Applying Machine Learning Methods and Outlier Detection to Process and Analyse Incomplete Heat Meter Data

Ulrich Trabert, Felix Pag, Janybek Orozaliev, Klaus Vajen

University of Kassel, Institute of Thermal Engineering, Kassel (Germany)

Abstract

Digitalization plays an important role to achieve optimised operation of renewable-based energy supply systems coupling the sectors heat, electricity, and mobility. The benefit derived from optimization methods, however, also depends on the quality of the used data. Therefore, the present study analyses methods which can be applied for processing of low-quality heat meter data. Machine learning algorithms are used to fill the gaps of incomplete load and return temperature multi-year profiles retrieved from a dataset of five heat meters installed in a heating grid of an industrial consumer. On average, the coefficient of determination R^2 for the machine learning algorithm is 0.92. Additionally, two statistics-based filters, that are applied to detect outliers in continuous profiles, are compared. Those filters slightly improve the accuracy of the models and help to obtain more consistent profiles from the heat meter data. The completed datasets support enhanced data analysis, which lays the ground for redesigning fossil-fuelled and optimizing smart renewable energy systems.

Keywords: Digitalization, machine learning, outlier detection, heat meter data

1. Introduction

In the wake of digitalization, the amount of available datapoints from energy systems is strongly increasing. Analysis of that data gives detailed knowledge about the systems, which is useful for improving operational stability, energy efficiency and for remodelling of existing systems towards renewable energies.

While digitalization in the electricity sector is already well advanced, the heat sector lags behind as measuring heat flows requires more complex sensor technology with higher installation costs. In combination with historically lower costs for heat than for electricity, it is less attractive to monitor heat systems in detail. Therefore, accurate measured heat consumption data in a high temporal resolution (i.e. hourly) is usually rare. Additionally, the quality of the data depends on installation and signal transmission conditions. Adverse conditions may lead to effects like incomplete or irregular heat meter data, which is also true for the dataset used in this work.

2. Heat Meter Data Processing

The following chapters propose a procedure to improve low quality heat meter data. This includes the completion of fragmentary timeseries by predicting the missing data with a machine learning regression algorithm as well as two methods for statistical outlier detection. Furthermore, the impact of outlier detection on the accuracy of the models is analyzed.

2.1 Original dataset

The original dataset used in this study consists of three years of data which come from five heat meters measuring load, supply and return temperature in hourly resolution. They belong to a heating grid of an industrial complex with several buildings. The timeseries of the example dataset have gaps that, on average, account for 29 % of the time within the three-year period. These gaps are caused by an error-prone IT infrastructure, which is used to transfer the measured values of the heat meters and store them in a central database.

As an example, Fig. 1 shows the original heat load, supply and return temperature profiles for one heat meter for the year 2020. The heat load was normalized using its maximum value within the three-year period, as the actual values of the profile are irrelevant to the methods applied in this paper. The gaps are visible as zero values in the shown graphs. They mostly occur simultaneously in all three profiles but show no obvious patterns. The duration of missing periods varies from a few hours up to multiple weeks. All five heat meters have similar characteristics in terms of missing data, but the course of load and temperature profiles deviates according to the type of consumers. Both heat load and return temperature are strongly dependent on ambient temperature, because space heating accounts for the largest share in heat demand.

doi:10.18086/eurosun.2022.16.12 Available at http://proceedings.ises.org

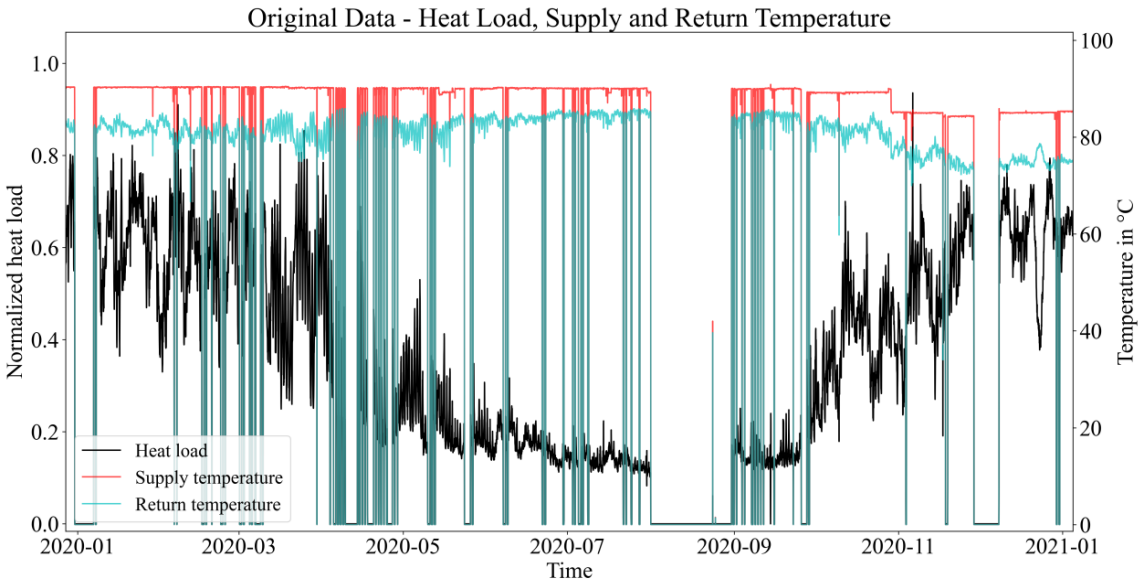

Fig. 1: Original data for a one-year period of one heat meter (heat load, supply and return temperature)

2.2 Completing missing heat meter data with machine learning models

Data analysis helps to understand and improve energy systems. However, continuous timeseries are required to prevent a bias from availability. In this work a machine learning algorithm is used to predict values for the gaps in the described profiles.

It has already been shown by Saloux and Candanedo (2018) and Bünning et al. (2020), that the use of machine learning algorithms can improve the accuracy of heat load forecasts. Filling gaps in historic load profiles is a rather similar use case and it is additionally applied to fill gaps in return temperature profiles here. The present work has been conducted using the scikit-learn machine learning library (version 0.24.0) in a python environment (Pedregosa et al., 2011). After comparing different available algorithms in scikit-learn, the Extreme Gradient Boosting (XGBoost) regression algorithm turned out to be the most suitable for filling gaps in measured data.

The general procedure for each timeseries is depicted in Fig. 2. In the first step, the input variables which are relevant for prediction of the desired outputs are chosen and included into the raw data. A preliminary visual analysis showed that ambient temperature, time, and solar radiation have a relevant influence on the load. In case of the return temperature, the supply temperature is an additional input variable. The raw data is pre-processed to ensure unbiased training of the machine learning algorithm. The timestamp is encoded (1, 0) into columns for each category (day of week, hour of day). Ambient and supply temperature show an almost normal distribution so that they are standardized, while solar radiation is normalized between 0 and 1.

Next, the gaps are identified by defining validity ranges that are applied to each profile individually. The validity ranges are:

- Heat load: $\dot{Q} > 0$
- Supply temperature: $T_{supply} = T_{supply,setpoint} \pm 2K$
- Return temperature: $T_{return} > 40\ °C$

The gaps are removed from the training dataset, which is fed into the machine learning algorithm (XGBoost) to build a black box model of the system. The trained model is then used to predict the missing values in the original data. Finally, the accuracy of the models is evaluated in order to determine their reliability. Common regression indicators are determined, including the coefficient of determination (R^2), the root mean squared error (RMSE) and the mean absolute percentage error (MAPE). They are calculated as the mean of five random data splits to ensure cross validation.

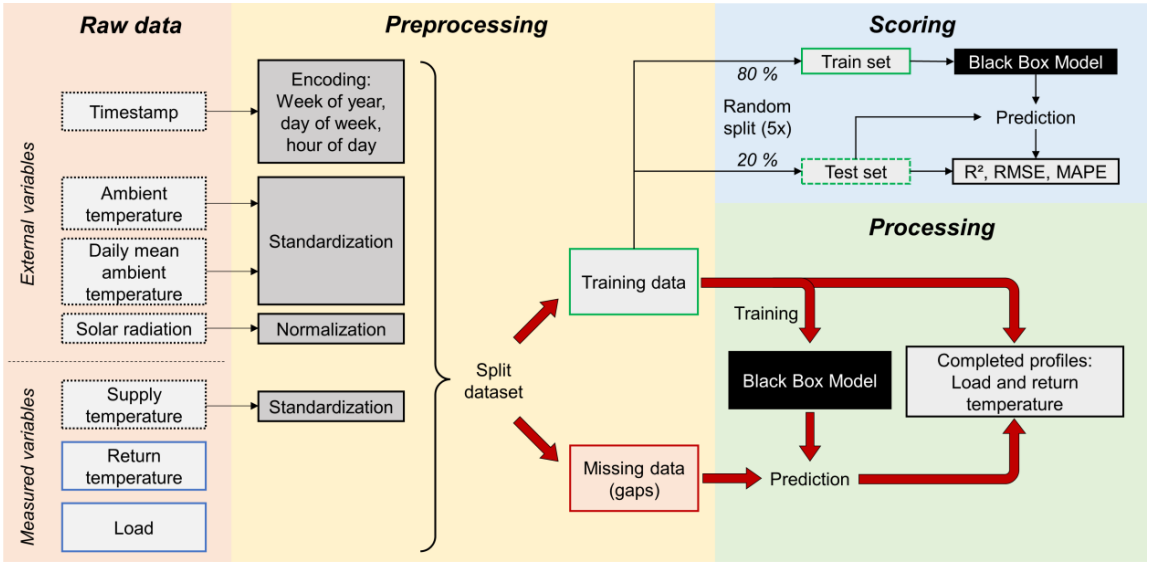

Fig. 2: Procedure diagram for completing missing data

2.3 Outlier detection through statistical deviations

Besides the obviously missing data points in a timeseries, there are also outliers where the values abnormally deviate from the course of a profile. Those can be caused for example by temporary external effects on the sensors or failures in data logging that last shorter than the actual temporal resolution of the timeseries. In this study two different methods based on statistics are developed and tested to detect outliers.

Z-filter

The Z-filter detects a value as an outlier, if it deviates from the mean of a floating period (i.e. one week) by a predefined number of standard deviations (see eq. 1). The period for the floating values (mean and standard deviation) is centred around each tested value. The period and the number of standard deviations (n_{std}) are free parameters adjustable to the relevant use case.

Z-criteria for outliers: $$\frac{|Y_i - \bar{Y}|}{\sigma} > n_{std} \qquad \text{(eq. 1)}$$

For this paper the parameters were chosen according to a visual analysis of the profile. Ideally the parameters are set by using a validation dataset, where the number of outliers is known. However, this was not available for this study.

Grubbs-filter

The Grubbs-filter uses the maximum normalized residual test (Grubbs-test) to detect outliers within a floating period (i.e. one week) centred around the tested value. An outlier is located by the maximum deviation of a value in terms of number of standard deviations from the mean within the given period (see eq. 2). This procedure is repeated until no outliers are detected at a significance level α.

Grubbs-criteria for outliers: $$G = \frac{max_{i=1,\dots,N}|Y_i - \bar{Y}|}{\sigma} \qquad \text{(eq. 2)}$$

The filter can be adjusted through the length of the floating period and the significance level. Grubbs-test only detects outliers by their value but does not consider the order of datapoints. However, the order is relevant when considering timeseries. Therefore, the Grubbs-test is applied to the second-degree gradient of a timeseries in this case. It can be used to locate outliers that deviate strongly from the course of a profile, because the first-degree gradient is low right before the outlier and high just after the outlier. Additionally, the timeseries of the second-degree gradient is approximately normally distributed, which is a prerequisite for using Grubbs-test. (Adikaram et al., 2015)

3. Results

The following sections show excerpts from the profiles to demonstrate the presented methods. The machine learning models are evaluated using R^2, RMSE and MAPE and the effects on the scores through the filters is shown. Additionally, the importance of the several input parameters is discussed.

3.1 Comparison of Grubbs-filter and Z-filter for Outlier Detection

The filters are applied after the gaps in the profiles are completed by the machine learning algorithm described in section 2.2. For the supply temperature profiles, the gaps are filled with the setpoint value, which is constant throughout the year and independent of ambient temperatures.

The floating period in both filters is set to one week (168 h) to account for the fact that the profiles are retrieved from an industrial consumer with characteristics related to the day of the week. The Z-filter is applied with n_{std} set to 2.5 and the Grubbs-filter with a significance level α at 0.05.

The supply temperature is only filtered with the Z-filter, as it works well to detect outliers in an almost constant profile. The values filtered in supply temperature profiles are replaced by the median of the two adjacent values.

The total number of detected outliers in the 15 profiles is documented in Tab. 1. It slightly differs between the five meters and between the three profile types (supply temperature, return temperature, heat load), but is mostly similar regarding the applied filters. The detected outliers amount to 0.6 % up to 1.8 % of the total dataset

Tab. 1: Comparison of the number of outliers detected by the two filtering methods in processed return temperature and heat load profiles

	Number of detected outliers				
Meter ID	**1**	**2**	**3**	**4**	**5**
Supply Temperature					
Z-filter	474	428	445	484	423
Return Temperature					
Z-filter	237	320	242	268	326
Grubbs-filter	226	298	257	149	308
Heat Load					
Z-filter	330	375	363	304	307
Grubbs-filter	385	353	332	280	304

Fig. 3 depicts an excerpt from the return temperature profile of meter 1 for May 2020 in hourly resolution. The graph shows that the Z-filter (black markers) only detects the outliers that deviate more than 2.5 standard deviations (green dashed line) from the mean value (grey dashed line) of the floating period, while the Grubbs-filter (red makers) also detects outliers that deviate from the course of the profile. This can also be observed in the excerpt of the heat load profile during the same period (Fig. 4).

On the one hand outliers are sometimes detected in both profiles (heat load and return temperature) at the same time, on the other hand some occurrences do not overlap. Some outliers are detected by both filters, others by only one of them. In order to confirm if the detected outliers are actually faulty meter data, a validation dataset is needed. As this was not available in this study, the accuracy of the filters cannot be evaluated quantitatively. Nevertheless, the most obvious outliers that can be detected visually, are mostly detected by the filters.

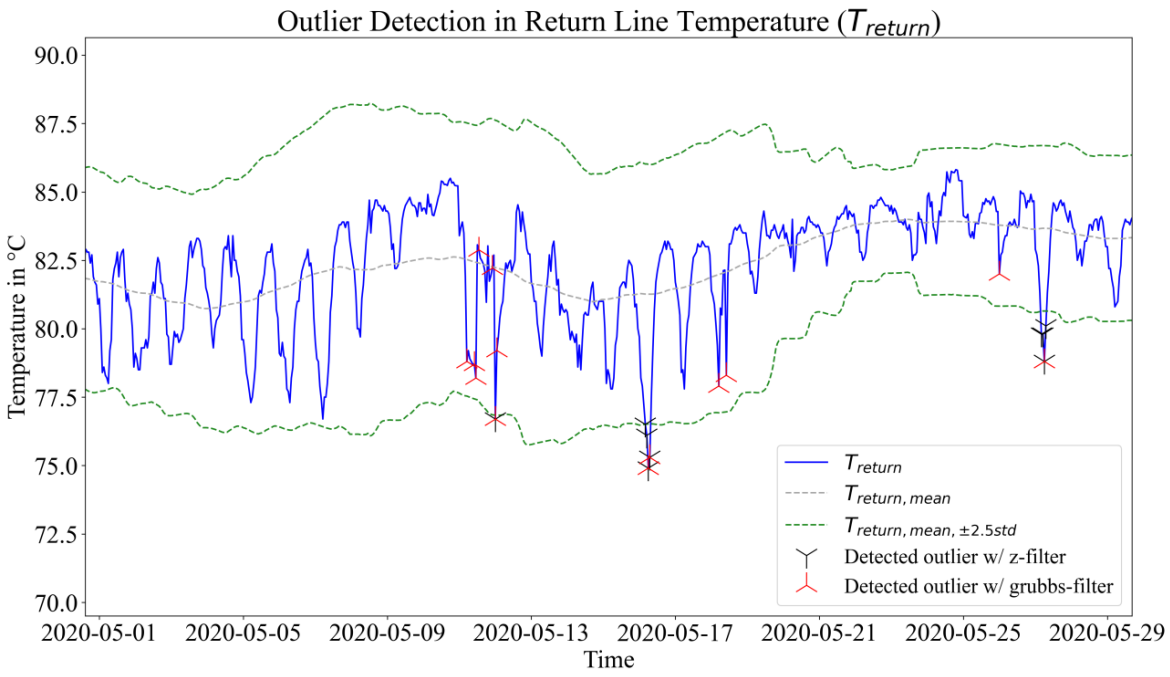

Fig. 3: Excerpt of a return temperature profile (meter 1, May 2020, hourly resolution) with Grubbs- and Z-filter applied

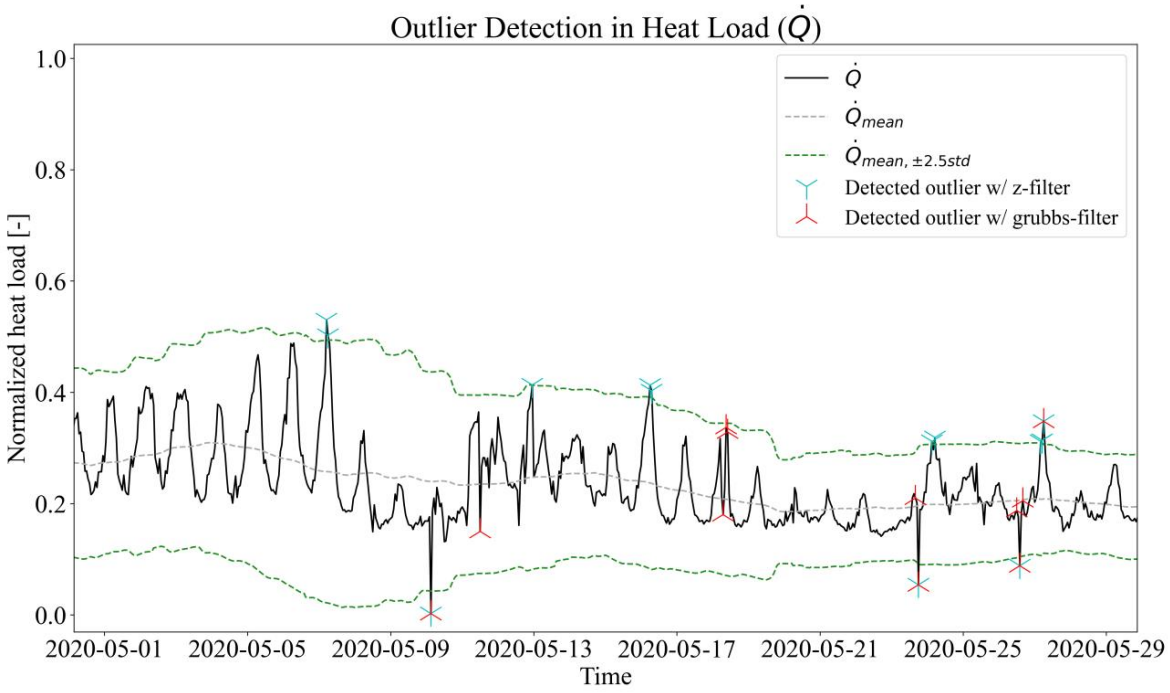

Fig. 4: Excerpt of a heat load profile (meter 1, May 2020, hourly resolution) with Grubbs- and Z-filter applied

3.2 Performance of the machine learning algorithm

The two following figures (Fig. 5 and Fig. 6) show the completed three-step processing of the profiles. The missing values in the original profiles (red and grey dashed lines) are visible as zero values and periods with constant supply temperature. In the first step, the missing values are replaced by predictions using the machine learning algorithm (cyan dashed lines). In the second step, the Grubbs-filter is applied (red markers). The figures show that detected outliers are frequently adjacent to missing periods. In the final profiles (solid lines), the detected outliers are removed from the training data of the machine learning algorithm and treated as missing values. Therefore, the course of the profiles created with the regression models in the third step slightly differs from those that in the first step.

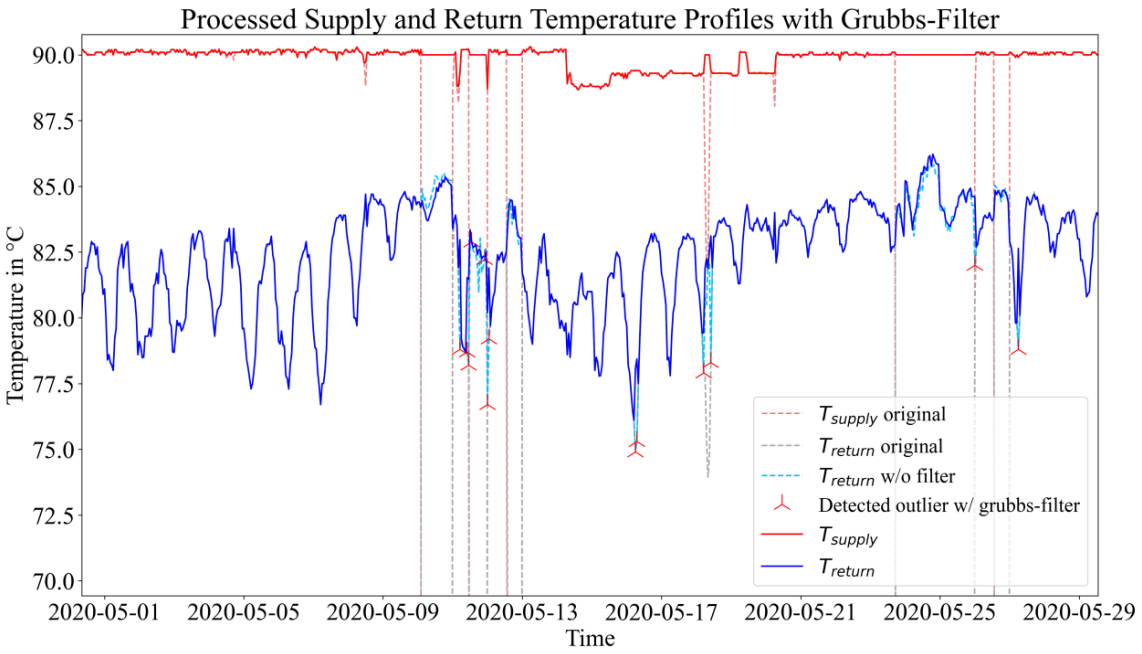

Fig. 5: Visualization of the three-step processing of supply and return temperature profiles – original (grey), processed profile with detected outliers (cyan), final processed profile (blue) (May 2020, hourly resolution)

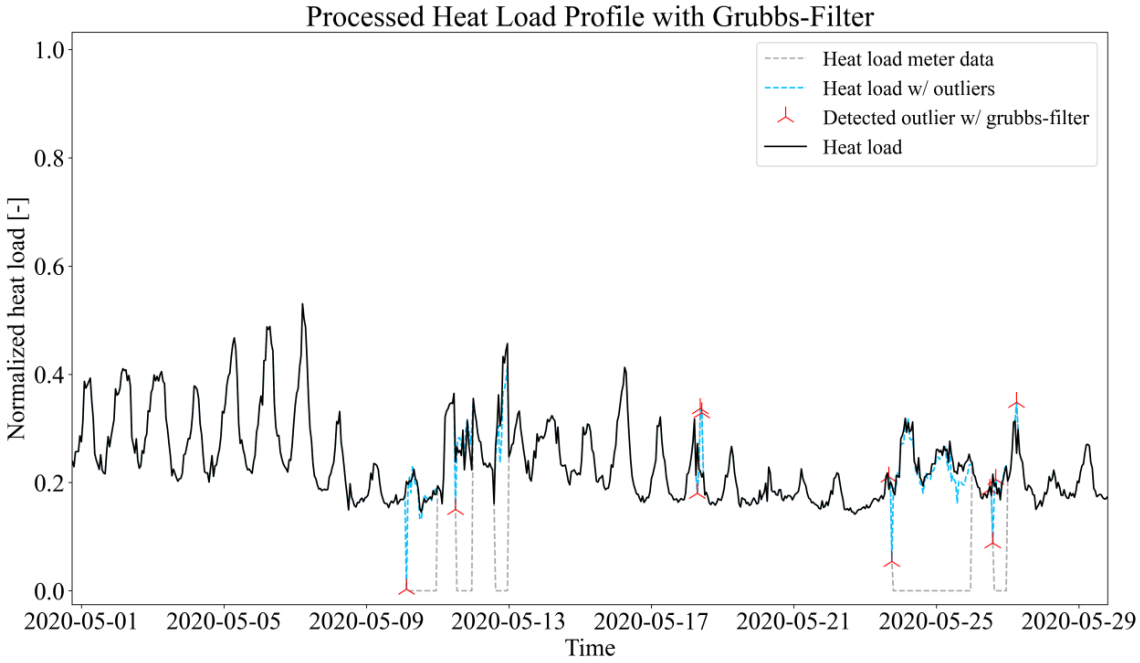

Fig. 6: Visualization of the three-step processing of supply and return temperature profiles – original (grey), processed profile with detected outliers (cyan), final processed profile(black) (May 2020, hourly resolution)

The accuracy of the presented machine learning algorithm is evaluated using the indicators R^2, RMSE and MAPE presented in Tab. 2 as the mean of the individual scores of the five meters for the base case (first step) and for either filter applied.

In general, the indicators confirm the visual impression from Fig. 5 and Fig. 6 that the models are able to predict realistic profiles for the missing values in the timeseries. The influence of the filters on the accuracy of the models is below 5 %. However, in all cases they improve the models by increasing R^2 but decreasing RMSE and MAPE. Since the number of detected outliers is relatively low (0.6 – 1.8%), the training data is only slightly altered by the filters compared to the base case. Therefore, impact of the filters on the overall performance of the algorithm is quite low. Additionally, there is no significant trend for favoring Grubbs- or Z-filter on the basis of the calculated indicators.

Tab. 2: Average performance of the machine learning algorithm as mean of the five meters and influence of the two filters

	Evaluation criterion								
	R^2			**RMSE in K resp. kW**			**MAPE in %**		
	w/o filter	*Grubbs*	*Z*	**w/o filter**	*Grubbs*	*Z*	**w/o filter**	*Grubbs*	*Z*
Return temperature	**0.917**	*0.921*	*0.922*	**0.94**	*0.91*	*0.92*	**5.41**	*5.23*	*5.23*
Heat load	**0.925**	*0.929*	*0.930*	**140.0**	*133.9*	*135.3*	**13.18**	*12.58*	*12.71*

In addition to the performance indicators, the processed profiles are evaluated in terms of annual mean supply and return temperature and total heat amount. Tab. 3 shows that the supply temperature in processed profiles barely deviates from the original data, while the return temperature is more than 1.8 % higher. Since the temperatures are volume flow weighted, no strong deviation is expected. However, the heat amount is more than 17 % higher than in the original data, where missing values occur in 29 % of the time. This is consistent with the visual impression that the gaps occur more frequently in summer when the heat load is below average.

Tab. 3: Average deviation of annual mean supply and return temperature and total heat amount of the five meters in processed profiles compared to original profiles

	Deviation		
Value	**w/o filter**	**Grubbs**	**Z**
Supply temperature	+0.27%	+0.27%	+0.27%
Return temperature	+1.84%	+1.87%	+1.81%
Heat amount	+17.34%	+17.66%	+17.45%

3.3 Importance of input parameters

Finally, the input parameters to the black box models, which are also called features in the machine learning domain, are analysed. The relative importance of the six features is retrieved from the machine learning algorithm in the scikit-learn library and is depicted in Fig. 7 for the several processed profiles of return temperature and heat load that are discussed in the previous section. It clearly shows that ambient temperature is the most influential input parameter with more than 35% and is followed by the hour of the day and the day of week. The supply temperature is only used as an input parameter to the return temperature models and has a medium relevance with about 15 %. There is no significant change in the feature importance when comparing the models with and without a filter.

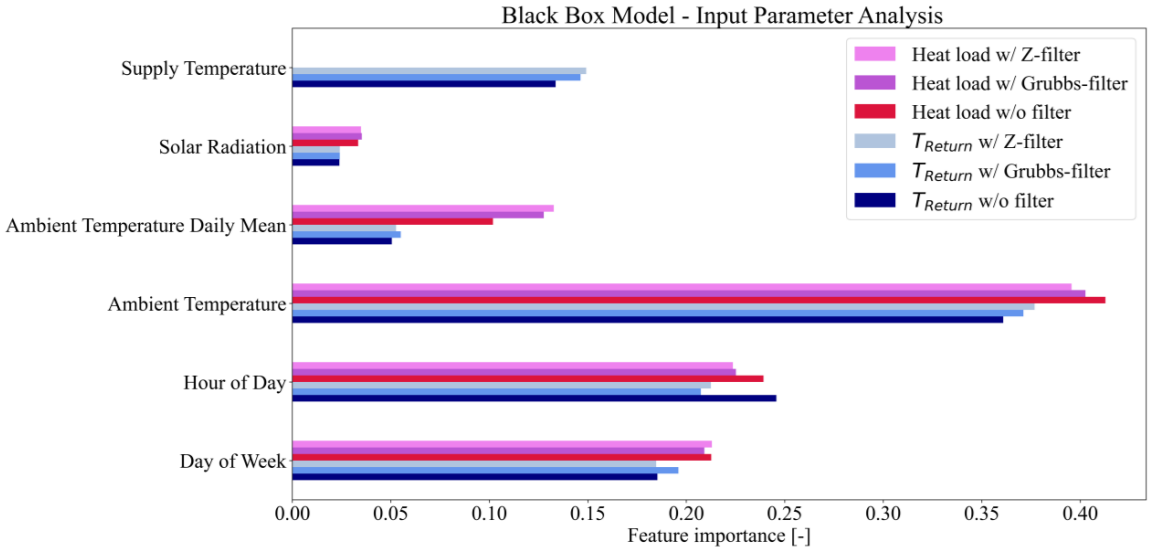

Fig. 7: Input parameter (feature) importance for return temperature and heat load prediction models with and without filters for outlier detection

4. Conclusion

The methods applied in this study enable to process faulty heat meter data and improve its consistency. Missing datapoints and detected outliers are replaced by using machine learning regression models. The performance of the models is evaluated with common indicators like R^2, RMSE and MAPE. A coefficient of determination R^2 of about 0.92 is achieved for both heat load and return temperature profiles. It should be noted that the indicators are calculated as mean values from five different meters to ease readability. However, the quality and availability of the training data of the profiles vary and strongly influence the accuracy of the individual models. It confirms the thesis that the better the data basis, the better the results of the model predictions.

The applied filters detect the most visually identifiable outliers, although some datapoints are detected which are difficult to distinguish from actual peaks. In a further step, it is therefore necessary to test the filters on data, where a corresponding validation dataset with known outliers exists. Nevertheless, the performance evaluation shows that the filters slightly improve the accuracy of the machine learning models.

The evaluation of the annual mean values points out that the processed profiles deviate from the original data, which is especially true for the heat amount with more than 17% deviation. This will likely cause errors when redesigning or continuously optimizing this energy system with the original profiles. After all, redesigning and optimizing are vital aspects for transforming the heat sector from conventional fossil-fuelled systems to smart energy-efficient systems based on renewable energies. Therefore, this work is aiming to contribute to a more efficient transformation process.

5. Acknowledgements

The project partners greatly acknowledge the financial support of the project by the Federal Ministry for Economic Affairs and Climate Action. (Funding Code: 03EN3023)

6. References

Adikaram, K.K.L.B., Hussein, M.A., Effenberger, M., Becker, T., 2015. Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation. Journal of Applied Mathematics 2015, 1–9.

Bünning, F., Heer, P., Smith, R.S., Lygeros, J., 2020. Improved day ahead heating demand forecasting by online correction methods. Energy and Buildings 211, 109821.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 11 (34), 785–794.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2012. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12, 2825–2830.

Saloux, E., Candanedo, J.A., 2018. Forecasting District Heating Demand using Machine Learning Algorithms. Energy Procedia. 149, 59–68.