Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses

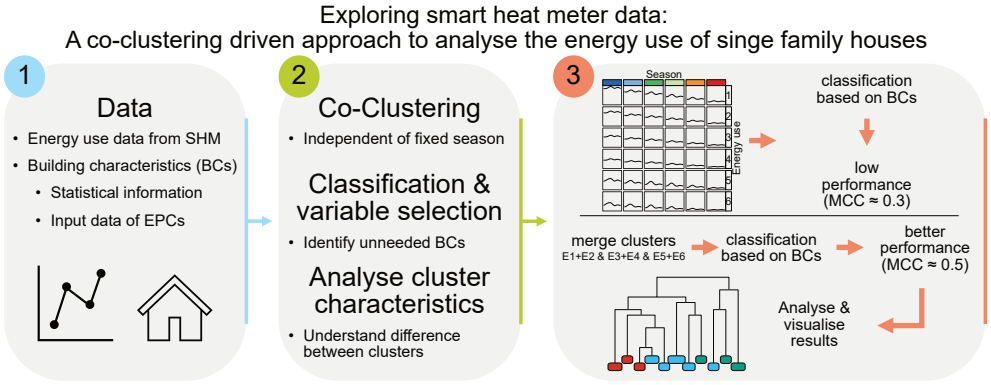
Schaffer, Markus; Vera-Valdés, J. Eduardo; Marszal-Pomianowska, Anna

Publication date:
2023

Graphical Abstract

Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses

Markus Schaffer, J. Eduardo Vera-Valdés, Anna Marszal-Pomianowska



Highlights

Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses

Markus Schaffer, J. Eduardo Vera-Valdés, Anna Marszal-Pomianowska

- Co-clustering of smart heat meters to establish season-independent energy clusters
- Analysis of energy use clusters based on 26 building characteristics
- Classification and variable selection to identify the minimum information needed
- Statistical data leads to the same insight as detailed building data
- Prediction of energy use clusters with building characteristics has a low accuracy

Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses

Markus Schaffer[a,*], J. Eduardo Vera-Valdés[b], Anna Marszal-Pomianowska[a]

[a]*Department of the Built Environment, Aalborg University, Aalborg, 9220, Denmark*
[b]*Department of Mathematical Sciences, Aalborg University, Aalborg, 9220, Denmark*

Abstract

The ongoing digitalisation of the district heating (DH) sector opens new doors for data-driven methods. Remotely readable meters (smart meters) create data at an unprecedented extent and temporal resolution, which allows gaining inside into the energy use of buildings. This insight can support the needed renovation wave and the transformation of the current DH networks to low-temperature 4th generation DH networks. This work contributes by proposing a novel workflow to establish energy use clusters without relying on fixed season definition, addressing the challenge that climate change makes static seasonal definitions difficult to establish. Further, these clusters are analysed regarding their relationship with respect to 26 building characteristics (BCs) to understand why a building is within a specific cluster using classification and variable selection methods. The results, based on two years of data of 4798 single-family houses, show that the used co-clustering approach establishes well-separated energy use clusters. While correlated to the exterior temperature, the found season variation does not follow commonly used fixed season definitions and further varies across years, showing that fixed season definitions do not correctly capture the individual seasonal variation of energy use in single-family houses. The results of the variable selection and classification approaches show that even highly detailed BCs are not sufficient to explain why a building is in its respective cluster (Matthew's correlation coefficient (MCC) ≈0.3). By artificially simplifying the found energy use clusters based on similarities in their energy use profile and magnitude, the performance of the classification could be significantly increased (MCC ≈0.5). For both the not simplified and simplified energy use clusters, simple BCs, which are inexpensive to collect and in most cases, already available, lead to a similar understanding as detailed BCs.

Keywords: smart meter data, district heating, co-clustering, classification, variable selection

1. Introduction

In light of the recent geopolitical changes and the accompanying need to reduce the dependency of the European Union (EU) on natural gas, district heating (DH) has come to the fore in some EU countries [1, 2, 3]. DH can not only play an essential role in reducing the dependency on natural gas but also in the needed future reduction

*Corresponding author
Email address: msch@build.aau.dk (Markus Schaffer)

of CO_2 emissions if renewable energy sources and residual energy sources (waste and biomass) are implemented [4]. However, to facilitate the required share of 100 % renewable energy, existing DH networks must undergo severe changes to become low-temperature 4th generation DH networks [4]. Yet, to enable such a transformation, the building stock must also transform as such networks must interact with low-energy buildings [4]. However, nowadays, nearly 75 % of the EU´s building stock is energy inefficient [5]. In most EU countries, half of the residential buildings were built before the first thermal regulations (in 1970), while the renovation rate remains at low 1 % to 2 % per year [6]. Thus, highlighting the challenges the building sector must face to enable a fully decarbonised building stock by 2050 [5]. To facilitate this transition, in-depth knowledge about the building stock is required.

In ten EU countries, more than 20 % of the residential sector's heating demand is covered by DH, with five countries having a share of more than 50 % [7]. At the same time, since the end of 2020, newly installed heat meters must be smart heat meters (SHMs) (remotely readable meters), and from 2027 also previously installed heat meters must be remotely readable [8]. This, already nowadays, available data from SHM opens up the door for new data-driven methods to transform both DH networks and the building stock.

This potential was confirmed by recent research, demonstrating the broad spectrum of possible use cases of building/apartment level SHM data. On a large scale by calibrating and validating urban building energy models [9, 10, 11, 12, 13], by optimising the DH network temperature control [14], and by forecasting heat loads in the DH network [15]. On a smaller scale, it was shown that SHM data can be used to derive building characteristics [16, 17], to apply clustering to identify typical consumption patterns [18, 19, 20] and relate these to buildings and occupants' characteristics [21, 22, 17], to analyse the peak consumption and evaluate the potential of peak load shifting [23, 24], and to identify abnormal operation [25, 26, 27, 28]. Overall, it is expected that inspiration about the application possibilities can also be drawn from the mature field of smart electricity meters (SEMs), such as summarised in Wang et al. [29]. However, there are differences between SEM and SHM data. The most distinct ones are that SEMs have a higher reporting frequency of 15 min or less for most EU countries [30], while SHMs commonly report only in 1 h intervals. Additionally, SHM data are commonly rounded down to integer kWh values [11] meaning that e.g., any value between 1.0 kWh and $1.\overline{9}$ kWh is transmitted as 1.0 kWh. Hence the applicability of SEM research to SHM is not necessarily given.

While the principal usability of SHM data-driven approaches was shown in recent research, limitations remain. For clustering, with the notable exception of do Carmo and Christensen [21], approaches rely until now either on fixed dates for season definitions or do not take season into account [e.g. 18, 26, 19, 23]. Especially as traditional seasonal patterns shift and dissolve in the face of climate change [31, 32, 33]. This makes it harder and less reliable to identify predefined seasons. Furthermore, in depth analyses of, at large-scale, available building characteristics (BCs) in relation to obtained clusters to gain more insight into, e.g., the cause for different energy use patterns, have been limited in terms of number of studied BCs [21, 22]. Thus, it remains unknown if more or other BCs, than the studied ones, would give more insight into, e.g., the cause for different energy use patterns or which BCs are overall important.

For these reasons, this work establishes a novel workflow to overcome the mentioned limitations in the area of clustering SHM data and understanding these clusters. Therefore, first, easy to interpret representative daily heat energy use curves without fixed season definitions are derived. Thereafter, the derived energy use clusters are analysed in relation to both high-level BCs as well as BCs describing a building in a level of detail. Thereby the aim is to first identify the most important/useful BCs and secondly to understand the difference in buildings between clusters based on the identified BCs. Thereby focus is set on communicating the differences between the energy use clusters in a way that allows for layman level communication of the findings to decision-makers and non-experts. Furthermore it is analysed if not yet to the DH network-connected buildings can be classified into the established clusters based on the selected most important BCs. The suitability of the whole process is demonstrated on a large-scale dataset of two years of data from 4798 SHM installed in single-family houses in Aalborg, Denmark. This analysis aims to gain more knowledge of the DH network and connected buildings. The derived significant BCs can guide stakeholders such as DH utility companies in collecting such data if such information is yet unavailable.

The paper is organised as follows. Section 2 highlights the contribution of this work, before Section 3 describes each step of the proposed method in detail and relates it to state-of-the-art research. Section 4 outlines the used SHM and BCs data. In Section 5, the results of the case study are presented before the results are discussed, and conclusions are drawn in Section 6

2. Contribution

- Clustering of smart heat meter data for the first time, taking into account seasonality without relying on fixed season definitions.
- Analysis of heat energy use clusters in relation to building characteristics at an unprecedented scale and level of detail.
- Identifying which building characteristics are useful to explain the difference between the heat energy use clusters.
- Evaluating whether classification can be used to predict building energy use clusters based on building characteristics.

3. Method

For better clarity in following each step, its aim and the respective state-of-the-art research are outlined separately. First, for the energy use clustering and after that for the variable selection and classification before the analysis of cluster characteristics. An overview of the proposed method is given in Figure 1.

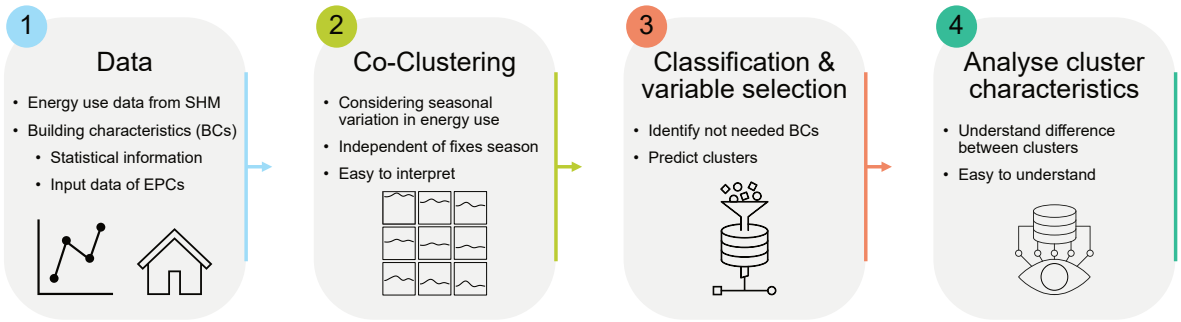


Figure 1: Overview of the proposed method.

3.1. Energy use clustering

3.1.1. State of the art

The first step intents to derive representative energy use curves from the SHM data. The task of deriving representative energy use curves can generally be seen as time-series clustering, of which exhaustive overviews are given in Warren Liao [34] and Aghabozorgi et al. [35], and more specifically for the DH sector, partly in Mbiydzenyuy et al. [36]. Most works related to clustering SHM data to establish daily load profiles rely either on fixed season definitions by date or do not account for seasonal variation at all [18, 26, 19, 23]. One work that studied the seasonal variation is Gianniou et al. [22], who used k-means clustering based on the *KSC-distance* to derive energy use density and energy use patterns for about 8300 single-family buildings in Denmark. They analysed how the energy use density changes over the year and found that, while the various consumption densities showed overall the expected seasonal variation, deviations were also visible. Notable is the work of do Carmo and Christensen [21], who,in order to be independent of fixed seasons, proposed to first cluster the daily profiles of each building into three groups (high, medium, low), which can be seen as a representation of seasons, before clustering each group across all buildings. Applying this approach with k-means clustering on the heating and domestic hot water energy use of 139 Danish dwellings equipped with heat pumps, they demonstrated that the seasonal groups do not precisely follow the anticipated seasons. Thus, both studies highlight the need to use methods which define seasons independent of fixed dates. This also agrees with the before mentioned consideration that due to climate change fixed season definitions are becoming increasingly difficult to establish [31, 32, 33]. In the field of SEM data, Bouveyron et al. [37] has identified this problem of clustering the customer's electricity use while considering the season variation as a co-clustering problem of individuals (customers) and a feature (time). While mainly focused on the mathematical development of their method (a functional latent block model), they could demonstrate the applicability of their method by clustering the residuals of a regression of the energy use against the outdoor temperature, from about two years of half-hourly SEM energy data from 1481 households in France. Following the same principle idea, Divina et al. [38] developed a different co-clustering approach (Sequential Multi-Objective Bi-clustering) to identify groups of buildings that behave similarly during a time period, which they applied to 15 min resolution electricity data of five university buildings in Spain.

3.1.2. Used Co-clustering method

Based on the recent research results, this work aims to establish representative daily energy use profiles without relying on a fixed season definition. While the approach of do Carmo and Christensen [21] is appealing because it allows using any clustering algorithm, it has the drawback that the information about the season length for each building is lost, i.e., if buildings have a similar energy use profile for each energy use level (low, medium, high) but the length of each energy use level differs significantly this information is lost. Consequently, the co-clustering approach developed by Bouveyron et al. [37] and implemented in the R [39] package `FunLBM` [40] is used for this work. This approach, as mentioned, allows to cluster customers (the individual buildings) while taking the days of observation, the seasonality, as a feature into account. The approach assumes that the data can be summarised in a few exhaustive co-cluster. Hence, it is assumed that all data belongs to any of the found cluster. The obtained clusters follow a checkerboard like structure, i.e. the season pattern is identical for all energy use clusters. From a mathematical perspective, the algorithm is an extension of the latent block model [41] to functional data using a model-based approach, which assumes that functional principal components of the curves are block specific. From this also, the name FunLBM (functional latent block model) is derived, which is used from here on. For a more detailed explanation of the algorithm, the interested reader is referred to Bouveyron et al. [37]. To select the most suitable model, so the optimal number of both customer and time cluster, FunLBM uses the *integrated information likelihood criterion* (ICL) (also referred to as *integrated completed likelihood*, or *integrated classification likelihood*), whereby the highest ICL value indicates the most suitable model [37]. Thus, a grid search must be performed to find the optimal number of cluster. *FunLBM* transforms the discrete energy use data into functional data using basis expansions based on Fourier or Spline basis. Given the expected periodic nature of the data as in Bouveyron et al. [37] and as recommended by Ramsay and Silverman [42], Fourier basis functions are seen as a more appropriate choice for SHM data. The number of basis functions, which the user must supply to *FunLBM*, influences the degree of smoothing and thus the resulting cluster. The number of basis functions is a problem of bias/variance trade-off (excluding random or ignorable variation in the data while keeping important one), and one common approach to solve this is to use *generalised cross-validation (GCV)* [42] as a criterion. This approach was also chosen for this work. The overall outcome of this step are representative mean energy use curves and co-cluster.

3.2. Variable selection and Classification

The main two aim of this step is to understand why buildings fall into their respective energy use cluster (Section 3.1) and to classify yet not to the DH network connected buildings based on their BCs into energy clusters.

3.2.1. State of the art

A similar idea was recently shown by do Carmo and Christensen [21] (described before in 3.1.1), who used logistic regression to analyse the influence of building and socioeconomic parameters for heat energy use cluster. They analysed the influence of eleven BCs and four household characteristics and concluded that BCs such as the

building area and age are significant as well as the space heating distribution system for medium and low energy use seasons. However, they also pointed out that the limited number of household and building characteristics available to them, limited the potential of the analysis. Further, their small homogeneous group of buildings (139 dwellings all with a ground source and air-water heat-pumps) makes it unknown if their results can be generalised. A similar analysis on a large sample of about 8300 single family houses but with less characteristics (building area and age, and number of adults, teenager and children) was conducted by Gianniou et al. [22] for clusters of daily energy use curves based on SHM data. Based on logit regression models they built for each of the five daily energy use clusters (thus analysing if a building belongs to an cluster or not) they concluded that all of the building area and age as well as the number of teenagers are significant. Further, they concluded that more building characteristics should be investigated. Focusing more on social economic characteristics, such as income, job type etc. Hansen et al. [43] analysed their relation to energy use peaks of about 800 households based on SHM data. They computed energy usage profiles for average working days for different characteristics, and used linear regression, once controlled and once uncontrolled for selected BCs, to analyse the significance of the different social economic parameters. They concluded that when controlled for the BCs, only the household income remains significant. Thus, they concluded that BCs and the household income seem to be the most important factors for the energy usage peaks. Further studies with a similar intend were conducted in the field of electricity data [44, 45, 46, 47], but given the difference in driving forces for electricity consumption they are expected to be less applicable to this work.

3.2.2. Used classification and variable selection methods

Based on the current research, two key research gaps can be identified. The first is that, until now, only limited BCs have been used which limits the general validity of the found significant BCs as a parameter can become insignificant if other (better) information is available to the model. However, at the same time it is expected that if many BCs are available, at least some are redundant. Thus, variable selection to minimise noise and obtain the simplest model possible [48] is seen as a necessary step. In addition, a reduction in the number of BCs required for the model can also be seen as a reduction in costs, as the collection of additional and more detailed BCs is associated with considerable costs if it is to be done at the city or country level. Consequently, a simpler model can also be seen as easier and cheaper to implement, and thus more applicable for ”real world” applications. The second gap is that the potential of multiclass classification has not been explored, i.e., can a building based on BCs be classified into one of the found energy use clusters.

The first approach used to address both aims is a multinomial logistic regression fitted using group least absolute shrinkage and selection operator (group lasso) [49] (MLRGL). This can be seen as an extension of the current research, which used logistic or logit regression. Lasso in general is an established regularisation and feature selection approach which was also applied in recent research in the context of sensitivity analysis in the building sector to identify significant parameters [50, 51, 52]. The benefit of group lasso [53] is that it allows grouping variables together, which is beneficial for e.g. categorical variables, which have to be encoded with dummy variables, as it prevents that one level

of a categorical variable while another level is not included. As categorical BCs are not expected to have many levels, sparse grouped lasso [54], which allows to group predictors but also that predictors within a group are not included, was not considered. As MLRGL requires to select the penalty coefficient lambda (λ), nested cross-validation with five outer and ten inner folds is used for model selection and assessment. For the assessment, the Matthews correlation coefficient (MCC) is used, which was shown to be superior over e.g. the accuracy, particularly for unbalanced classes [55, 56]. The MCC can be interpreted as the Pearson correlation coefficient, ranging from -1 to 1, with 1 being perfect agreement. BCs are scaled as recommended by Gelman [57], by subtracting the mean and dividing by two standard deviations for continuous BCs and centering binary BCs. For this work the MLRGL as implemented in the R package `msgl` package [58] was used.

The second used approach is variable selection based on random forests (RF) and was proposed by Genuer et al. [59] and is implemented in the R package called VSURF (Variable Selection Using Random Forests) [60, 61] (this name will be also used in the remainder of this paper to refer to this method). VSURF is a two step procedure, whereby in the first step (*threshold*) variables are ranked based on their permutation-based importance and unimportant variables are excluded. The second step consists of two sub-steps, in the first sub-step called *interpretation* a nested collection of RF is constructed, starting with one that includes only the most important variable (based on the threshold step) to one that includes all variables retained from the first step. and the most accurate is kept (based on the out-of-bag (OOB) error) For the second sub-step called *prediction* the by importance sorted variables (based on the previous step) are sequentially introduced and the variable is only kept if it decreases the error significantly. For a more detailed explanation the interested reader is referred to the aforementioned references. VSURF was shown to identify a smaller subset of important variables compared to lasso [62] and to be the generally best performing RF variable selection technique for classification [63]. To reduce the risk of bias, VSURF is used within outer five cross validations. It is to be noted that the MCC could not be easily implemented within VSURF and consequently the OOB error is used as a criterion within the variable selection procedure, while the MCC is used to evaluate the performance on the test data. As VSURF does not perform any hyperparameter optimisation based on the chosen BCs, a random forest model is constructed and its hyperparameters are tuned using the automated tuning strategy tuneRanger [64] with MCC as the optimisation criterion and within an an outer five fold cross validation.

3.3. Analysis of cluster characteristics

From the above derived models MLRGL and VSURF + optimised RF, it is not necessarily straightforward to understand why buildings fall into their respective clusters. One could for example analyse the coefficients of MLRGL. However, this requires statistical knowledge and experience, which might not be available at, e.g., DH utility companies. Thus a more graphical visualisation is preferred in this work. For the results of MLRGL, a nomogram as proposed by Zhang and Kattan [65] could be used.

For VSURF + optimised RF, respectively the RF only (as VSURF is only used for variable selection), one can visualise the individual decision trees of the RF, which are well-known to be suitable for visual data analysis [66, 67,

68]. As the RF in its used implementation can not be directly visualised, it was decided to build a separate decision tree for visualisation. As the purpose this step is not generalised prediction no splitting into training and test data is performed and over fitting to the data at hand is deliberately taken into account. Decision trees, as implemented in the R package `rpart` [69] in combination with the dedicated visualisation package `rpart.plot` [70] were used.

4. Data description

The data used in this work is a subset of the extensive dataset of SHM data, and BCs described in Schaffer et al. [71]. This dataset consists of processed hourly data from about 35 000 SHMs installed mainly in residential buildings in Aalborg Municipality, Denmark, with varying lengths and, where available, accompanying BCs. From this dataset, only SHM data of single-family houses for which data are available for 2020 and 2021 were selected. Further, only buildings/SHMs where all accompanying BCs are available were considered. Based on this, SHM data of 4798 single-family houses were selected. It is to be noted that the BCs originating from the Energy Performance Certificates (EPCs) reports were recomputed using the in described procedure with the change that the validity period was set to 2020 and 2021 only, which allowed retrieving more valid data. In the following, a short overview of the data is given. A more extensive description is given in Schaffer et al. [71].

For this work, only the energy use data from SHM is considered. The energy use data is the hourly aggregated energy use for space heating and domestic hot water. As SHM data used in previous research, the energy use data is transmitted as cumulative kilowatt-hour values, which are rounded down to the next integer [72, 11]. To mitigate this problem not the original data but the energy use data processed by the the by Schaffer et al. [73] developed approach called SPMS is used, which is also available in the dataset. SPMS uses a moving average smoothing combined with a ruleset and scaling approach. Thereby, SPMS obeys the cumulative trend of the data on a daily basis, i.e., every day accumulates to the same amount as the unprocessed data. The data obtained from the dataset was normalised by the buildings total area to accommodate the well known influence of the building size on the heat energy use while still allowing to incorporate the energy use intensity. In the remainder of the paper, the term *energy use* always refers to the with SPMS processed and by the area normalised energy use data.

For each of the selected buildings, BCs from two sources in Denmark are available. The first source is the Danish Building and Dwelling Register (BBR) [74], which is a publicly accessible database which the Danish Customs and Tax Administration operates. It contains statistical/high-level information about every Building in Denmark to a unit level, i.e., an apartment for a multifamily house or the whole house for a single-family house, and the building owner must provide some of the information. The information in the dataset originating from the BBR includes, e.g., the unit size, the number of rooms and the unit use. The second source from which data is available is the input data for Energy Performance Certificates (EPCs). This data is not publicly available and contains detailed information about a building down to a component level, e.g., u-value, total solar transmittance, orientation, and size of a window. In Schaffer et al. [71], the data was summarised from this component level, so every building has the same features. For

the remainder of the paper the BCs originating from the BBR and the EPCs are always analysed once separately and once combined. Thus three different situations are considered:

- BBR only
- EPC only
- BBR + EPC

4.1. Building characteristic data treatment

In total, 86 BCs are available in the dataset developed by Schaffer et al. [71]. As for the selected 4798 single-family houses, the BBR data unit level is identical to the building level. Thus, the first step of the processing was to select only non-redundant variables, whereby unit-level variables were preferred over building-level variables. Consequently, seven BBR-based BCs were dropped. In the second processing step, only BCs were kept if they clearly varied within the selected buildings. After this step only one building had one missing BC (rent status), which was imputed with the most frequent value. The last step had two aims, on the one hand, to reduce collinearity between parameters assessed using the Pearson correlation coefficient and, on the other hand, to simplify BCs and incorporate possible known interaction between parameters. Additionally, where appropriate BCs were, as the used energy use, normalised by the total building area. An overview of all resulting 26 BCs (ten originating from the BBR data, 16 from the EPCs) is given in Table 1. The correlation matrix of these BCs is shown in Figure A.16. From this, it can be seen that a correlation exists for some BCs, particularly in relation to the representative year. Therefore as next step an analysis of possible multicollinearity was conducted.

4.1.1. Analysis of multicollinearity of building characteristic

As mentioned above in Section 4.1, some of the used BCs are partially correlated (Figure A.16). To further analyse this and identify possible multicollinearity, the generalised variable inflation factor (GVIF) [75], an extension of the variable inflation factor (VIF) for categorical variables was used. Further, as suggested by Fox and Monette [75], to make the GVIF comparable across dimensions, the degrees of freedom of the coefficients were taken into account: $GVIF^{(1/(2\times DF))}$. Additionally, to make the result comparable with the VIF, the result was squared: $\left(GVIF^{(1/(2\times DF))}\right)^2$. Figure 2 shows the result of this analysis considering all BCs. Given that no value exceeds 5, a commonly used rule of thumb, each BC is assessed to be not multicollinear with the remaining BCs.

5. Results

5.1. Co-Clustering

The first step in the co-clustering process is to determine the optimal number of basis functions as a variance-bias trade-off based on GCV. The results (Figure 3) show that seven basis functions give the lowest GCV with an apparent

Table 1: Used BCs based on the database described in Schaffer et al. [71]. If the levels of categorical BCs were changed the new levels are highlighted in italic while the original ones as stated in Schaffer et al. [71] are written in parentheses.

		BC name	BC description
Combined	BBR	developed_area_ratio	Developed area divided by the total building area
		ext_wall_mat_code	Exterior wall cladding material simplified to five levels: *brick*(0), *concrete*(2,3,6), *wood*(4,5), *others*
		nr_bathroom	Number of bathrooms
		nr_floor	Number of floors
		nr_room	Number of rooms
		nr_toilet	Number of toilets
		renovation_code	Binary variable indicating if the building was renovated (*TRUE*) or not (*FALSE*)
		rent_status_code	Indicating if the building is rented (*rented*) or used by the owner (*self_use*) or not used (*not_used*)
		representative_year	If the building was renovated the renovation year, otherwise the construction year
		roof_mat_code	Roof material cladding summarised to seven levels: *not_stated*(0), *built_up*(1), *roofing_felt*(2), *fiber_cement*(3,10), *cement_tile*(4), *tile*(5), *metal*(6), *others*
	EPC	dhw_average_consumption	Total Domestic hot water demand - building area normalised
		dhw_pipes	Total heat losses through DHW pipes - building area normalised
		dhw_tank_heat_loss	Total heat losses from domestic hot water tanks - area normalised
		has_heat_pump_code	Binary variable indicating if a building has a heat pump (*TRUE*) or not ((*FALSE*))
		heat_capacity	Simplified heat capacity of the building per unit gross area
		heating_pipes	Total heat losses through heating pipes - building area normalised
		heating_temp_diff	Calculate temperature difference between supply and return temperature of the heat distribution system
		skylight_solar	Total pseudo solar factor of skylights - building area normalised
		thermal_bridge_total	Total heat losses through thermal bridges - building area normalised
		total_transmission	Total heat losses through opaque and transparent building envelope - building area normalised
		vent_mech_winter	Total equivalent mechanical ventilation in winter - building area normalised
		vent_nat_winter	Total equivalent natural ventilation in winter - building area normalised
		window_solar_east	Total pseudo solar factor of windows facing east - building area normalised
		window_solar_north	Total pseudo solar factor of windows facing north - building area normalised
		window_solar_south	Total pseudo solar factor of windows facing south - building area normalised
		window_solar_west	Total pseudo solar factor of windows facing west - building area normalised

decrease in GCV compared to 5 basis functions. This is therefore considered to be the optimal choice for transforming the discrete energy use data into functional data.

As the next step, the optimal number of clusters had to be determined for energy use and time. Therefore, a grid search was performed over 2 to 9 time clusters and 2 to 12 energy use clusters. One cluster for time or energy use can not be used as the algorithm requires at least two clusters. For each combination, convergence, defined as the change

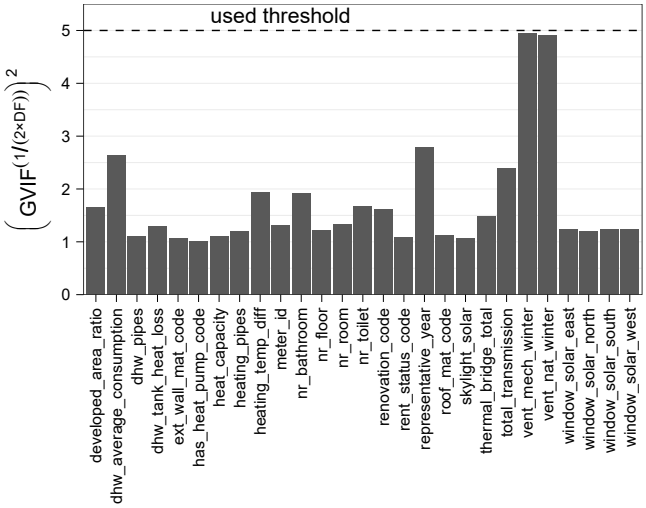


Figure 2: For the the degrees of freedom of the coefficients adjusted generalised variable inflation factor of all BCs.

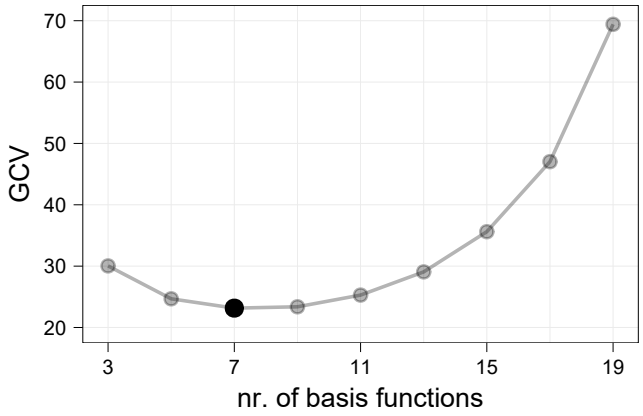


Figure 3: Generalised cross-validation for different number of Fourier basis functions

of the loglikelihood between the current and the current minus ten iterations of smaller than 1×10^{-5}, was ensured. As shown in Figure 4, the ICL has its maximum at six energy use and six time clusters. Thus, this was chosen as the optimal result for all further analyses. Further, it can be seen that this is, at the same time, the result with the most partitions where a solution could be found without a cluster being empty, and consequently, FunLBM failing.

As the second step of the clustering results analyses, the six time clusters were analysed to understand how they relate to known seasonal variations of the exterior conditions. As the naming of the clusters is arbitrary, the cluster names were chosen to represent the season pattern to ease the understanding. In Figure 5 a), the distribution of the time clusters is shown. Considering only the time clusters, a principle pattern is visible. T1 and T2 seem to be clusters of the winter season, T3 and T4 and partly T5 of the transitional season, and T6 is clearly in the summer season, but a clear reason for this distinction is missing. However, considering the daily exterior temperature of the closest public

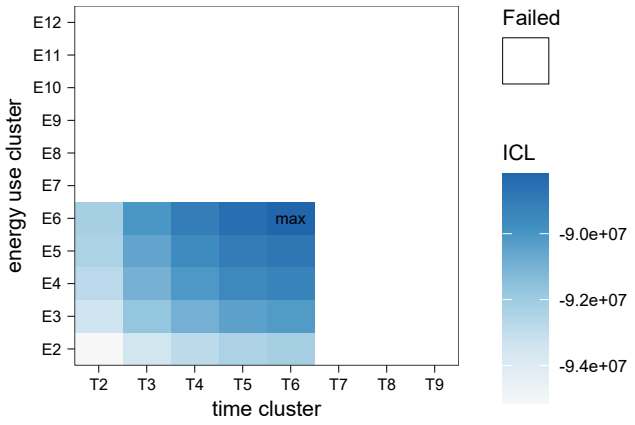


Figure 4: ICL of the different combinations of energy use and time clusters. Failed combinations are due to at least one empty cluster.

weather stations for the two years (Figure 5 b), a clear correlation between the external temperature and the time clusters becomes evident. A clear example of this correlation can be seen when comparing January 2020 and 2021. January 2021 was considerably colder than January 2020 and is thus, in another time cluster. However, one can see that the few warmer days in January 2021, which have a similar mean exterior temperature to January 2020, are assigned to the same time cluster as January 2020. Thus, the results capture variations of even a single day well. The mean temperature per cluster was computed to confirm further this correlation between daily mean exterior temperatures and time clusters (Figure 5 c). From this, it is visible that most clusters have a distinct mean exterior temperature and that the temperature change between clusters is not uniform. However, for some clusters e.g., T3 and T4 the difference in mean daily external temperature is minimal, indicating that exterior temperature alone does not sharply separate the time clusters. Further analyses against the mean global radiation (not shown), which is also correlated to the external temperature (Pearson correlation coefficient = 0.567), revealed no additional information. It was further analysed if restrictions due to COVID-19 had sufficient influence to lead to different seasonal clusters. Comparing, e.g., the last week of January 2021, where measures such as strongly recommended working from home were in place [76], to December 2021, where no restrictions were imposed, no apparent difference can be seen. Thus, it seems that restrictions due to COVID-19 had not an influence which would ”break” the seasonal pattern. Nevertheless, further analyses are necessary to identify possible minor impacts. Additionally, as no social-economic information is available for the used buildings, the job type of the occupants is also unknown and, thus, to which degree they were affected by COVID-19 restrictions. Overall, these results indicate that clustering based on a fixed season definition, e.g., based on a fixed date, does not lead to optimal clusters and thus, does not capture the season variation of the data correctly. At the same time, the results show the method's capability to capture seasonal variations even on daily granularity without relying on prior knowledge.

With the season variation analysed, as the next step, the obtained cluster mean curves for the different energy clusters were analysed. Figure 6 shows the average energy use curves as estimated by the SEM–Gibbs algorithm for

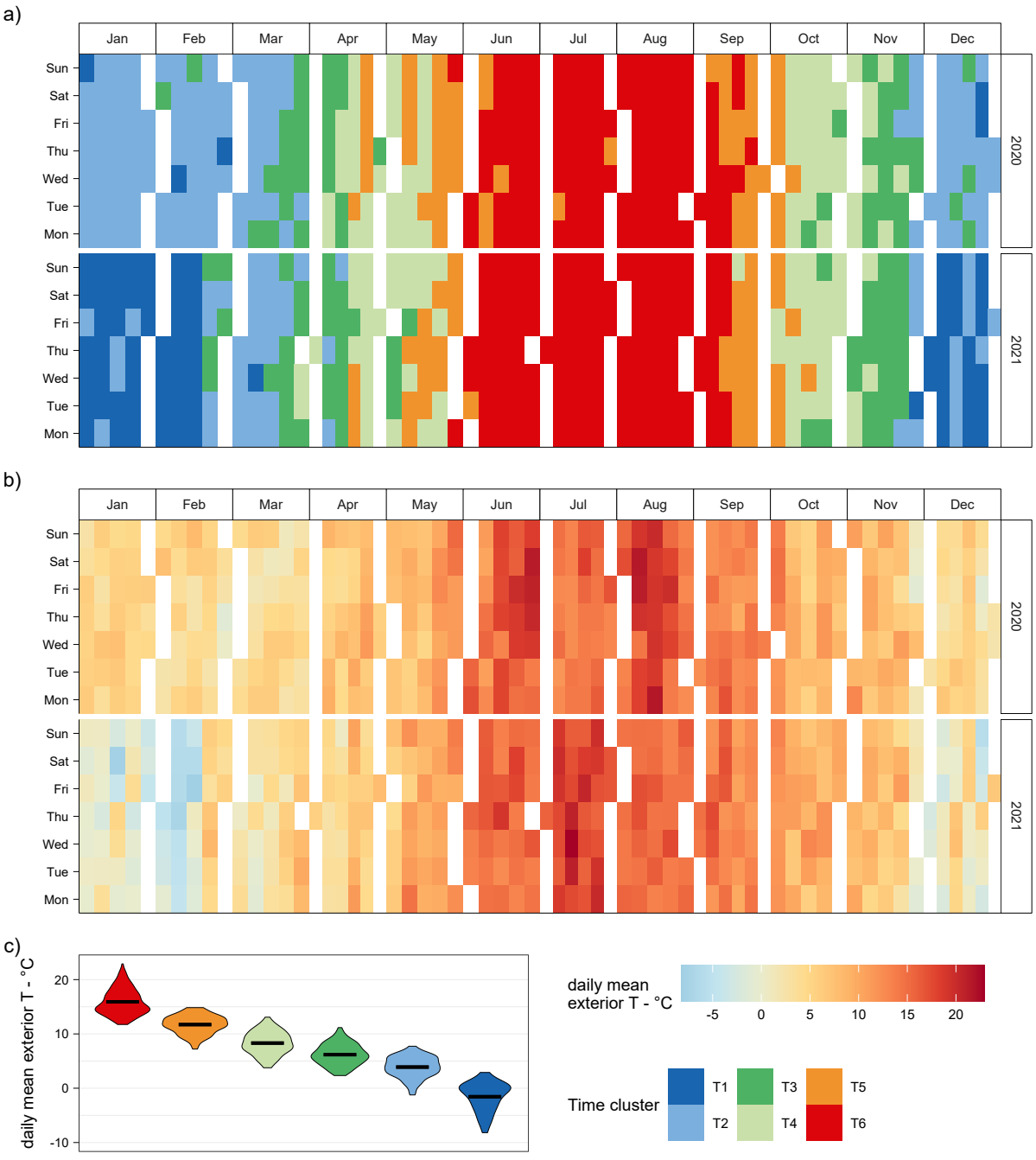


Figure 5: a) Distribution of time clusters for the selected two years. b) Daily mean exterior temperature at the SHM location c) daily mean exterior temperature per time cluster.

the six energy use clusters and their seasonal variation sorted based on the mean daily external temperature of the time clusters. As for the time clusters, the naming of the clusters was chosen to ease any further analysis. First, the seasonal trend in energy use is visible, and as expected, the energy use decreases with increasing external temperature. Between

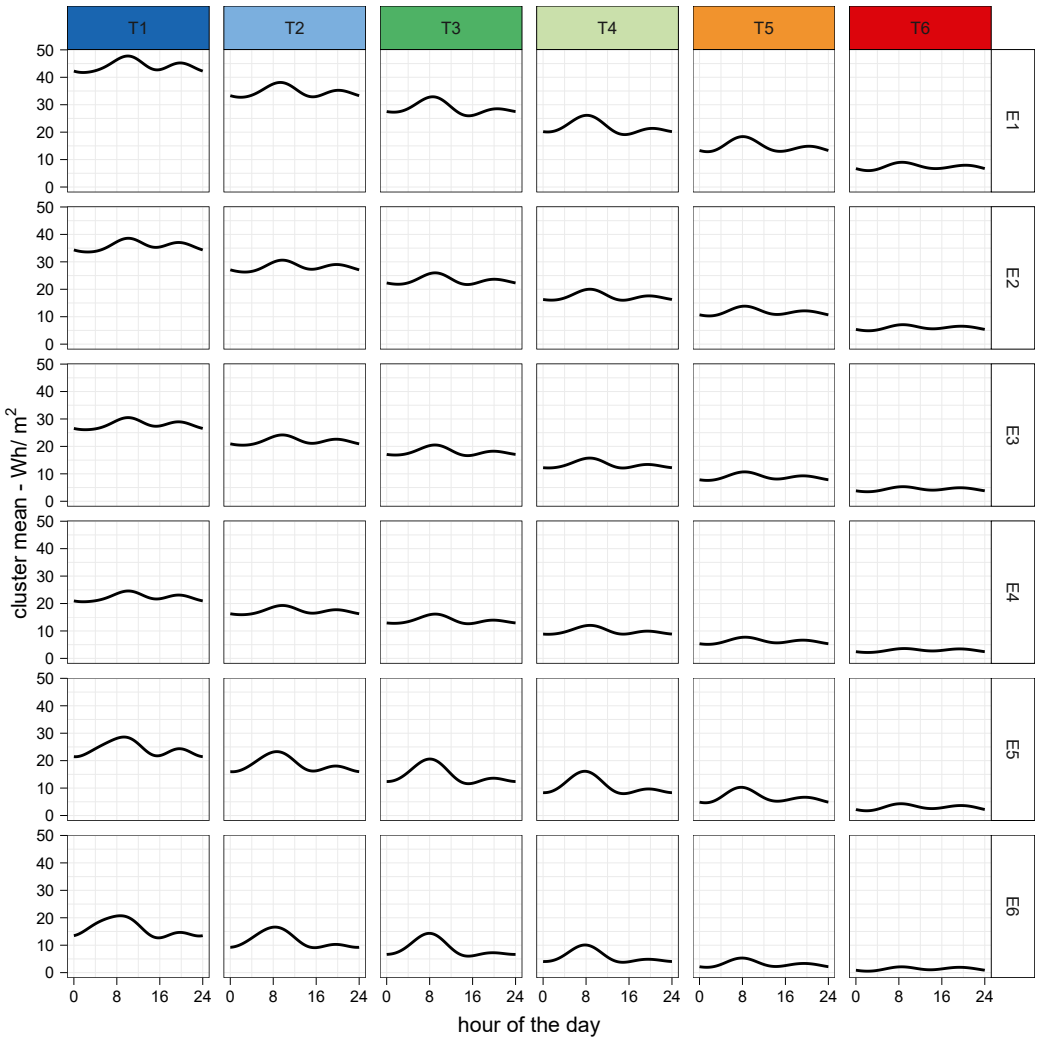


Figure 6: Average energy use curves as estimated by the SEM–Gibbs algorithm for the used co-clustering

the energy use clusters (E1-E6), clear differences in magnitude and shape but also similarities are visible. Overall, for all clusters, two peaks, one larger one in the morning around 8am and one smaller one in the afternoon around 8pm, are visible. Clusters E5 and E6 have different profiles compared to the remaining four clusters (E1-E4), with a more pronounced peak with a different shape in the morning. The four other clusters (E1-E4) mainly differentiate in the magnitude of the energy use. For time cluster T6, the energy use clusters show little to no variation, and the clearly visible peaks in the colder periods seem to diminish. Assuming that DHW mainly causes the peaks, one can explain this by the fact that in this period (mainly June and August), the occupant behaviour is not as regular due to, e.g., holidays and thus, the peaks are overall stronger evened out, when considering all buildings. This hypothesis also agrees with the cluster sizes (Figure 7), as E1 is by the smallest cluster, has thus, a minor equating effect and has the most pronounced pattern in T1. Further, recent research clearly shows that the energy for DHW decreases when

the exterior air temperature increases, as the cold water temperature increases and additionally, the user's comfort temperature likely decreases, reducing the needed energy for DHW [77, 78, 79]. Overall it can be said that the energy use magnitude seems to be the main differentiating criteria between the energy use clusters, with only E5 and E6 showing different patterns.

In terms of the cluster sizes (Figure 7), it can be said that besides the fact mentioned above, that E1 is significantly smaller than the other clusters (about one-third in size), the energy use clusters are relatively balanced. For the time clusters, all clusters but the summer cluster (T6) are fairly equal in size.

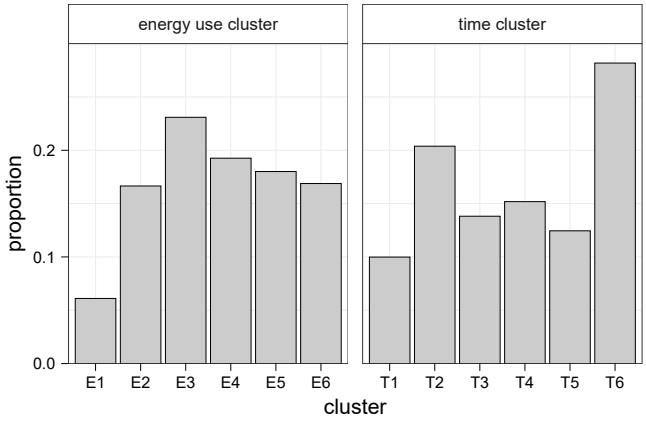


Figure 7: Cluster proportions for energy use and time clusters

5.2. Variable selection and Classification

With the energy clusters established, the next step was to perform variable selection and build the respective MLRGL and VSURF + optimised RF models. Therefore, first the BCs' variation across energy use clusters was visually analysed. After that, the results of the two used variable selection and classification techniques, MLRGL and VSURF + optimised RF, are presented.

5.2.1. Distribution analysis of building characteristic

Figure 8 shows the distribution of three selected BCs. The distributions of the remaining BCs are provided in the supplementary material. For the representative year (Figure 8 a), it can be seen that the most significant difference is shown for cluster E6 followed by E5, which both include the largest share of new or renovated buildings. These two energy clusters also showed a different daily pattern (Figure 8) from the other four energy clusters. This suggests that the pattern of energy use observed for these two clusters is more likely to be seen in new or refurbished buildings. For clusters E2 to E4, only a slight variation can be seen, with E4 including slightly more newer or renovated buildings while E2 has more buildings from the 1950s. E1 clearly includes the most old buildings, which corresponds well to the high energy use. For the number of bathrooms (Figure 8 b), which were normalised for the cluster size, it is to

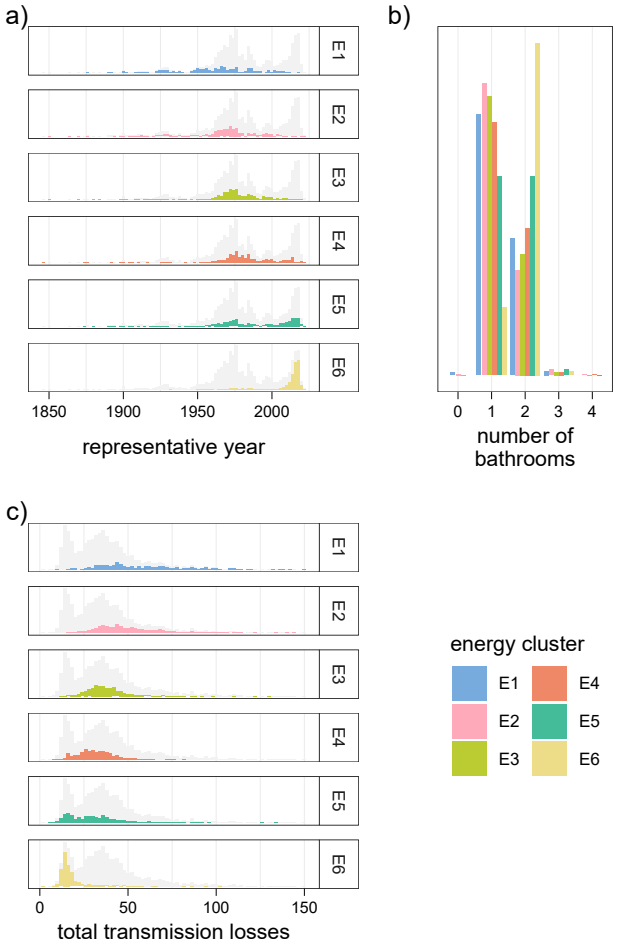


Figure 8: Distribution of selected variables across the energy use clusters - a) representative year, b) number of bathrooms, c) transmission losses (Table 1). The distribution of the number of bathrooms was normalised by the cluster size.

be highlighted that a few buildings have zero bathrooms. It is assumed that this is due to incorrect data, but this can not be determined with certainty, highlighting the also in Schaffer et al. [71] mentioned uncertainty in the BCs. E6 shows the most significant difference in distribution for the number of bathrooms, being the only cluster with more buildings with two than one bathroom, followed by E5, which has about equally many buildings with one and two bathrooms. Other than that, no clear trend can be seen. For the third shown BC, the total transmission losses (Figure 8 c)), a similar pattern as before can be seen. E6 is the most distinct cluster, and E5 shows a distribution between E2 to E4 and E6. For E1 and partly for E2, the influence of the old buildings is visible, showing the highest transmission losses. Between E3 and E4, hardly any difference is visible. The trend observed for these three BCs also holds for all the other BCs. It can therefore be said that, overall, E6 and, to a lesser extent, E5 and E1 show the most significant difference, while E2 to E4 show only slight variation between them.

5.3. Multinomial logistic regression with group lasso penalty

The MCC of the best MLRGL model on the test data of each of the five folds of the outer CV loop (Table 2) for all three tested situations (only BCs originating from BBR, from EPC and all BCs combined) is low with little variation between both the outer CV and the different sets of BCs. Consequently, including more detailed BCs does not seem to increase the MCC significantly, which means that nearly the same classification performance can be achieved with high-level statistical information than with in-depth detailed information about the building. Further, the average number of BCs (excluding the intercept) with non-zero coefficients shows that for the BCs originating from the BBR, only one BC was included, while only a few were excluded for the other two BC sets. The detailed breakdown of the variable selection of each of the five folds (Figure B.17) shows that the variable selection between the folds shows some variation for the BCs originating from the EPC and the combined set particularly for Fold1 while it was constant for the BCs from the BBR only. Thus indicating a sensitivity to the used subset of data.

Table 2: Mean MCC of the best MLRGL model on the test dataset of the five fold of the outer CV loop. Mean number of BCs (excluding the intercept) with non-zero coefficients averaged over the five folds.

dataset	mean MCC	mean nr. of BCs
BBR	0.258	1.0
EPC	0.308	13.8
Combined	0.308	21.6

A normalised confusion matrix (Figure 9) averaged over all five outer CVs was used to analyse the performance in more detail. From this it can be first seen that the BBR BCs differ significantly from the other two sets. There only two clusters E3 and E6 are predicted correctly, but therefore with a high accuracy, while non of the other clusters is predicted correctly. The other two BCs sets (EPC and combined) show a more even and to each other more similar pattern. For these two BCs sets, particularly E3 and to a lesser extend E6 are less often predicted correctly, therefore the other clusters show a more favourable pattern. Energy use clusters E6, which also showed the most distinct distribution (Section 5.2.1), has overall the best performance. Surprisingly E5 and E1, which both showed some difference in the BCs compared to the other energy use clusters, are the most difficult to predict correctly. From this, it can be concluded that some energy use clusters seem more directly related to specific BCs than others, where unknown parameters have a more significant influence. Consequently, the used BCs, which can be seen as extensive, are insufficient to correctly classify buildings in the found energy use clusters with MLRGL.

As the last step, it was analysed how the MCC changes over the number of included BCs. To reduce computational cost and as only small differences between the MCC on the test data and the MCC from the inner CV were observed, the MCC based on the inner CV was used for this analysis. Figure 10 shows that a simpler model can be obtained for the EPC and the combined BCs while decreasing the MCC only minorly. Based on this an alternative definition of the best model, as the simplest model with an MCC higher than the best model minus one standard deviation based on the

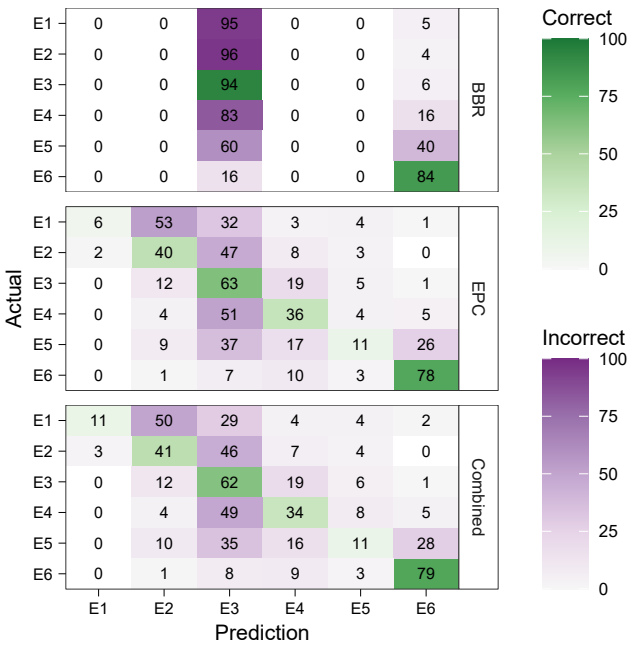


Figure 9: Normalised confusion matrix averaged over the the best model of MLRGL of each fold of the outer CV.

inner CV was tested. However, the results lead then to similar results as seen for the BBR BCs (Figure 9), so clusters E3 and E6 were predicted correctly more frequently while the other clusters were predicted correct significantly less frequent. As this is not seen as a desirable his was not further investigated. Furthermore, additional investigations showed that if more BCs for the BBR set are included, a similar confusion matrix as for the other two sets can be obtained with only a minor decrease in MCC. Thus highlighting that a more complex model is necessary for MLRGL to achieve an more even performance across all energy use clusters.

5.4. VSURF and optimised random forest

First, the MCC of VSURF on test data of the five folds of the outer CV was analysed for both sub-steps (interpretation and prediction) of the second step (Table 3). These results show that the achieved MCC differs only minimally from the one obtained from MLRGL and that the difference between the two steps is only minor (particularly considering the overall low MCC). Further focusing on the mean number of BCs used, a significant difference is visible between the interpretation and prediction steps. The prediction step includes significantly fewer BCs for data originating from BBR and the combined data with only a marginal reduction in mean MCC for the combined data while the MCC for BBR increases even slightly. It is to be highlighted that the combined BCs have a lower MCC for the prediction step than the BCs from the EPCs and BBR alone, which is counterintuitive. It is assumed that this is due to increased noise and redundant information in BCs originating from BBR and EPCs which increases the variance. Both steps include fewer BCs than MLRGL for BCs from EPC and the combined set. Based on these results, it was decided to use selected BCs from the predictor step of VSURF for further analysis.

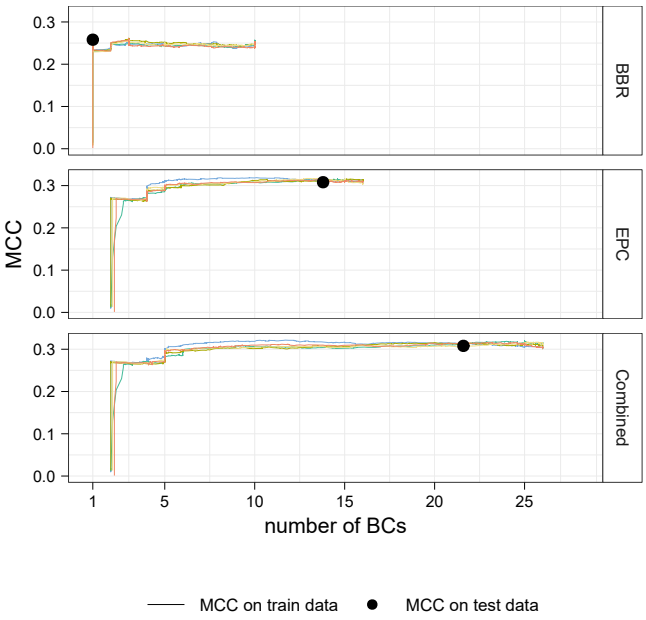


Figure 10: MCC of MLRGL for each of the five outer CV folds as a function of the number of BCs with non-zero coefficient

Table 3: Mean MCC and accuracy of the best VSURF model for both sub-steps (interpretation and prediction) on the fivefold outer CV loop test dataset and mean number of BCs used by these models.

	dataset	mean MCC	mean nr. of BCs
interpretation	BBR	0.268	5.8
	EPC	0.302	2.6
	Combined	0.305	15.6
prediction	BBR	0.276	2.0
	EPC	0.303	2.6
	Combined	0.279	3.0

Analysing the selected BCs of the predictor step in more detail (Figure 11), it can be seen that for the BBR BCs, the selection is consistent across all five outer folds and only the representative year and the information if the building was renovated or not is used. For the EPC BCs, variation can be seen for the natural ventilation in winter, which is included three out of five times, while the total transmission losses and the temperature difference of the heating system are always included. For the combined BCs, the total transmission losses and the representative year are always used, while the heating system's natural ventilation and temperature difference alternate. These results show that the information about the construction and renovation year already leads to a nearly as high MCC as detailed information about a building's heating system or ventilation use. However, given the low MCC, it is questionable

whether these BCs would be included if the necessary unknown information to predict the energy use clusters correctly were available. Nevertheless, for the hyperparameter-optimised RF, all at BCs were considered (two for BBR, three for EPC and four for the combined data).

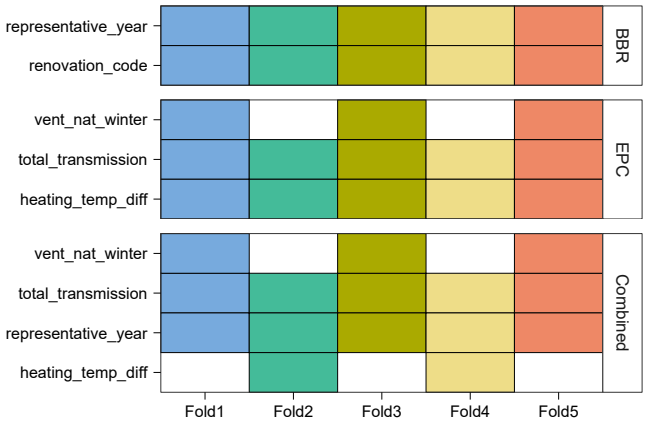


Figure 11: By the predictor step of VSURF selected BCs

The results of the tuned RF are only marginally better than the ones obtained from VSURF, with MCCs of 0.277, 0.313 and 0.318 for BCs from BBR, EPC, and the combined data. The main difference is that the increase by about 0.04 for the combined BCs, which now as expected lead to a slightly better result than only using information from EPCs. Again, a normalised confusion matrix is used for a more detailed analysis of the results (Figure 12). Overall,

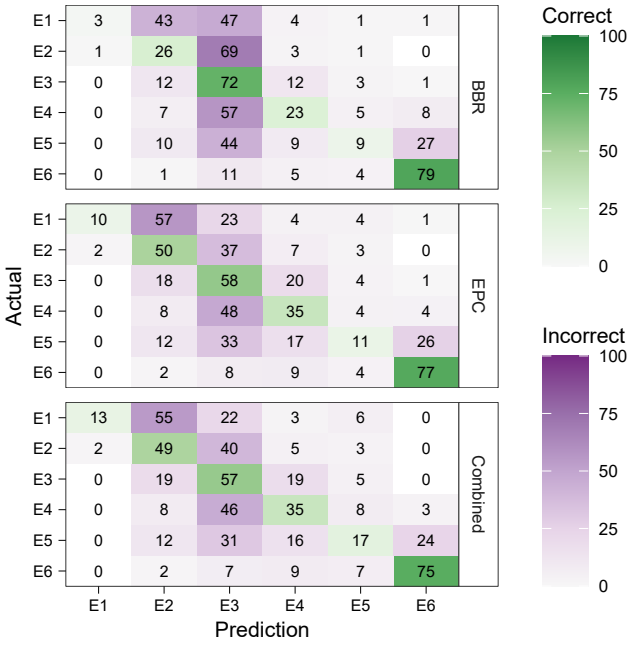


Figure 12: Normalised confusion matrix averaged over the the best model of the RF of each fold of the outer CV.

the same trends for MLRGL for the BCs based on EPC and the combined set are visible. E5 and E1 are the most challenging energy-use clusters to predict, while E6 is most often predicted correctly. The main difference can be observed for the BCs from the BBR. Here with the the additional information if a building is renovated more even performance across the clusters is while the MCC is higher then the one from MLRGL. Thus, these results indicated that using overall fewer BCs this approach performs overall better than MLRGL.

5.5. Reduced cluster analysis

Based on the results of the conducted analyses, it was decided to merge some of the energy use clusters based on expert knowledge to analyse whether this would lead to higher classification performance and consequently to a better understanding of the different energy use clusters. This simplification was done as it is expected that even if artificially simplified, the obtained information can still be valuable to stakeholders such as utility companies with no comparable possibilities at the moment. Based on similarities in their daily profiles in shape, magnitude and variations across time clusters, the clusters were merged as follows:

- E12: merged cluster E1 and E2
- E34: merged cluster E3 and E4
- E56: merged cluster E5 and E6

As the approach of VSURF combined with the optimised RF showed more promising results than MLRGL, overall fewer BCs used at comparable to superior performance, only this approach was used for the simplified clusters.

Firstly, the performance of VSURF is analysed (Table 4), which shows, as expected, a significant increase in MCC which is now in the range of 0.42 to 0.51. Again the MCC changes only minorly between the interpretation and prediction steps. However, more BCs are selected on average in the thresholding step compared to the non-simplified clusters. Nevertheless, in the prediction step, there is again an apparent reduction. Consequently, it was decided to use, as for the non-simplified clusters, the result of the prediction step for further analysis.

The detailed analysis of the selected BCs (Figure 13) shows for the BCs from the BBR that the representative year is chosen for each of the five folds, while the renovation code is now only selected two out of five times. Additionally, the roof material code was selected once, which was never selected for the not simplified clusters. For the BCs from the EPCs, the natural ventilation in winter is now chosen every time, while it was only selected three times for the not simplified clusters. The total transmission losses and the heating system temperature difference are again used in every fold. For the combined BCs, the representative year and the total transmission losses were again chosen for each fold. However, not a single time the heating temperature difference was selected, which was selected two times before. Additionally, once the number of rooms was included, which was never considered before. From these results, it is concluded that overall some variation to the not simplified clusters is visible no significant changes are observed. For the hyperparameter-optimised RF, the representative year and the renovation code will be used for the BCs from

Table 4: Accuracy of best VSURF model for both sub-steps (interpretation and prediction) on the test dataset of the five fold of the outer CV loop. Number of BCs refers to the number of BCs used by the best model.

	dataset	mean MCC	mean nr. of BCs
interpretation	BBR	0.438	5.0
	EPC	0.495	10.8
	Combined	0.514	19.6
prediction	BBR	0.421	1.6
	EPC	0.484	3.0
	Combined	0.473	3.2

the BBR. For the EPC-based BCs, all three selected BCs are used, and for the combined BCs, all but the number of rooms are considered.

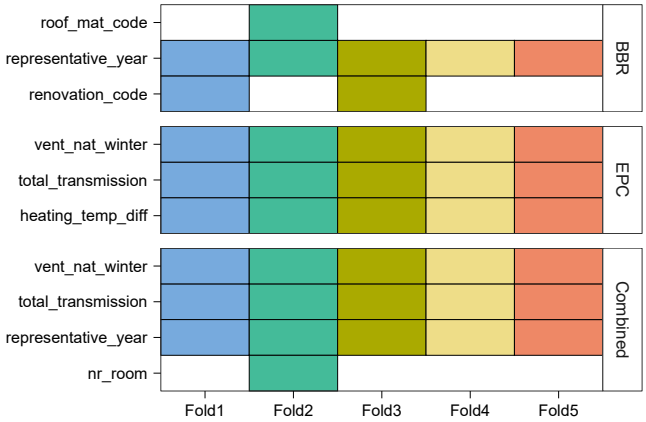


Figure 13: By the predictor step of VSURF selected BCs based on the simplified clusters

The results of the tuned RF are again only marginally better than the ones obtained from VSURF, with MCCs of 0.437, 0.499 and 0.501 for BCs from BBR, EPC, and the combined data. Again, a normalised confusion matrix is used for a more detailed analysis of the results (Figure 12). Form this it can be seen that now cluster E34 is the one most frequently predicted correctly while E12 and E56 are predicted about equally often correctly. Thereby both E12 and E56 are mainly misspredcited as E34. Overall the results show that this artificial simplification of the energy use clusters could be one option to increase the classification performance.

5.6. Analysis of cluster characteristics

Based on the above-presented results, it was decided to focus only on the simplified clusters, given the low MCC obtained for the not simplified clusters. Furthermore, based on the superior performance of VSURF in combination

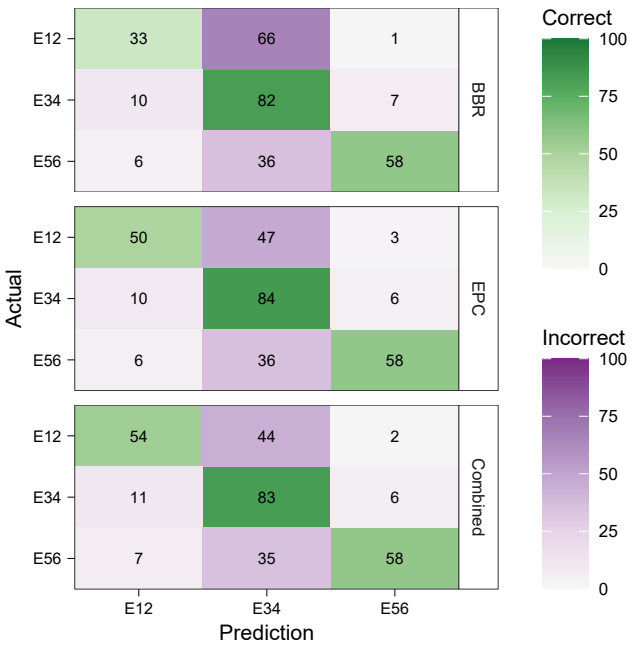


Figure 14: Normalised confusion matrix averaged over the the best model of the RF of each fold of the outer CV for the simplified clusters.

with the optimised RF, the clusters were only explored based on partition trees. In the following,exemplary the results for the BCs based on the BBR are shown. The results for the BCs based on EPC and combined data are shown in the supplementary materials

Figure 15 shows the resulting decision tree. The number under each cluster number indicates the number of correct classifications vs the number of node observations. From this, one can see that cluster E12 are mainly buildings till the beginning of the 1960s, independent of their renovation status. Cluster E34 are mainly buildings built or renovated from the beginning of the 1960s till 2002 or renovated till 2014. Cluster E56 are buildings built after 2009 or renovated after 2014. However, one must remember that the MCC of the shown classification tree is only 0.45 (accuracy = 0.64) for the data also used for constructing it (the training data). Consequently, this means that still a significant number of buildings is misclassified and thus do not follow the shown ”rules” of the decision tree.

6. Discussion & Conclusion

This work has proposed a novel approach to obtain firstly representative daily energy use curved from SHM data without relying on fixed season definitions using a recently developed co-clustering approach. Secondly, two different classification and variable selection approaches were used to investigate whether building characteristics originating from the BBR and EPCs can be used to explain the cluster characteristics. Furthermore, it was analysed how the BCs differ between the established energy use clusters to understand why buildings are in their respective cluster. This approach was showcased on two years of data from 4798 SHM single-family houses in Denmark, for which 26

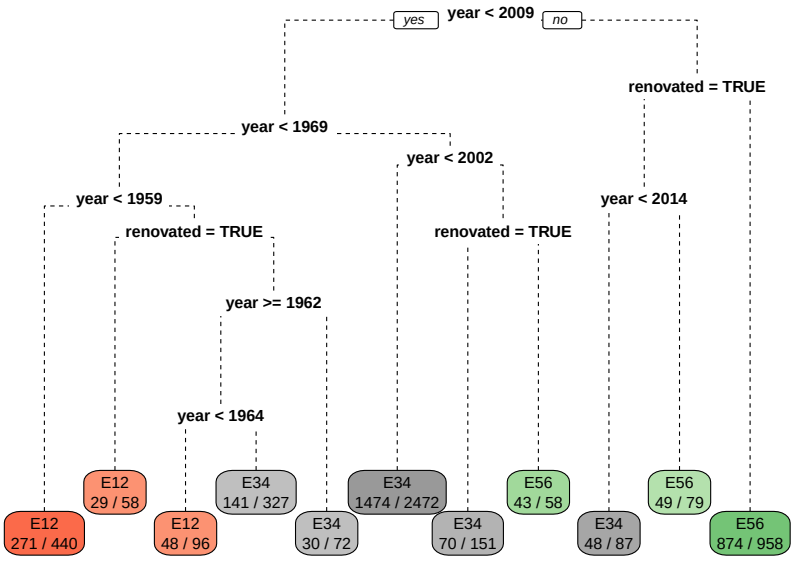


Figure 15: Pruned decision tree for the simplified energy use clusters using only the representative year (year) and renovation code (renovated)

different BCs were available. The results for the used co-clustering approach demonstrated that the found six time clusters, representing the seasonal variation while correlated to the exterior temperature, do not follow commonly used fixed season definitions and further vary across years. Consequently, showing that fixed season definitions do not correctly capture the seasonal variation of energy use in single-family houses. Additionally, the results indicated that other factors than exterior conditions, which could not be determined, influence the seasonal variation of the energy use profiles. It could also be demonstrated that restrictions due to COVID-19, such as a strong recommendation to work from home, did not lead to strong enough effects to break the general season pattern. However, further research is needed to confirm these results or identify more minor influences. The six found energy use clusters varied mainly in magnitude, and only two different profile shapes were observed. Further, the results seem to confirm that energy for DHW usage decreases with increasing external temperature. However, further investigations of energy use separated into heating and DHW, which is not readily available from commercial SHM data, are needed to confirm this. Overall, the clustering results showed encouraging results, which pave the way for further research. It is expected that the same clustering approach can give inside into different research and application-oriented questions either by using other from SHM meter available data, such as the instantaneous supply and return temperature readings or by pre-processing data differently, e.g., by first regressing the energy use against the external temperature.

The two used classification and variable techniques to identify important BCS and to analyse whether BCs can be used to predict energy clusters for buildings showed a comparable low MCC in the range of 0.25 to 0.31 across all three tested BCs subsets. MLRGL showed an overall lower performance and tendency to predict only clusters E3 and

E6 correctly if few BCs were used. VSURF lead to a more consistent performance across the energy use clusters and a higher MCC while considering only a few BCs. For both approaches, it is to be highlighted that the BCs from the EPC, which describe a building with a high level of detail, lead only to a minor improvement of the MCC: Thus, these results clearly indicate that BCs, even in the used level of detail, are insufficient to predict the energy use cluster of a building correctly. Consequently, they are also insufficient to understand why a building is in a particular energy use cluster with high certainty. Further research is needed to determine which additional information is needed, but is it hypothesised based on previous research [28, 80, 81] that occupant practices, such as the use of the heating system, daily routines and heating setpoint, and possible faults in the building could have a significant influence. Further, it can be concluded that the VSURF + optimised RF approach shows superior performance and more robust results and is thus seen as more suited if the presented approach is used by, e.g. utility companies.

Due to the low MCC, the six found energy use clusters were merged based on their similarity of energy use profiles and energy use magnitude into three clusters to investigate if this increases the MCC. Applying the VSURF and random RF approach to these simplified clusters showed a clear improvement in the MCC to about 0.44 to 0.50. Overall, the same BCs were selected for the not simplified clusters. Thus, confirming that using only the representative year and the information if a building was renovated leads to nearly the same MCC as detailed information about a building.

Based on the results for the simplified clusters, decision trees were used to understand the difference in BC across the clusters. The results for BCs based on BBR data showed that one can obtain insight into the different clusters in an easy-to-commute way understandable, also for non-experts. However, the results also showed that the MCC remained relatively low at 0.45 (accuracy = 0.64). Thus, such an explanation must be used cautiously as it is not valid for many buildings.

Overall it is expected that these results, despite their limitations, can be of high value for stakeholders such as DH utility companies, which at the moment have no comparable possibility. The proposed method allows to easily obtain daily energy use clusters and offers the possibility to gain insight into the difference between the buildings in the energy use clusters. Furthermore, the approach does not rely on expert knowledge for, e.g. hyperparameter optimisation or model selection, and is thus expected to be well suited for stakeholders commonly involved in the DH network. Furthermore, based on the proposed workflow, utility companies, can predict the theoretical energy use cluster of not yet to the DH-network connected buildings, which should support the planning for new areas.

7. CRediT authorship contribution statement

8. Funding

This work was funded by the Independent Research Fund Denmark under FOREFRONT project (0217-00340B).

Appendix A. Detailed building characteristic description

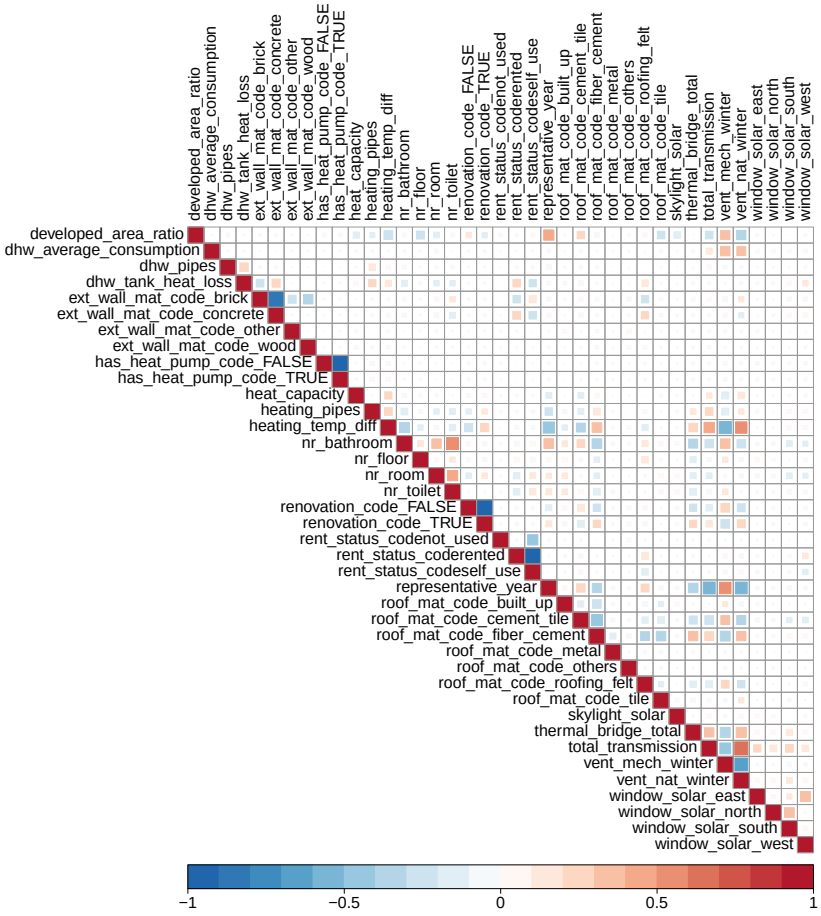


Figure A.16: Pearson correlation of the combined BCs originating from BBR and EPC data.

Appendix B. Detailed results of variable selection of the multinomial logistic regression with group lasso penalty

References

[1] Republic of Austria, Bundesgesetz zum Ausstieg aus der fossil betriebenen Wärmebereitstellung (Erneuerbare, 2022. URL: https://www.parlament.gv.at/PAKT/VHG/XXVII/ME/ME_00212/fname_1451879.pdf.

[2] M. L. Maach, Bred aftale i Folketinget: Fra 2035 skal ingen boliger opvarmes af gas, 2022. URL: https://www.dr.dk/nyheder/indland/bred-energiaftale-skal-goere-danskerne-fri-af-russisk-gas-fra-2035.

[3] European Commission, State aid: Commission approves €2.98 billion German scheme to promote green district heating, 2022. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_4823.

[4] H. Lund, S. Werner, R. Wiltshire, S. Svendsen, J. E. Thorsen, F. Hvelplund, B. V. Mathiesen, 4th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems, Energy 68 (2014) 1–11. URL: http://dx.doi.org/10.1016/j.energy.2014.02.089. doi:10.1016/j.energy.2014.02.089.

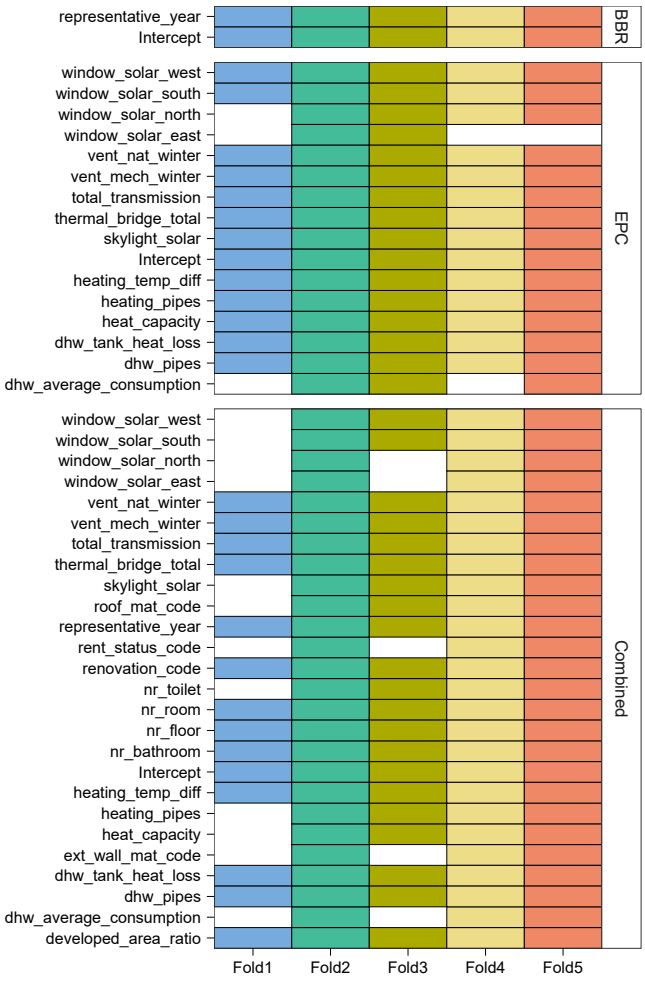


Figure B.17: Detailed results of the variable selection of the best model of MLRGL of each fold of the outer CV including the intercept

[5] European Commission, Energy performance of buildings directive, Retrieved: 2020-08-03, 2022. URL: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en.

[6] European Commission, EU Buildings Factsheets — Energy, 2022. URL: https://ec.europa.eu/energy/eu-buildings-factsheets_en.

[7] Rambøll, D2.3 - District Heating and Cooling Stock at EU level, Technical Report, 2020. URL: https://www.wedistrict.eu/wp-content/uploads/2020/11/WEDISTRICT_WP2_D2.3-District-Heating-and-Cooling-stock-at-EU-level.pdf.

[8] European Parliament, Directive (EU) 2018/2002 amending Directive 2012/27/EU on energy efficiency, Official Journal of the European Union (2018). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018L2002&from=EN.

[9] M. H. Kristensen, R. Choudhary, S. Petersen, Bayesian calibration of building energy models: Comparison of predictive accuracy using metered utility data of different temporal resolution, Energy Procedia 122 (2017) 277–282. URL: https://linkinghub.elsevier.com/retrieve/pii/S1876610217329259. doi:10.1016/j.egypro.2017.07.322.

[10] M. H. Kristensen, R. E. Hedegaard, S. Petersen, Long-term forecasting of hourly district heating loads in urban areas using hierarchical archetype modeling, Energy 201 (2020) 117687. URL: https://doi.org/10.1016/j.energy.2020.117687https://linkinghub.elsevier.com/retrieve/pii/S0360544220307945. doi:10.1016/j.energy.2020.117687.

[11] M. H. Kristensen, R. E. Hedegaard, S. Petersen, Hierarchical calibration of archetypes for urban building energy modeling, Energy and Buildings 175 (2018) 219–234. URL: https://doi.org/10.1016/j.enbuild.2018.07.030https://linkinghub.elsevier.com/retrieve/pii/S0378778818312532. doi:10.1016/j.enbuild.2018.07.030.

[12] R. E. Hedegaard, M. H. Kristensen, T. H. Pedersen, A. Brun, S. Petersen, Bottom-up modelling methodology for urban-scale analysis of residential space heating demand response, Applied Energy 242 (2019) 181–204. URL: https://doi.org/10.1016/j.apenergy.2019.03.063https://linkinghub.elsevier.com/retrieve/pii/S0306261919304726. doi:10.1016/j.apenergy.2019.03.063.

[13] M. Lumbreras, R. Garay-Martinez, B. Arregi, K. Martin-Escudero, G. Diarce, M. Raud, I. Hagu, Data driven model for heat load prediction in buildings connected to District Heating by using smart heat meters, Energy 239 (2022) 122318. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360544221025664. doi:10.1016/j.energy.2021.122318.

[14] H. G. Bergsteinsson, P. B. Vetter, J. K. Møller, H. Madsen, Estimating temperatures in a district heating network using smart meter data, Energy Conversion and Management 269 (2022) 116113. URL: https://linkinghub.elsevier.com/retrieve/pii/S0196890422008974. doi:10.1016/j.enconman.2022.116113.

[15] S. Idowu, S. Saguna, C. Åhlund, O. Schelén, Applied machine learning: Forecasting heat load in district heating system, Energy and Buildings 133 (2016) 478–488. doi:10.1016/j.enbuild.2016.09.068.

[16] P. Gianniou, C. Reinhart, D. Hsu, A. Heller, C. Rode, Estimation of temperature setpoints and heat transfer coefficients among residential buildings in Denmark based on smart meter data, Building and Environment 139 (2018) 125–133. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360132318302762. doi:10.1016/j.buildenv.2018.05.016.

[17] D. Leiria, H. Johra, A. Marszal-Pomianowska, M. Z. Pomianowski, P. Kvols Heiselberg, Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case, Smart Energy 3 (2021) 100035. URL: https://doi.org/10.1016/j.segy.2021.100035https://linkinghub.elsevier.com/retrieve/pii/S2666955221000356. doi:10.1016/j.segy.2021.100035.

[18] Z. Ma, R. Yan, N. Nord, A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings, Energy 134 (2017) 90–102. URL: https://linkinghub.elsevier.com/retrieve/pii/S0360544217309878. doi:10.1016/j.energy.2017.05.191.

[19] H. Johra, D. Leiria, P. Heiselberg, A. Marszal-Pomianowska, T. Tvedebrink, Treatment and analysis of smart energy meter data from a cluster of buildings connected to district heating: A Danish case, E3S Web of Conferences 172 (2020) 12004. URL: https://www.e3s-conferences.org/articles/e3sconf/abs/2020/32/e3sconf_nsb2020_12004/e3sconf_nsb2020_12004.htmlhttps://www.e3s-conferences.org/10.1051/e3sconf/202017212004. doi:10.1051/e3sconf/202017212004.

[20] C. Wang, Y. Du, H. Li, F. Wallin, G. Min, New methods for clustering district heating users based on consumption patterns, Applied Energy 251 (2019) 113373. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261919310475. doi:10.1016/j.apenergy.2019.113373.

[21] C. M. R. do Carmo, T. H. Christensen, Cluster analysis of residential heat load profiles and the role of technical and household characteristics, Energy and Buildings 125 (2016) 171–180. URL: http://dx.doi.org/10.1016/j.enbuild.2016.04.079https://linkinghub.elsevier.com/retrieve/pii/S0378778816303565. doi:10.1016/j.enbuild.2016.04.079.

[22] P. Gianniou, X. Liu, A. Heller, P. S. Nielsen, C. Rode, Clustering-based analysis for residential district heating data, Energy Conversion and Management 165 (2018) 840–850. URL: https://doi.org/10.1016/j.enconman.2018.03.015https://linkinghub.elsevier.com/retrieve/pii/S019689041830236X. doi:10.1016/j.enconman.2018.03.015.

[23] Y. Yang, R. Li, T. Huang, Smart meter data analysis of a building cluster for heating load profile quantification and peak load shifting, Energies 13 (2020) 4343. URL: https://www.mdpi.com/1996-1073/13/17/4343/htmhttps://www.mdpi.com/1996-1073/13/17/4343. doi:10.3390/en13174343.

[24] E. Guelpa, S. Deputato, V. Verda, Thermal request optimization in district heating networks using a clustering approach, Applied Energy 228 (2018) 608–617. URL: https://doi.org/10.1016/j.apenergy.2018.06.041. doi:10.1016/j.apenergy.2018.06.041.

[25] S. Abghari, V. Boeva, J. Brage, C. Johansson, H. Grahn, N. Lavesson, Higher order mining for monitoring district heating sub-

stations, in: Proceedings - 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019, IEEE, 2019, pp. 382–391. URL: http://www.energimyndigheten.se/en/sustainability/households/https://ieeexplore.ieee.org/document/8964173/. doi:10.1109/DSAA.2019.00053.

[26] E. Calikus, S. Nowaczyk, A. Sant'Anna, H. Gadd, S. Werner, A data-driven approach for discovering heat load patterns in district heating, Applied Energy 252 (2019) 113409. URL: https://linkinghub.elsevier.com/retrieve/pii/S0306261919310839. doi:10.1016/j.apenergy.2019.113409.

[27] E. Calikus, S. Nowaczyk, A. Sant'Anna, S. Byttner, Ranking Abnormal Substations by Power Signature Dispersion, Energy Procedia 149 (2018) 345–353. URL: https://linkinghub.elsevier.com/retrieve/pii/S1876610218304946. doi:10.1016/j.egypro.2018.08.198.

[28] H. Gadd, S. Werner, Fault detection in district heating substations, Applied Energy 157 (2015) 51–59. doi:10.1016/J.APENERGY.2015.07.061.

[29] Y. Wang, Q. Chen, T. Hong, C. Kang, Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges, IEEE Transactions on Smart Grid 10 (2019) 3125–3148. URL: https://ieeexplore.ieee.org/document/8322199/. doi:10.1109/TSG.2018.2818167.

[30] F. Tounquet, C. Alaton, Benchmarking Smart Metering Deployment in EU-28, December, 2020. URL: https://op.europa.eu/en/publication-detail/-/publication/b397ef73-698f-11ea-b735-01aa75ed71a1/language-en. doi:10.2833/492070.

[31] H. Pörtner, D. Roberts, E. Poloczanska, K. Mintenbeck, M. Tignor, A. Alegría, M. CRAIG, S. LANGSDORF, S. LÖSCHKE, V. MÖLLER, A. OKEM, IPCC Sixth Assessment Report, Technical Report, 2022.

[32] EPA, Seasonality and Climate Change: A Review of Observed Evidence in the United States, Technical Report December, U.S. Environmental Protection Agency, 2021.

[33] European Environment Agency, What will the future bring when it comes to climate hazards? - Overview — European Environment Agency, 2023. URL: https://www.eea.europa.eu/publications/europes-changing-climate-hazards-1/what-will-the-future-bring.

[34] T. Warren Liao, Clustering of time series data - A survey, Pattern Recognition 38 (2005) 1857–1874. URL: www.elsevier.com/locate/patcog. doi:10.1016/J.PATCOG.2005.01.025.

[35] S. Aghabozorgi, A. Seyed Shirkhorshidi, T. Ying Wah, Time-series clustering – A decade review, Information Systems 53 (2015) 16–38. doi:10.1016/J.IS.2015.04.007.

[36] G. Mbiydzenyuy, S. Nowaczyk, H. Knutsson, D. Vanhoudt, J. Brage, E. Calikus, Opportunities for Machine Learning in District Heating, Applied Sciences 11 (2021) 6112. URL: https://www.mdpi.com/2076-3417/11/13/6112/htmhttps://www.mdpi.com/2076-3417/11/13/6112. doi:10.3390/app11136112.

[37] C. Bouveyron, L. Bozzi, J. Jacques, F. X. Jollois, The functional latent block model for the co-clustering of electricity consumption curves, Journal of the Royal Statistical Society. Series C: Applied Statistics 67 (2018) 897–915. URL: https://rss-onlinelibrary-wiley-com.zorac.aub.aau.dk/doi/10.1111/rssc.12260https://onlinelibrary.wiley.com/doi/10.1111/rssc.12260. doi:10.1111/rssc.12260.

[38] F. Divina, F. Vela, M. Torres, Biclustering of Smart Building Electric Energy Consumption Data, Applied Sciences 9 (2019) 222. URL: https://www.mdpi.com/2076-3417/9/2/222/htmhttps://www.mdpi.com/2076-3417/9/2/222http://www.mdpi.com/2076-3417/9/2/222. doi:10.3390/app9020222.

[39] R Core Team, R: A Language and Environment for Statistical Computing, 2022. URL: https://www.r-project.org/.

[40] C. Bouveyron, J. Jacques, A. Schmutz, funLBM: Model-Based Co-Clustering of Functional Data, 2022. URL: https://cran.r-project.org/package=funLBM.

[41] G. Govaert, M. Nadif, Co-Clustering, volume 9781848214, John Wiley & Sons, Inc., Hoboken, USA, 2013. URL: http://doi.wiley.com/10.1002/9781118649480. doi:10.1002/9781118649480.

[42] J. Ramsay, Silverman, Functional data analysis, 2 ed., Springer New York, 2005.

[43] A. R. Hansen, D. Leiria, H. Johra, A. Marszal-Pomianowska, Who Produces the Peaks? Household Variation in Peak Energy Demand for Space Heating and Domestic Hot Water, Energies 15 (2022) 9505. URL: https://www.mdpi.com/1996-1073/15/24/9505/htmhttps://www.mdpi.com/1996-1073/15/24/9505. doi:10.3390/en15249505.

[44] J. P. Gouveia, J. Seixas, Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys, Energy and Buildings 116 (2016) 666–676. URL: http://dx.doi.org/10.1016/j.enbuild.2016.01.043. doi:10.1016/j.enbuild.2016.01.043.

[45] F. McLoughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study, Energy and Buildings 48 (2012) 240–248. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778812000680. doi:10.1016/j.enbuild.2012.01.037.

[46] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior, Energy 55 (2013) 184–194. URL: http://dx.doi.org/10.1016/j.energy.2013.03.086https://linkinghub.elsevier.com/retrieve/pii/S0360544213002831. doi:10.1016/j.energy.2013.03.086.

[47] A. Albert, R. Rajagopal, Smart Meter Driven Segmentation: What Your Consumption Says About You, IEEE Transactions on Power Systems 28 (2013) 4019–4030. URL: https://ieeexplore.ieee.org/document/6545387/. doi:10.1109/TPWRS.2013.2266122.

[48] T. Hastie, R. Tibshirani, J. H. Friedman, J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, volume 2, Springer, 2009.

[49] M. Vincent, N. R. Hansen, Sparse group lasso and high dimensional multinomial classification, Computational Statistics & Data Analysis 71 (2014) 771–786. URL: http://dx.doi.org/10.1016/j.csda.2013.06.004https://linkinghub.elsevier.com/retrieve/pii/S0167947313002168. doi:10.1016/j.csda.2013.06.004.

[50] Q. Wang, G. Augenbroe, J. H. Kim, L. Gu, Meta-modeling of occupancy variables and analysis of their impact on energy outcomes of office buildings, Applied Energy 174 (2016) 166–180. doi:10.1016/j.apenergy.2016.04.062.

[51] W. Gang, G. Augenbroe, S. Wang, C. Fan, F. Xiao, An uncertainty-based design optimization method for district cooling systems, Energy 102 (2016) 516–527. URL: http://dx.doi.org/10.1016/j.energy.2016.02.107. doi:10.1016/j.energy.2016.02.107.

[52] Y. Sun, L. Gu, C. F. Wu, G. Augenbroe, Exploring HVAC system sizing under uncertainty, Energy and Buildings 81 (2014) 243–252. doi:10.1016/j.enbuild.2014.06.026.

[53] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society. Series B: Statistical Methodology 68 (2006) 49–67. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2005.00532.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.xhttps://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00532.x. doi:10.1111/j.1467-9868.2005.00532.x.

[54] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A Sparse-Group Lasso, Journal of Computational and Graphical Statistics 22 (2013) 231–245. URL: https://www.tandfonline.com/doi/abs/10.1080/10618600.2012.681250http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.681250. doi:10.1080/10618600.2012.681250.

[55] G. Jurman, S. Riccadonna, C. Furlanello, A comparison of MCC and CEN error measures in multi-class prediction, PLoS ONE 7 (2012) 1–8. doi:10.1371/journal.pone.0041882.

[56] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 1–13. URL: https://link.springer.com/articles/10.1186/s12864-019-6413-7https://link.springer.com/article/10.1186/s12864-019-6413-7. doi:10.1186/s12864-019-6413-7.

[57] A. Gelman, Scaling regression inputs by dividing by two standard deviations, Statistics in Medicine 27 (2008) 2865–2873. URL: www.interscience.wiley.com. doi:10.1002/sim.3107.

[58] M. Vincent, N. R. Hansen, msgl: Multinomial sparse group lasso, 2019.

[59] R. Genuer, J. M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recognition Letters 31 (2010) 2225–2236. URL: http://dx.doi.org/10.1016/j.patrec.2010.03.014. doi:10.1016/j.patrec.2010.03.014.

[60] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, VSURF: Variable Selection Using Random Forests, 2022. URL: https://cran.r-project.org/package=VSURF.

[61] R. Genuer, J. M. Poggi, C. Tuleau-Malot, VSURF: An R package for variable selection using random forests, R Journal 7 (2015) 19–33. URL: http://cran.r-project.org/package=VSURF. doi:10.32614/rj-2015-018.

[62] J. S. Virdi, W. Peng, A. Sata, Feature selection with LASSO and VSURF to model mechanical properties for investment casting, in: 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), IEEE, 2019, pp. 1–6. URL: https://ieeexplore.ieee.org/document/8862141/. doi:10.1109/ICCIDS.2019.8862141.

[63] J. L. Speiser, M. E. Miller, J. Tooze, E. Ip, A comparison of random forest variable selection methods for classification prediction modeling, Expert Systems with Applications 134 (2019) 93–101. URL: /pmc/articles/PMC7508310//pmc/articles/PMC7508310/?report= abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508310/https://linkinghub.elsevier.com/retrieve/pii/S0957417419303574. doi:10.1016/j.eswa.2019.05.028.

[64] P. Probst, M. N. Wright, A. L. Boulesteix, Hyperparameters and tuning strategies for random forest, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) 1–15. doi:10.1002/widm.1301.

[65] Z. Zhang, M. W. Kattan, Drawing Nomograms with R: applications to categorical outcome and survival data, Annals of Translational Medicine 5 (2017) 211–211. URL: https://atm.amegroups.com/article/view/14736/htmlhttps://atm.amegroups.com/article/view/14736http://atm.amegroups.com/article/view/14736/15089. doi:10.21037/atm.2017.04.01.

[66] T. Barlow, P. Neville, Case study: visualization for decision tree analysis in data mining, in: IEEE Symposium on Information Visualization, 2001. INFOVIS 2001., IEEE, 2001, pp. 149–152. URL: http://ieeexplore.ieee.org/document/963292/. doi:10.1109/INFVIS.2001.963292.

[67] S. Van Den Elzen, J. J. Van Wijk, BaobabView: Interactive construction and analysis of decision trees, in: VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings, IEEE, 2011, pp. 151–160. URL: http://ieeexplore.ieee.org/document/6102453/. doi:10.1109/VAST.2011.6102453.

[68] O. Parisot, Y. Didry, P. Bruneau, B. Otjacques, Data Visualization using Decision Trees and Clustering, in: Proceedings of the 5th International Conference on Information Visualization Theory and Applications, SCITEPRESS - Science and and Technology Publications, 2014, pp. 80–87. URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0004740800800087. doi:10.5220/0004740800800087.

[69] T. Therneau, B. Atkinson, rpart: Recursive Partitioning and Regression Trees, 2022. URL: https://cran.r-project.org/package=rpart.

[70] S. Milborrow, rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart', 2022. URL: https://cran.r-project.org/package=rpart.plot.

[71] M. Schaffer, M. Veit, A. Marszal-Pomianowska, M. Frandsen, M. Zbigniew Pomianowski, E. Dichmann, C. Grau Sørensen, J. Kragh, Dataset of smart heat and water meter data with accompanying building characteristics, 2023.

[72] M. Schaffer, T. Tvedebrink, A. Marszal-Pomianowska, Three years of hourly data from 3021 smart heat meters installed in Danish residential buildings, Scientific Data 9 (2022) 420. URL: https://www.nature.com/articles/s41597-022-01502-3. doi:10.1038/s41597-22-01502-3.

[73] M. Schaffer, D. Leiria, J. E. Vera-Valdés, A. Marszal-Pomianowska, Increasing the accuracy of low-resolution commercial smart heat meter data and analysing its error, 2023 Accepted for: 2023 European Conference on Computing in Construction 40th International CIB W78 Conference

[74] Danish Property Assessment Agency, Bygnings- og Boligregistret, 2023. URL: https://bbr.dk/om-bbr.

[75] J. Fox, G. Monette, Generalized Collinearity Diagnostics, Journal of the American Statistical Association 87 (1992) 178–183. URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190. doi:10.1080/01621459.1992.10475190.

[76] M. Guldbrandt Brønnum, J. Skriver Steffensen, C. Monin, C. Henriksen, Coronarestriktioner er fortid: Se tidslinjen over pandemien, 2022. URL: https://www.tv2nord.dk/coronavirus/coronarestriktioner-er-fortid-se-tidslinjen-over-pandemien.

[77] D. George, N. S. Pearre, L. G. Swan, High resolution measured domestic hot water consumption of Canadian homes, Energy and

Buildings 109 (2015) 304–315. URL: http://dx.doi.org/10.1016/j.enbuild.2015.09.067https://linkinghub.elsevier.com/retrieve/pii/S0378778815303066. doi:10.1016/j.enbuild.2015.09.067.

[78] E. Fuentes, L. Arce, J. Salom, A review of domestic hot water consumption profiles for application in systems and buildings energy performance analysis, Renewable and Sustainable Energy Reviews 81 (2018) 1530–1547. URL: http://dx.doi.org/10.1016/j.rser.2017.05.229https://linkinghub.elsevier.com/retrieve/pii/S1364032117308614. doi:10.1016/j.rser.2017.05.229.

[79] I. Meireles, V. Sousa, B. Bleys, B. Poncelet, Domestic hot water consumption pattern: Relation with total water consumption and air temperature, Renewable and Sustainable Energy Reviews 157 (2022) 112035. URL: https://doi.org/10.1016/j.rser.2021.112035https://linkinghub.elsevier.com/retrieve/pii/S1364032121012971. doi:10.1016/j.rser.2021.112035.

[80] T. S. Larsen, H. N. Knudsen, A. M. Kanstrup, E. T. Christiansen, K. Gram-Hanssen, M. Mosgaard, H. Brohus, P. Heiselberg, J. Rose, Occupants Influence on the Energy Consumption of Danish Domestic Buildings (2010) 77.

[81] R. Andersen, The influence of occupants' behaviour on energy consumption investigated in 290 identical dwellings and in 35 apartments, 10th International Conference on Healthy Buildings 2012 3 (2012) 2279–2280.