.

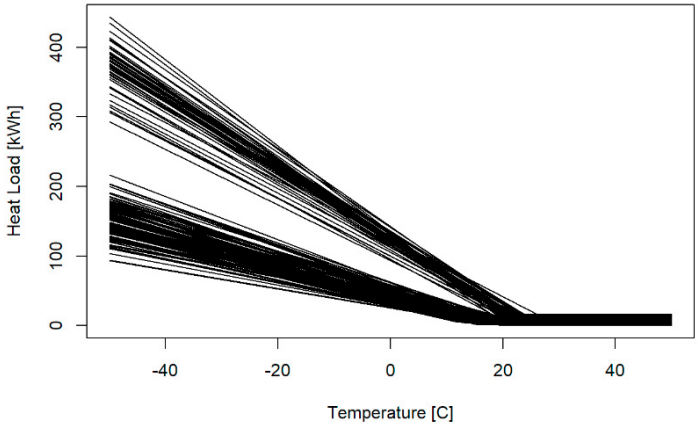

Fig. 4 Changepoint model results (before outlier removal)

Interpolation is applied to impute missing data values, especially for time intervals where data is absent. For instance, for gaps in the data that are less than 1 hour, interpolation utilizes data from the -1 to 1 hour time slot in the previous and following weeks. Similarly, for longer gaps, specifically those less than 5 hours, interpolation relies on data from the -5 to 5 hour slot in the previous and following weeks.

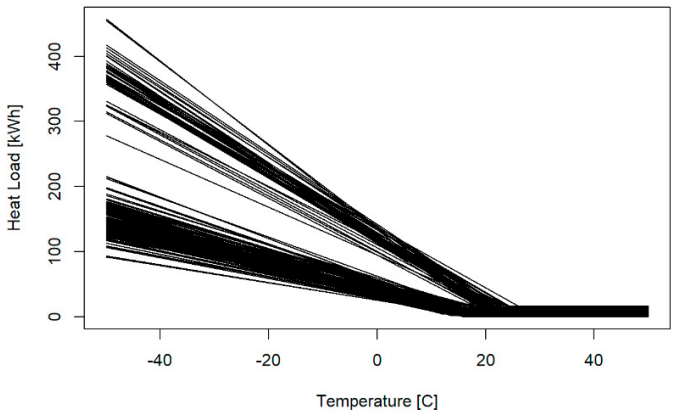

Fig. 5 Changepoint model results (after outlier removal)

This interpolation process is conducted in a stepwise manner, addressing increasingly longer time gaps. It begins by addressing 2-hour gaps, then progresses to 3-hour gaps, and continues in this incremental fashion until 30% of the outlier data is reconstructed. After this repair, the changepoint model is recalculated to produce the result presented in figure 5.

To evaluate the effectiveness of this approach towards outlier removal, R-square, Root mean square error (RMSE) and mean absolute error (MAE) are used to analyze the results (figure 6). These metrics are applied on four

different datasets developed during this study. The first dataset is the original one with all the outliers included. The second dataset is one with all the outliers removed using the changepoint model. The third dataset has the outliers removed and values obtained through interpolation filled as part of the repair process. Lastly, the fourth dataset is obtained after retraining the changepoint model.

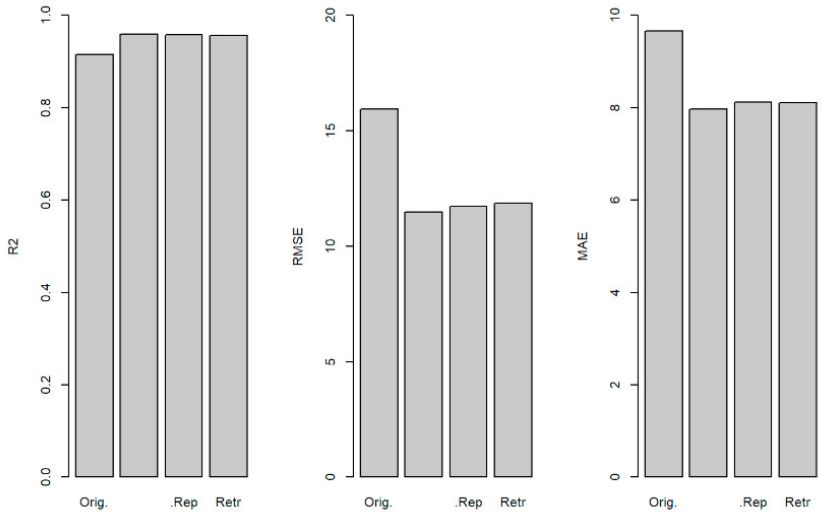

Fig. 6 Model statistics

Overall, in terms of model statistics, the performance is good. R2>0.9; RMSE<15.9, MAE<10 in all cases. Removing outliers leads to a slight improvement in performance. However, there doesn't appear to be a significant impact from repairing outliers or retraining the model.

5. Conclusions

In this paper, we have presented a novel model-based approach for the outlier detection on heat load data for buildings, based on a changepoint model, as per the work in [27]. It is proven that the model is a good approximation of the heat load already prior to the outlier removal and data imputation, with reasonable R2, RMSE and MAE values.

The identification of outliers (5% of the original dataset) and their removal produces a substantial improvement in model metrics.

The data imputation techniques used to repair the dataset, seem to properly represent the main trends of the data, as the error metrics remain constant, even for models trained before the imputation process.

We believe that the presented model-based outlier detection is at the same time simple, extendable (to more complex changepoint model) and robust to apply in real-life cases. And to our best knowledge, no such methods have been reported in the state of the art.

Code availability

The sample data and the code are available in Github and Zenodo [30].

All the code is written in R, and the scientific outcomes are directly inspectable in html output files created through rmarkdown files.

Acknowledgements

This study has been carried out in the context of the ATELIER project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 864374

The research leading to this study received funding from the Basque Goverment grant for Research Groups, “Grupos de investigación del Sistema Universitario Vasco, Departamento de Educación, Universidades e Investigación” (Research group: IT1677-22).

References

[1] Luis Pérez-Lombard, José Ortiz, Christine Pout (2008) “A review on buildings energy consumption information.” *Energy Build* 40 (3): 394-398, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2007.03.007

[2] Google Scholar. (s.f.). Directive (EU) 2018/844 of the European Parliament and of the Council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency.

[3] Orkesterjournalen. (2022) “Directive 2012/27/EU of the European parliament and of the council of 25 october 2012 on energy efficiency, amending directives 2009/125/EC and 2010/30/EU and repealing directives 2004/8/EC and 2006/32/EC text with EEA relevance”.

[4] Sven Werner (2017) “International review of district heating and cooling.” *Energy*, 137: 617 631, ISSN 0360-5442, 10.1016/j.energy.2017.04.045

[5] Henrik Lund, Sven Werner, Robin Wiltshire, Svend Svendsen, Jan Eric Thorsen, Frede Hvelplund, & Brian Vad Mathiesen (2014) “4th Generation District Heating (4GDH): integrating smart thermal grids into future sustainable energy systems.” *Energy* 68:1-11, ISSN 0360-5442, 10.1016/j.energy.2014.02.089.

[6] Haoran Li, & Natasa Nord (2018) “Transition to the 4th generation district heating - possibilities, bottlenecks, and challenges.” *Energy Procedia* 149: 483-498, ISSN 1876-6102, 10.1016/j.egypro.2018.08.213

[7] Mikel Lumbreras, Roberto Garay (2020) “Energy & economic assessment of façade-integrated solar thermal systems combined with ultra-low temperature district-heating” *Renew Energy* 159: 1000-1014, ISSN 0960-1481, 10.1016/j.renene.2020.06.019

[8] Mikko Wahlroos, Matti Pärssinen, Jukka Manner, Sanna Syri (2017) “Utilizing data center waste heat in district heating Impacts on energy efficiency and prospects for low-temperature district heating networks” *Energy* 140: 1228-1238, ISSN 0360-5442, 10.1016/j.energy.2017.08.078

[9] Jelena Ziemele, Kalnins Roberts, Girts Vigants, Edgars Vigants, Ivars Veidenbergs (2018) “Evaluation of the industrial waste heat potential for its recovery and integration into a fourth-generation district heating system.” *Energy Procedia* 147:315-321, ISSN 1876-6102, 10.1016/j.egypro.2018.07.098

[10] Google Scholar. (s.f.). Open district HeatingTM (2019). Obtenido de https://www.opendistrictheating.com

[11] Jaume Fitó, Sacha Hodencq, Julien Ramousse, Frédéric Wurtz, Benoit Stutz, François Debray, Benjamin Vincent (2020) “Energy- and exergy-based optimal designs of a low-temperature industrial waste heat recovery system in district heating.” *Energy Convers Manag*, 211 ISSN 0196-8904, 10.1016/j.enconman.2020.112753

[12] S. Darby. (2010) “Smart metering: what potential for householder engagement?” *Build Res Inf*, 38 (5): 442-457

[13] X. Liu, W. Golab, W. Golab, I.F. Ilyas. (2015) “Benchmarking smart meter data analytics.” *Proc of the 18th international conference on extending database technology* :385-396

[14] Klaus Lichtenegger, David Wöss, Christian Halmdienst, Ernst Höftberger, Christoph Schmidl, Pröll Tobias (2017) “Intelligent heat networks: first results of an energy-information-cost-model.” *Sustainable Energy, Grids and Networks*, 11: 1-12, ISSN 2352-4677, 10.1016/j.segan.2017.05.001

[15] Mattias Vesterlund, Andrea Toffolo, Jan Dahl. (2017) “Optimization of multi-source complex district heating network, a case study” *Energy* 126: 53-63, ISSN 0360-5442, 10.1016/j.energy.2017.03.018

[16] Vic Barnett, Toby Lewis (1994). Outliers in statistical Data ,3rd edition, ISBN: 978-0-471-93034-5

[17] P.-O. Gutman and B.Nilsson (1998). Modelling and prediction of bending stiffness for paper board manufacturing, *J.Process Contr* 8: 229-237

[18] R.K. Pearson (2002). Outliers in process modeling and identification, *IEEE Transactions on Control Systems Technology* 10 (1): 55-63, doi: 10.1109/87.974338

[19] R Core Team (2013), R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria

[20] Li Aihua, Feng Mengyan, Li Yanruyu, Liu Zhidong (2016). Application of outlier mining in insider identification based on boxplot method. *Procedia Computer Science* 91:245-251, ISSN 1877-0509, doi.org/10.1016/j.procs.2016.07.069

[21] Schwertman Neil C, Ann Owens Margaret, Robiah Adnan (2004). A simple more general boxplot method for identifying outliers, *Comput Stat Data Anal* 47(1):165-174, ISSN 0167-9473, doi.org/10.1016/j.csda.2003.10.012.

[22] H.Ferdowsi, S.Jagannathan, M.Zawodniok (2014). An online outlier identification and removal scheme for improving fault detection performance, *IEEE Transactions on Neural Networks and Learning Systems* 26(5): 908-919, doi: 10.1109/TNNLS.2013.2283456

[23] Jieqi Yu, Haipeng Zheng, Sanjeev R. Kulkarni, HVincent Poor (2010). Two-Stage Outlier Elimination for Robust Curve and Surface Fitting, EURASIP Journal on Advances in Signal Processing, Article number 154891, doi.org/10.1155/2010/154891

[24] Margaret F. Fels (1986). PRISM: An introduction, *Energy and buildings* 9(1-2):5-18, ISSN 0378-7788, doi.org/10.1016/0378-7788(86)90003-4

[25] Kissock, J. K.; Haberl, J. S.; Claridge, D. E. (2002). Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models, ASHRAE Research Project 1050-RP, Final Report. Energy Systems Laboratory, Texas A&M University. Available electronically from https://hdl. handle.net /1969 .1 /2847

[26] Beñat Arregi, Roberto Garay (2017) “Regression analysis of the energy consumption of tertiary buildings” CISBAT 2017, Lausanne, Switzerland. *Energy Procedia* 122: 9-14, ISSN 1876-6102, https://doi.org/10.1016/j.egypro.2017.07.290.

[27] Mikel Lumbreras, Roberto Garay-Martinez, Beñat Arregi, Koldobika Martin-Escudero, Gonzalo Diarce, Margus Raud, Indrek Hagu (2022) “Data driven model for heat load prediction in buildings connected to District Heating by using smart heat meters” *Energy* 239, Part D, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2021.122318

[28] J. L. Mathieu, P. N. Price, S. Kiliccote and M. A. Piette (2011) "Quantifying Changes in Building Electricity Use, With Application to Demand Response," in IEEE Transactions on Smart Grid 2 (3) 507-518, Sept. 2011, https://doi.org/10.1109/TSG.2011.2145010

[29] Mikel Lumbreras, Gonzalo Diarce, Koldobika Martin, Roberto Garay-Martinez, Beñat Arregi (2023) Unsupervised recognition and prediction of daily patterns in heating loads in buildings, Journal of Building Engineering, 65,105732, https://doi.org/10.1016/j.jobe.2022.105732

[30] Roberto Garay-Martinez, Heat Load Characterisation, 2023, https://github.com/robgaray/Building_Heat_Load_Characterisation, https://doi.org/10.5281/zenodo.7692351

[31] Scrucca, L. (28-01-2024). Package ‘GA’. Obtenido de https://cran.r-project.org/web/packages/GA/index.html

[32] Olaia Eguiarte, Antonio Garrido-Marijuan, Roberto Garay-Martinez, Margus Raud, Indrek Hagu (2022) “Data-driven assessment for the supervision of District Heating Networks.” *Energy Reports* 8: 34-40, ISSN 2352-4847, https://doi.org/10.1016/j.egyr.2022.10.212