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

Published in:
Applied Energy

DOI (link to publication from Publisher):
10.1016/j.apenergy.2024.123586

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

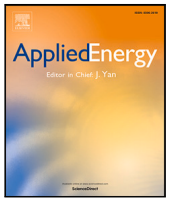
Citation for published version (APA):
Schaffer, M., Vera-Valdés, J. E., & Marszal-Pomianowska, A. (2024). Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses. *Applied Energy*, *371*, Article 123586. https://doi.org/10.1016/j.apenergy.2024.123586

Applied Energy 371 (2024) 123586

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

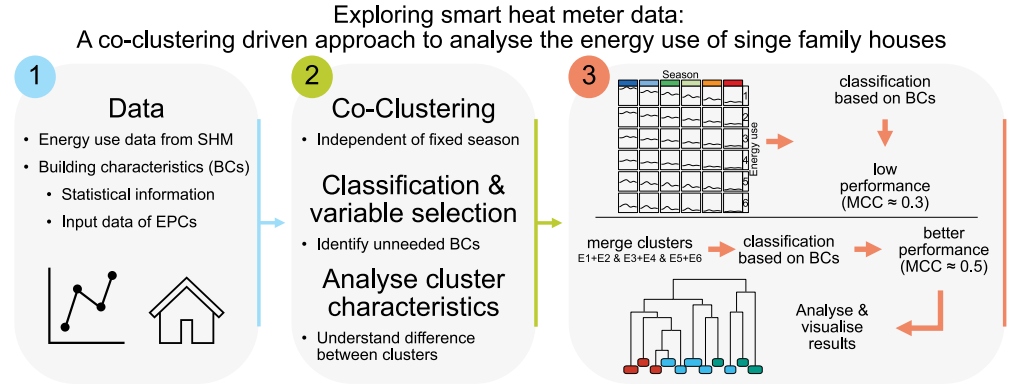


Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses

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

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Smart heat meter
District heating
Co-clustering
Classification
Variable selection

ABSTRACT

The ongoing digitalisation of the district heating sector, particularly the installation of smart heat meters (SHMs), is generating data with unprecedented extent and temporal resolution. This data offers potential insights into heat energy use at a large scale, supporting policymakers and district heating utility companies in transforming the building sector. Clustering is crucial for representing this wealth of data in human-understandable groups, necessitating consideration of seasonality.

Advancing current research in clustering SHM data, this work applies an established co-clustering approach, FunLBM, considering seasonal variation without fixed season definitions. Furthermore, to enhance the understanding of differentiating factors between clusters, the possibility to understand cluster memberships based on 26 building characteristics was analysed using classification and variable selection methods.

Applying FunLBM on a large-scale hourly dataset from single-family houses revealed six well-separated energy use clusters each distributed over six-temporal clusters, which are correlated with the exterior temperature, yet not following fixed seasons. Variable selection and classification showed that building characteristics describing the building with a high level of detail are insufficient to explain cluster membership (Matthew's correlation coefficient (MCC) ≈0.3).

By merging the energy use clusters based on profile and magnitude similarities, classification performance significantly improved (MCC ≈0.5). In both cases, simple and readily available building characteristics yield similar insights to detailed ones, emphasising their cost-effectiveness and practicality.

* Corresponding author.
E-mail address: msch@build.aau.dk (M. Schaffer).

https://doi.org/10.1016/j.apenergy.2024.123586
Received 21 September 2023; Received in revised form 4 April 2024; Accepted 27 May 2024
Available online 13 June 2024

Nomenclature

BBR	Danish Building and Dwelling Register
BCs	Building Characteristics
CV	Cross-Validation
DH	District Heating
DHW	Domestic Hot Water
EPC	Energy Performance Certificate
EU	European Union
FunLBM	Functional Latent Block Model
GCV	Generalised Cross-Validation
GMM	Gaussian Mixture Model
GVIF	Generalised Variance Inflation Factor
ICL	Integrated Information Likelihood Criterion
MCC	Correlation Coefficient
MLRGL	Multinomial Logistic Regression fitted using Group Least Absolute Shrinkage and Selection Operator (Group Lasso)
OOB error	Out-Of-Bag error
RF	Random Forest
SEM	Smart Electricity Meter
SHM	Smart Heat Meter
SH	Space Heating
VIF	Variance Inflation Factor
VSURF	Variable Selection Using Random Forests

1. Introduction

In light of recent geopolitical changes and the resulting need to reduce the European Union's (EU) dependence on natural gas, District Heating (DH) has come to the fore in some EU countries [1–3]. DH can not only play an essential role in reducing the dependency on natural gas, but also in the needed future reduction of CO_2 emissions if renewable 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], with a low return temperature as the return temperature of buildings directly influences the production efficiency and, consequently, the network efficiency. Networks relying on renewables and waste heat have thereby a 6 to 7 times higher supply temperature cost gradient (cost per MW h/°C) compared to traditional networks [5]. However, nowadays, nearly 75% of the EU's building stock is energy inefficient [6]. 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 [7], highlighting the challenges the building sector must face to enable a fully decarbonised building stock by 2050 [6]. In-depth knowledge of the building stock is essential for facilitating its transformation. This knowledge also empowers DH utility companies to gain insights into both connected and prospective buildings/customers on a smaller scale, thereby facilitating the development and operation of efficient DH networks.

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% [8]. 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 [9]. This already nowadays available data from SHMs opens the door for new data-driven methods to transform both DH networks and the building stock.

1.1. State-of-the-art

Considering the need to gain more insight into the heat energy use of buildings and the vast amount of data collected by SHMs, clustering is an essential step to represent the available information in human understandable representative groups. Clustering of SEM (like) data has been investigated for several purposes: to identify typical consumption patterns [10–15] and relate them to building and occupant characteristics [16,17], to analyse peak consumption, to evaluate the potential of peak load shifting [18], to identify abnormal operation [19].

Ma et al. [10] applied Partitioning Around Medoids clustering using a dissimilarity measure based on the Pearson correlation coefficient to cluster daily profiles from three years of hourly Space Heating (SH) energy usage data from 19 educational buildings in Norway. As they treated each daily profile individually, one building had profiles from multiple clusters. Similarly, Gianniou et al. [17] employed k-means clustering with KSC-distance to derive daily clusters from hourly data of approximately 8300 single-family households in Denmark, spanning varying durations (up to 81 months). They too considered each day independently, but analysed the temporal distribution of individual daily clusters, observing the anticipated seasonal variations in consumption densities alongside some deviations. Lumbreras et al. [20] utilised two clustering methodologies: DBSCAN to identify energy usage outliers based on the relationship with the outdoor temperature, followed by k-means with Euclidean distance to categorise daily profiles. They applied this approach to one year of hourly SEM data from an apartment building in Estonia. Their analysis also revealed that daily clusters predominantly varied in accordance with seasonal patterns.

Calikus et al. [19] implemented k-shape clustering of *z-score* normalised hourly SHM data from 1222 non-single-family buildings in Sweden to identify demand clusters and unusual consumers (demand profiles divergent from any recognised cluster), using four predetermined seasons based on calendar dates. Similarly utilising four fixed calendar date-based seasons, Johra et al. [15] utilised k-means clustering with Euclidean distance to derive daily profiles for approximately 1000 buildings in Denmark. Notably, they not only derived profiles for energy use but also for return temperature and the difference between supply and return temperature. Utilising three predefined seasons, Yang et al. [18] employed hierarchical agglomerative clustering with Ward's minimum variance method and Euclidean distance within their framework to assess peak load shifting potential, applied to *z-score* normalised hourly data from 61 municipality buildings in Denmark.

Lu et al. [14] employed Gaussian Mixture Model (GMM) based clustering as the initial step within a framework to predict heat load patterns, with the aim to the identify temperature and occupancy-related patterns within the data. They demonstrated their approach using four months (equating to one heating season) of hourly data collected from six commercial buildings in China. Tureczek et al. [12] investigated various data transformation methods, feature extractions, and wavelet transformations to incorporate autocorrelation information when employing k-means clustering. They utilised one month of hourly data from 49 DH network substations, illustrating that autocorrelation features resulted in more distinguishable clusters, with both autocorrelation features and wavelet transformation notably reducing the upper bound runtime. Wang et al. [11] utilised GMM clustering to group nearly one year of hourly data from 480 buildings in Sweden with diverse usage patterns, leveraging three distinct feature types. Each feature type enabled the identification of users or patterns with different focal points, such as users contributing to the DH network peak load, long-term variations, and usage patterns independent of outdoor temperature.

Le Ray and Pinson [13], acknowledging the continuous stream-like nature of SHM data, proposed an adaptive clustering algorithm based on online k-means clustering utilising a dynamic time warping-based distance. They demonstrated this methodology using one month of hourly SHM data from 97 buildings in Denmark. Seasonality was

Table 1
Overview of existing studies clustering SHM (like) data, and their approach in handling seasonal variation.

Clustering algorithm	Buildings	Handling of seasonality	Ref.
Partitioning Around Medoids	19 educational	Daily profile were treated individually	[10]
k-means	≈8300 single-family	Daily profile were treated individually and temporal distribution was analysed	[17]
k-means	1 apartment building	Daily profile were treated individually and temporal distribution was analysed	[20]
k-shape	≈1200 non-single-family	Four date based seasons	[19]
k-means	≈1000 single-family	Four date based seasons	[15]
hierarchical agglomerative	61 municipalities	Three date based seasons	[18]
GMM	6 commercial	Not considered	[14]
k-means	6 commercial	Not considered	[12]
k-means	480 diverse use	Not considered	[11]
based on k-means online	97 buildings	Smoothing coefficient controlling the transfer of information from one time step to the next	[13]
k-means	139 residential	Two clustering steps, first clustering season than buildings	[16]

addressed through a smoothing coefficient controlling the transfer of information from one time step to the next, with the assumption that similar usage patterns grouped in clusters would exhibit consistent dynamics over time. Noteworthy is the research conducted by do Carmo and Christensen [16], aiming to achieve independence from fixed seasons. They initially clustered the daily profiles of each building into three groups (high, medium, low), representing different seasonal characteristics, before clustering these groups across all buildings. Employing this approach with k-means clustering on the SH and Domestic Hot Water (DHW) energy usage of 139 Danish dwellings equipped with heat pumps, they revealed that the seasonal groups did not precisely align with expected seasonal patterns.

From the existing works, an overview is presented in Table 1, it can be concluded that the clustering of SHM data to establish daily profiles has been extensively studied; however, certain limitations persist. With the exception of do Carmo and Christensen [16], existing approaches either do not account for seasonal variations, rely on fixed season definitions derived from calendar dates, or treat daily profiles separately. These limitations make the interpretation of results on a large scale challenging and overlook the seasonal variation as a feature in the clustering process. Furthermore, as traditional seasonal patterns shift and dissolve in the face of climate change [21–23], the manual, predefined definition of seasons becomes more difficult and less reliable, emphasising the necessity of considering seasonal effects without relying on predefined definitions.

Secondly, the utilisation of clusters to gain insights into the heat energy use of the building stock requires the understanding of the differentiating factors. In this context, the aforementioned study by Gianniou et al. [17] incorporated two Building Characteristics (BCs): building area and age, along with the number of registered adults, the number of teenagers, and the number children. These factors were analysed using logit regression models constructed for each of the five daily energy use clusters, assessing the belonging of a building to a particular cluster. The findings indicated the significance of building area, age, and the number of teenagers. Moreover, it was suggested that additional building characteristics, particularly socioeconomic factors such as income and job type, should be investigated further. Similarly, the previously mentioned work by do Carmo and Christensen [16] employed logistic regression to examine the influence of building and socioeconomic parameters on the two identified heat energy use clusters. Their analysis encompassed eleven BCs and four household characteristics, revealing the significance of BCs such as building area, age, and the SH distribution system for medium and low energy use seasons. However, it was noted that the limited availability of household and building characteristics constrained the generalisation of their analysis.

From this, it can be said that in-depth analyses of, at large-scale, available 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 [16,17]. 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.

Finally, it is anticipated that inspiration can also be drawn from the well-established field of SEM research, as summarised in Wang et al. [24], and more specifically for electricity use profiles in Kang et al. [25]. Nonetheless, differences exist between SEM and SHM data. The most notable distinctions include the higher reporting frequency of SEMs, typically 15 min or less for most EU countries [26], compared to the commonly reported 1 h intervals for SHMs. Additionally, SHM data are frequently rounded down to integer kWh values [27,28], meaning that, for example, any value between 1.0 kWh and $1.\overline{9}$ kWh is transmitted as 1.0 kWh. For characteristics influencing daily profiles it is expected that the different driving factors (assuming electricity is solely used for appliances) reduce the applicability of SEM data research to SHM data.

1.2. Contribution

This work establishes a novel workflow to overcome the in Section 1.1 highlighted limitations in the area of clustering SHM data and understanding the differentiating factors for these clusters. First, 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 and BCs describing a building at a high level of detail. Thereby, the aim is to identify the most important/useful BCs and secondly to understand the difference in buildings between clusters based on the identified BCs. The 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 hourly data from 4798 SHMs installed in single-family houses in Aalborg Municipality, Denmark. This analysis aims to gain more knowledge of the demand side in the DH networks and connected buildings. The derived significant BCs can guide stakeholders such as DH utility companies or public entities in collecting such data if such information is yet unavailable. The contributions can be summarised as:

- Clustering of smart heat meter data 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.

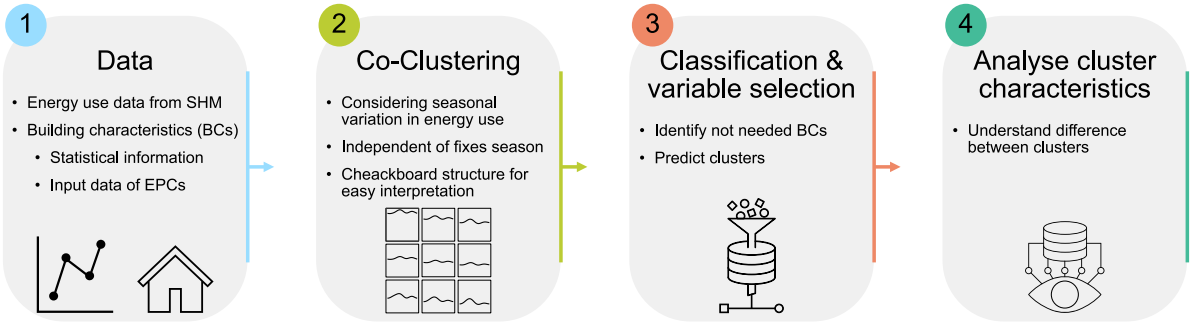


Fig. 1. Overview of the proposed method.

The paper is organised as follows: Section 2 describes each step of the proposed method in detail, before in Section 3 the used SHM data and BCs are outlined. In Section 4, the results of the case study are presented before the results are discussed, and conclusions are drawn in Section 5.

2. Method

For better clarity, each step is 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 Fig. 1.

2.1. Energy use clustering

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 [29] and Aghabozorgi et al. [30], and for functional data in Zhang and Parnell [31]. As mentioned in Section 1.1, most studies 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 and only the work by do Carmo and Christensen [16] has considered it by performing two consecutive clustering steps. In the domain of clustering SEM data Bouveyron et al. [32] 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 data from 1481 households in France. Following the same principle idea, Divina et al. [33] 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.

2.1.1. Used co-clustering method

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 [16] 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. [32] and implemented in the R [34] package `FunLBM` [35] is used for this work. It was chosen over the one by Divina et al. [33] due to the available implementation. 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 [36] 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. [32]. To select the most suitable model, so the optimal number of both customer and time cluster, the *Integrated information Likelihood Criterion* (ICL) (also referred to as *Integrated Completed Likelihood*, or *Integrated Classification Likelihood*) was used, whereby the highest ICL value indicates the most suitable model [32]. 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. [32], and as recommended by Ramsay and Silverman [37], 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)* [37] 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.

2.2. Variable selection and classification

The aim of this second step is to understand why buildings fall into their respective energy use cluster and to classify yet not to the DH network connected buildings based on their BCs into energy clusters. Based on the current research (Section 1.1), two key research gaps were 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 [38] 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.

2.2.1. Used classification and variable selection methods

The first approach used to address both aims is a Multinomial Logistic Regression fitted using group least absolute shrinkage and selection operator (Group Lasso) [39] (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 [40–42]. The benefit of group lasso [43] 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 [44], 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 (CV) 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 [45,46]. The MCC can be interpreted analogue to the Pearson correlation coefficient, ranging from -1 to 1, with 1 being perfect agreement and -1 total disagreement. BCs are scaled as recommended by Gelman [47], 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 [48] was used.

The second used approach is variable selection based on Random Forests (RFs), proposed by Genuer et al. [49] and implemented in the R package `VSURF` (Variable Selection Using Random Forests) [50,51] (this name will be also used in the remainder of this paper to refer to this method). VSURF is a two step procedure which is subsequently simplified outlined (for a detailed explanation, the interested reader is referred to the aforementioned references):

1. *threshold*: In this step variables are ranked based on their permutation-based importance and variables with small importance are excluded.
2. The second step consists of two sub-steps:
    - *interpretation*: a nested collection of RF models 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.
    - *prediction*: The, by importance sorted, variables (based on the *interpretation* step) are sequentially added under the restriction that the variable must decreases the OOB error significantly. The variables of the last model, i.e, the model, where introducing the next variable does not decrease the OOB error significantly.

VSURF was shown to identify a smaller subset of important variables compared to lasso [52] and to be the generally best performing RF variable selection technique for classification [53]. To reduce the risk of bias, VSURF is used within outer five-fold CV. 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 RF model is constructed and its hyperparameters are tuned using the automated tuning strategy tuneRanger [54] with MCC as the optimisation criterion and within an outer five-fold CV. The term VSURF + optimised RF is used to refer to this approach from here on.

2.3. 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 [55] could be used.

For VSURF + optimised RF, respectively the optimised 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 [56–58]. As the RF in its used implementation cannot 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` [59] in combination with the dedicated visualisation package `rpart.plot` [60] were used.

3. Data description

The data used in this work is a subset of the extensive dataset of smart meter data, and BCs described in and published by Schaffer et al. [28]. This dataset consists of processed hourly data from about 35 000 SHMs and 11 000 smart water meters installed mainly in residential buildings in Aalborg Municipality, Denmark, with varying lengths and, where available, accompanying BCs. This dataset has been used recently by Schaffer et al. [61] who developed a method to disaggregate to the total energy into SH and DHW. For this work, only SHM data of single-family houses for which data are available for 2020 and 2021 were selected from this dataset. Further, only buildings/SHMs where all accompanying BCs are available were considered. Based on this, SHM data of 4798 single-family houses with a total of about 8.418×10^7 hours were selected. It is to be noted that the BCs originating from the Energy Performance Certificates (EPCs) reports were recomputed using the in Schaffer et al. [28] 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. [28].

For this work, only the energy use data from SHM is considered. The energy use data is the hourly aggregated energy use for SH and DHW. The energy use data is transmitted as cumulative kilowatt-hour values, which are rounded down to the next integer, which is the common resolution for such data [28]. To mitigate this problem not the original data but the energy use data processed by the by Schaffer et al. [62] 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. Fig. 2 shows an overview of the daily energy use of all 4798 buildings, grouped per week.

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) [63], 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

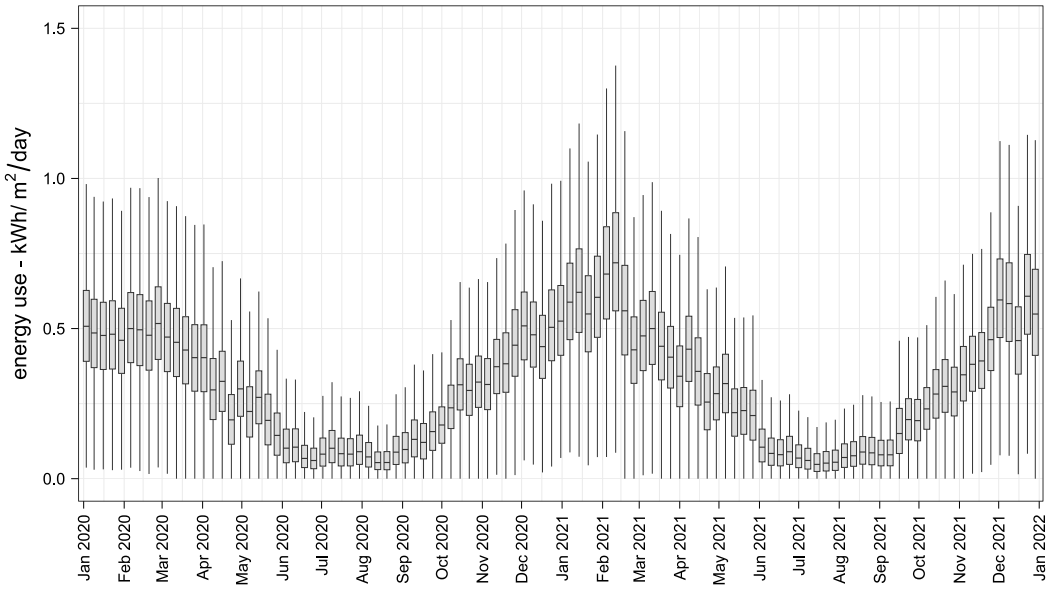


Fig. 2. Distribution of daily energy use of all 4798 buildings, grouped per week. To ease the visualisation data outside the range $Q1 - 1.5 \times IQR$ to $Q3 + 1.5 \times IQR$, where $Q1$ and $Q3$ are firsts and third quartile and IQR is the interquartile range are not shown.

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 EPCs. This data is not fully 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. [28], 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

3.1. Building characteristic data treatment

In total, 86 BCs are available in the dataset developed by Schaffer et al. [28]. 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 2. The Pearson correlation coefficient matrix of these BCs is shown in Fig. A.17 in Appendix A. 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.

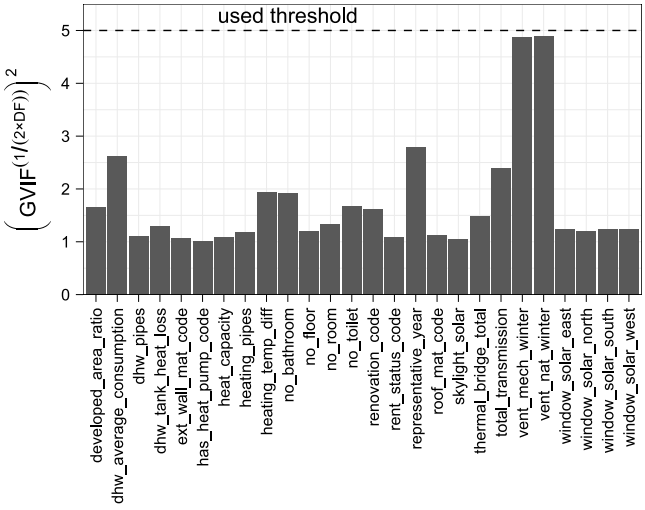


Fig. 3. For the degrees of freedom of the coefficients adjusted GVIF of all BCs.

3.1.1. Analysis of multicollinearity of building characteristic

As mentioned above in Section 3.1, some of the used BCs are partially correlated (Fig. A.17). To further analyse this and identify possible multicollinearity, the Generalised Variance Inflation Factor (GVIF) [64], an extension of the Variance Inflation Factor (VIF) for categorical variables was used. Further, as suggested by Fox and Monette [64], 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: $\left(GVIF^{(1/(2\times DF))}\right)^2$. Fig. 3 shows the result of this analysis considering all BCs. Given that no value exceeds five a commonly used rule of thumb, each BC is assessed to be not multicollinear with the remaining BCs.

4. Results

4.1. 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 the GCV.

Table 2
Used BCs based on the database described in Schaffer et al. [28]. 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. [28] 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*
		no_bathroom	Number of bathrooms
		no_floor	Number of floors
		no_room	Number of rooms
		no_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

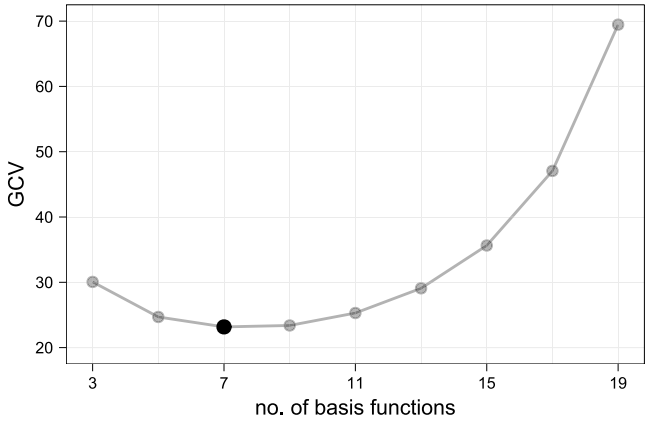


Fig. 4. Generalised cross-validation for different number of Fourier basis functions.

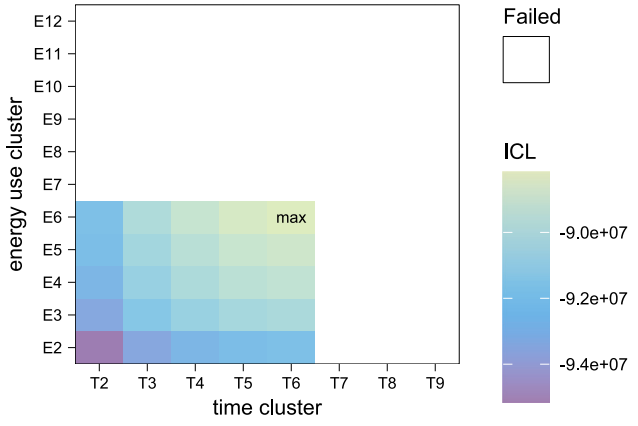


Fig. 5. ICL of the different combinations of energy use and time clusters. Failed combinations are due to at least one empty cluster.

The results (Fig. 4) show that seven basis functions give the lowest GCV with an apparent 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 cannot be used as the algorithm requires at least two clusters. For each combination, convergence, defined as the change of the loglikelihood between the current and the current minus ten iterations of smaller than 1×10^{-5}, was ensured. As shown in Fig. 5, 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 Fig. 6a, 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 weather stations for the two years (Fig. 6b), 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

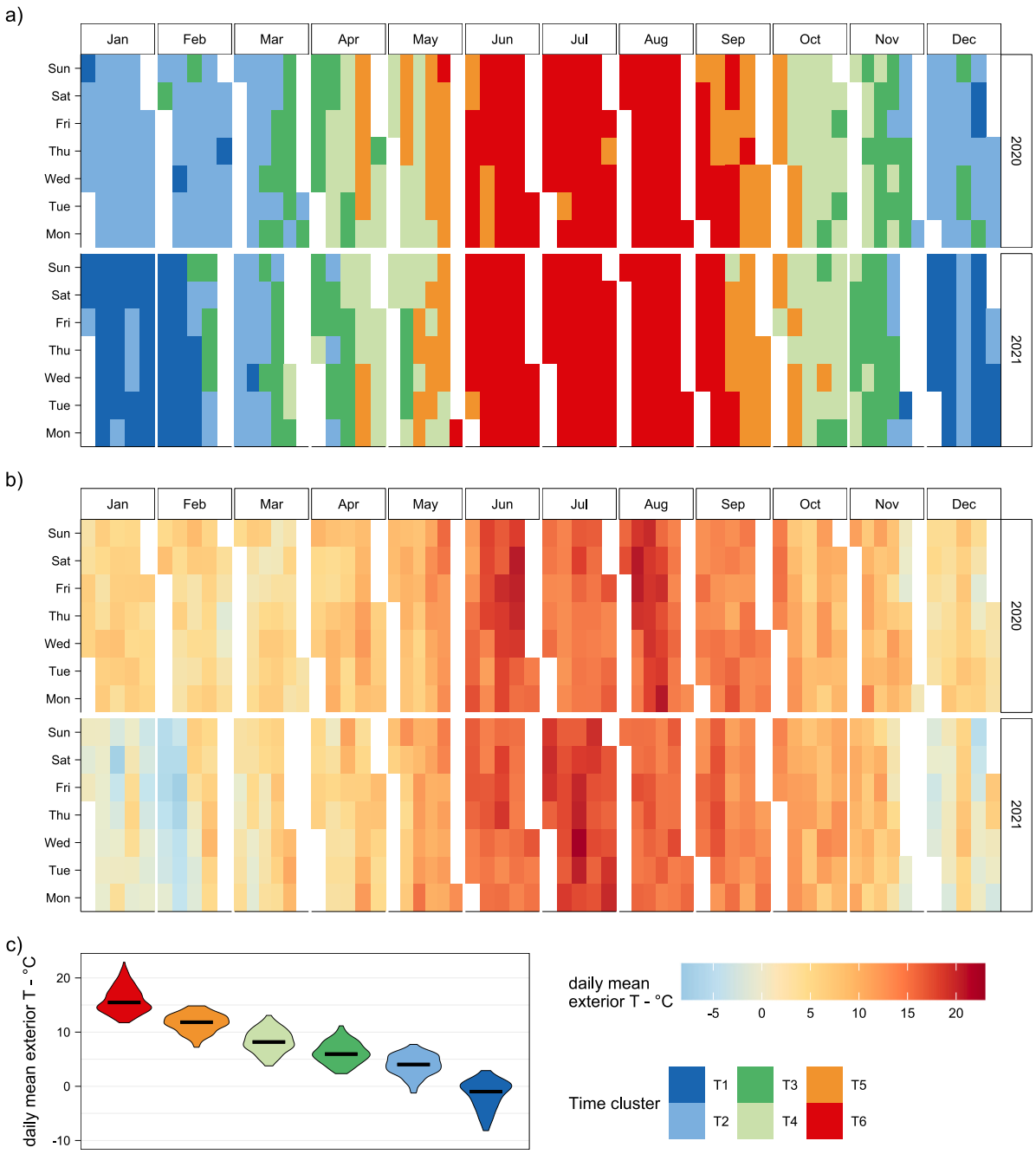


Fig. 6. (a) Distribution of time clusters for the selected two years. (b) Daily mean exterior temperature at the SHM location (c) daily mean exterior temperature per time cluster.

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 (Fig. 6c). 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 median daily external temperature is small, 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 [65], 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.

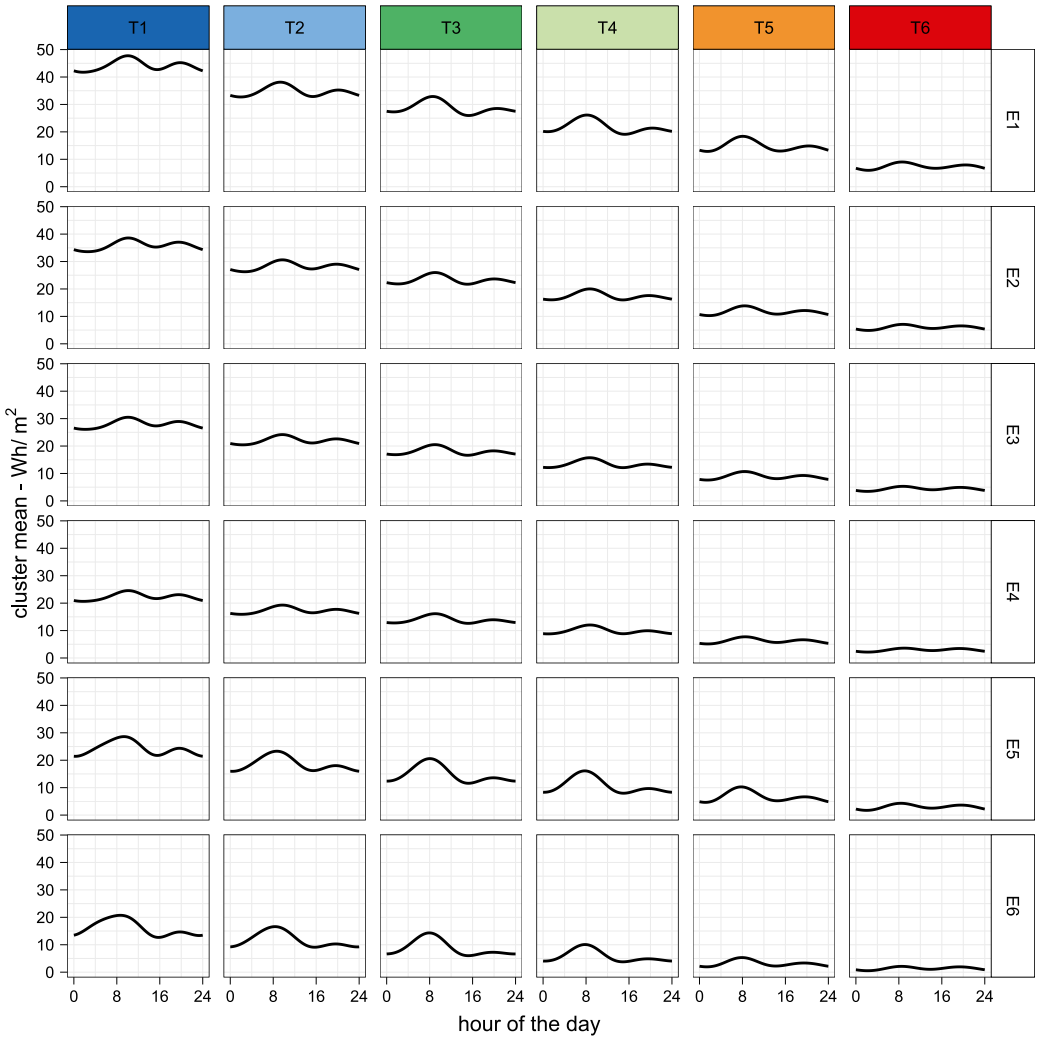


Fig. 7. Average energy use curves as estimated by FunLBM.

Fig. 7 shows the average energy use curves as estimated by FunLBM for 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 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 (Fig. 8), 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 [66–68]. 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 (Fig. 8), 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.

4.2. 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.

4.2.1. Distribution analysis of building characteristic

Fig. 9 shows the distribution of three selected BCs. The distributions of the remaining BCs are provided in the supplementary material. For the representative year (Fig. 9a), 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 (Fig. 7) 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

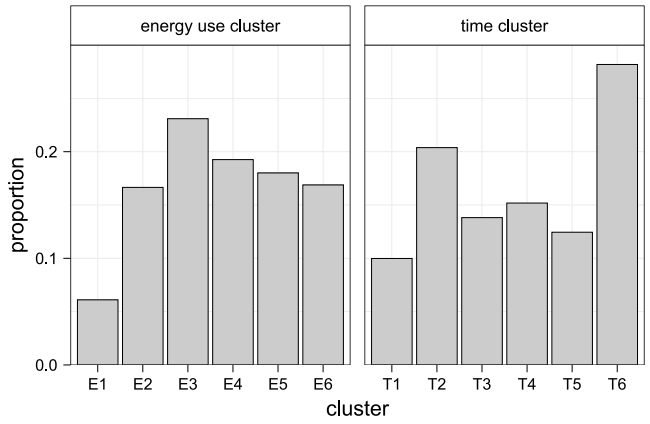


Fig. 8. Cluster proportions for energy use and time clusters.

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 second shown BC, the total transmission losses (Fig. 9b), 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. For the number of bathrooms (Fig. 9c), which were normalised for the cluster size, it is to be highlighted that a few buildings have zero bathrooms. It is assumed that this is due to incorrect data, but this cannot be determined with certainty, highlighting the also in Schaffer et al. [28] 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.

4.2.2. 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 3) 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 of the CV (Fig. B.18 in Appendix B) 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.

A normalised confusion matrix (Fig. 10) 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 none of the other clusters is predicted correctly. The other two BCs sets (EPC and combined) show

Table 3
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 no. of BCs
BBR	0.258	1.0
EPC	0.308	13.8
Combined	0.308	21.6

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 4.2.1), 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. Fig. 11 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 inner CV was tested. However, the results lead then to similar results as seen for the BBR BCs (Fig. 10), 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.

4.2.3. 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 4). 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.

Analysing the selected BCs of the *predictor* step in more detail (Fig. 12), 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,

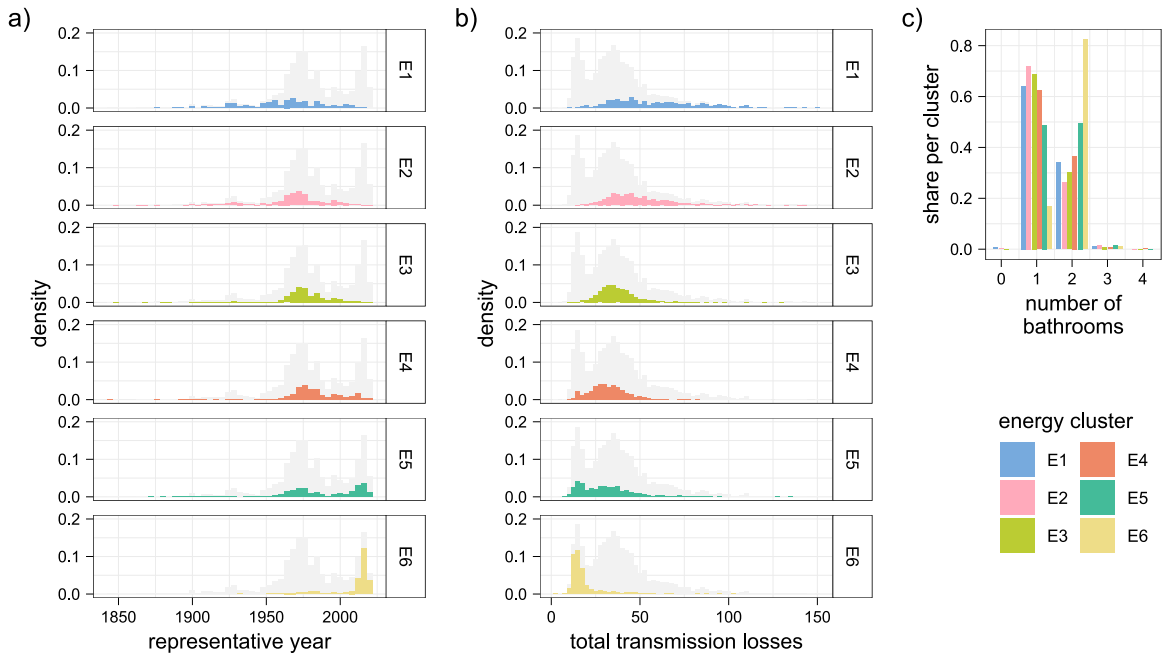


Fig. 9. Distribution of selected variables across the energy use clusters - (a) representative year, (b) transmission losses, (c) number of bathrooms (Table 2). The distribution of the number of bathrooms was normalised by the cluster size.

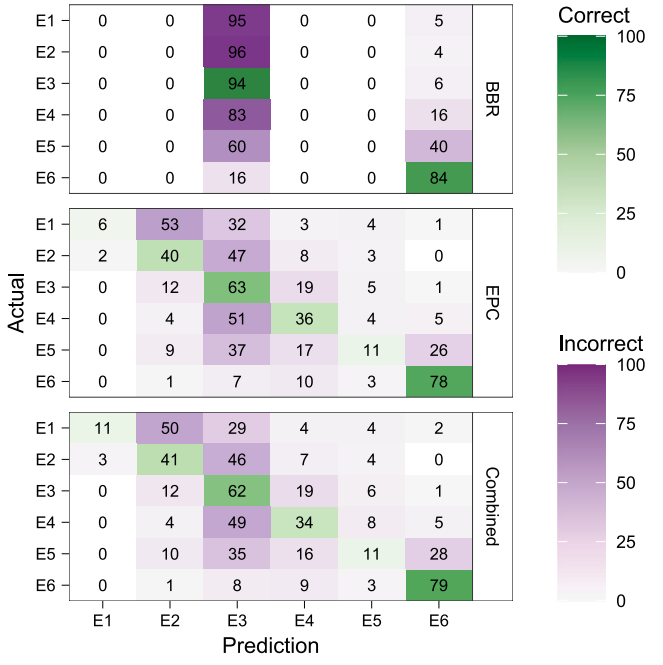


Fig. 10. Normalised confusion matrix averaged over the best model of MLRGL of each fold of the outer CV.

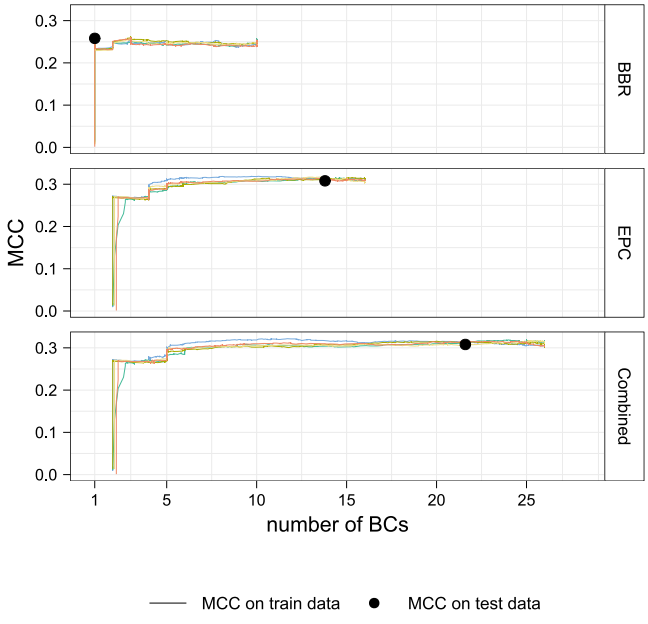


Fig. 11. MCC of MLRGL for each of the five outer CV folds as a function of the number of BCs with non-zero coefficient.

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

Table 4
Mean MCC and mean number of BC for VSURF for the sub-steps interpretation and prediction based on the outer CV loop.

	Dataset	Mean MCC	Mean no. 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

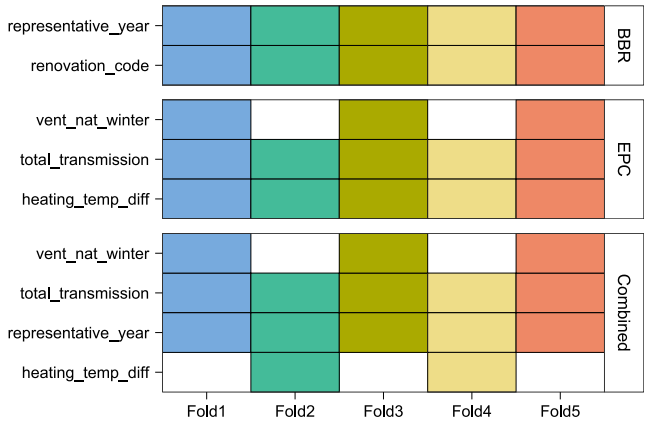


Fig. 12. By the *predictor* step of VSURF selected BCs.

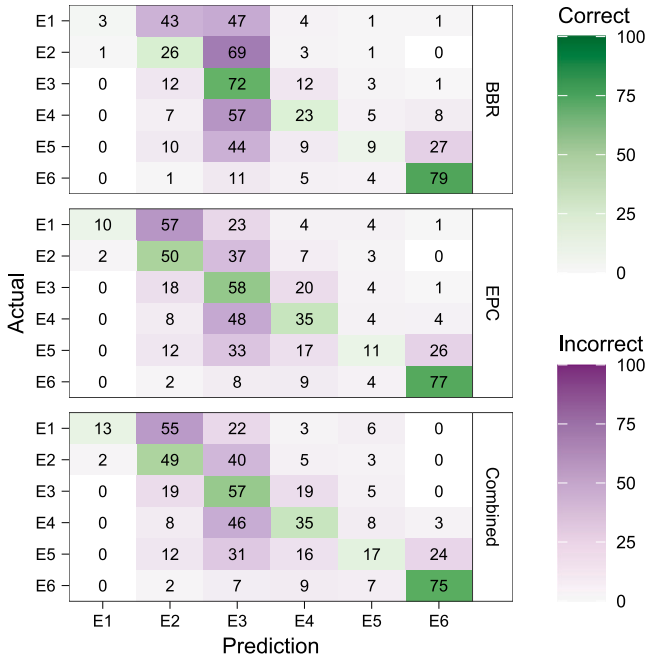


Fig. 13. Normalised confusion matrix averaged over the best model of the RF of each fold of the outer CV.

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).

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 (Fig. 13). Overall, 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 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.

Table 5
Mean MCC and mean number of BC for VSURF for the sub-steps interpretation and prediction based on the outer CV loop – based on the merged clusters.

	Dataset	Mean MCC	Mean no. 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

4.3. 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 5), 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 *interpretation* 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 (Fig. 14) 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 were used for the BCs from 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.

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 (Fig. 15). Form this it can be seen that now cluster E34 is the one most frequently

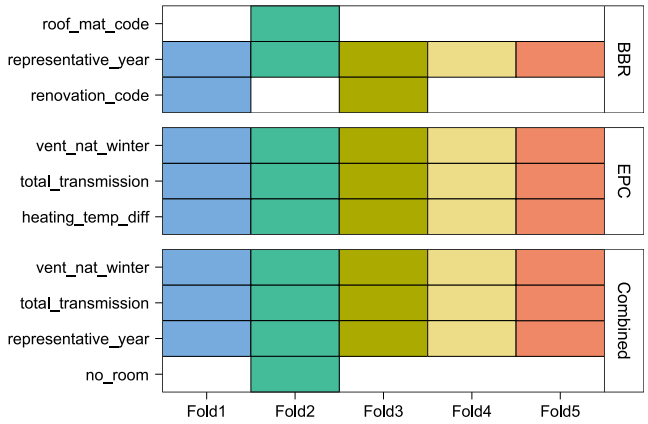


Fig. 14. By the predictor step of VSURF selected BCs based on the simplified clusters.

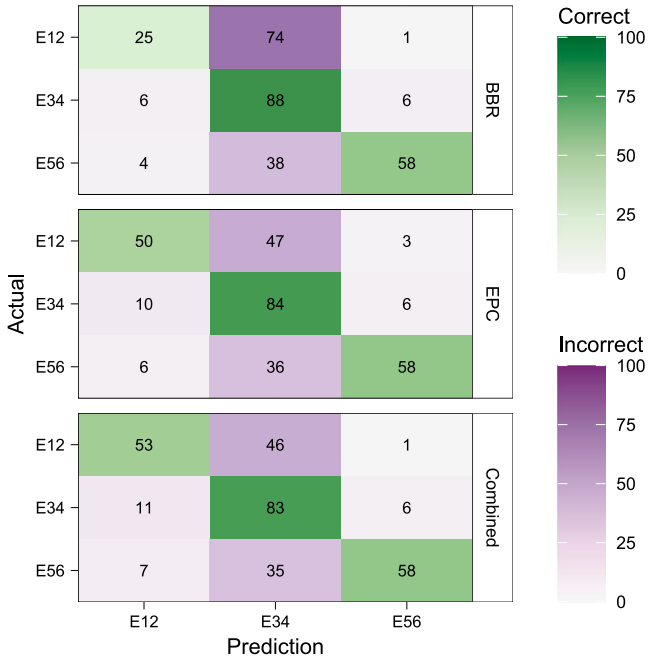


Fig. 15. Normalised confusion matrix averaged over the best model of the RF of each fold of the outer CV for the simplified clusters.

predicted correctly while E12 and E56 are predicted about equally often correctly. Thereby, both E12 and E56 are mainly mispredicted as E34. Overall the results show that this artificial simplification of the energy use clusters could be one option to increase the classification performance.

4.4. Visualisation 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 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

Fig. 16 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.

5. Discussion & conclusion

This work has established co-clusters of SHM data from 4798 single-family houses, utilising two years of data. The six identified time clusters highlight the limitations of using date-based season definitions, particularly when considering longer data periods. The optimisation of six time clusters, based on the ICL used for model/cluster selection, suggests a need to reconsider assumptions regarding the use of four or three seasons in previous research [15,18,19]. Additionally, results from the two-year period indicate considerable variations between the years, such as the warmer winter of 2020 compared to 2021 leading to different energy use intensities. This confirms the argument that with traditional seasonal patterns shifting and dissolving in the face of climate change [21–23] predefined seasons can incorrectly capture seasons.

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 SH and DHW, which is not readily available from commercial SHM data, are needed to confirm this. Overall, the clustering results showed that the chosen clustering method FunLBM leads to distinct energy use clusters and captures temporal variation well offering possible new insights compared to traditional clustering methods.

Limitations regarding the used co-clustering include that the use of basis expansions leads to smoothing of the data, which can result in inaccurate representations if the daily profiles are highly volatile. Additionally, it is suspected that the overall smoothing reduces the peaks caused by DHW, potentially altering the profile shapes. Furthermore, since the clustering is exhaustive, all data must be assigned to a cluster. As no detection of abnormal or atypical operation was performed, this means that buildings with unusual daily profiles are still assigned to clusters, potentially increasing noise in the data. It is expected that this issue can be mitigated in the future by applying methods capable of detecting such abnormal use patterns.

Finally, it is anticipated that the same clustering approach can provide insights into different research and application-oriented questions, either by using other available data from SHM meters, such as instantaneous supply and return temperature readings, or by preprocessing data differently, for example, by applying z-score normalisation to focus more on the shape rather than the magnitude of the profile.

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.

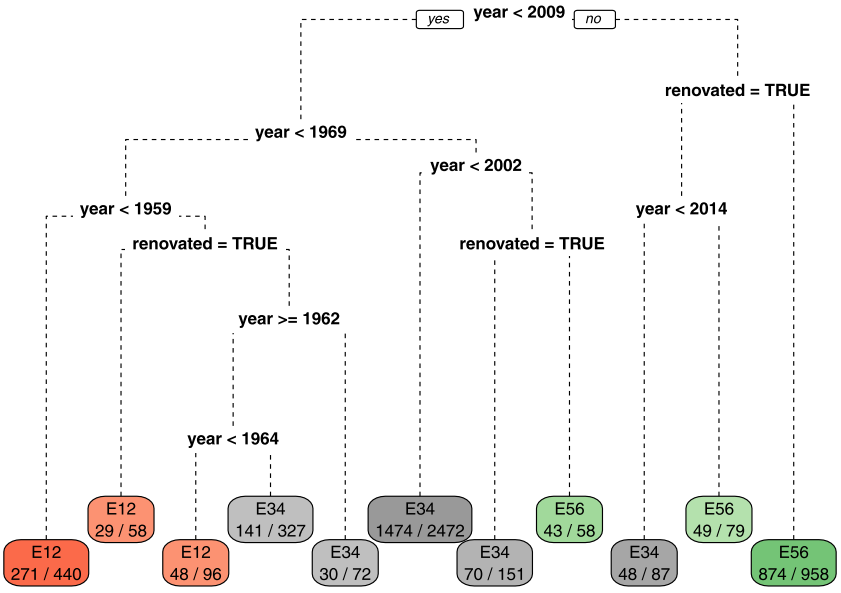


Fig. 16. Pruned decision tree for the simplified energy use clusters using only the representative year (year) and renovation code (renovated).

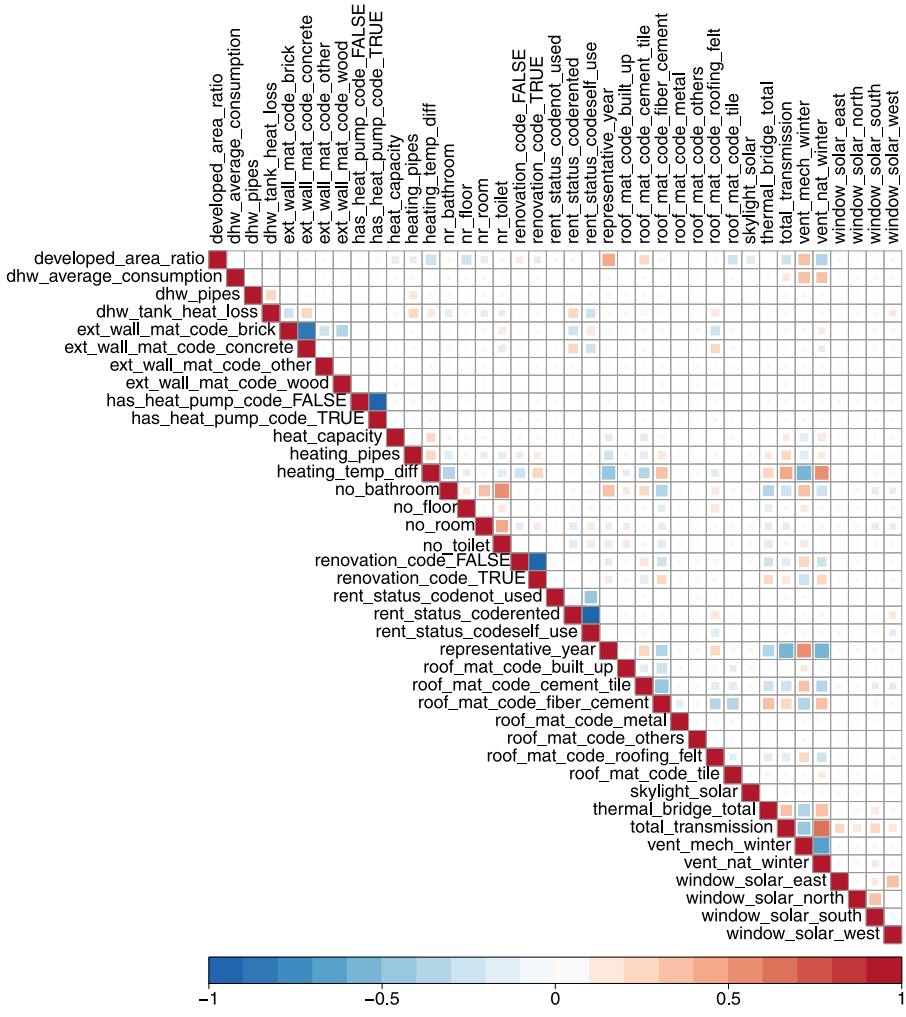


Fig. A.17. Pearson correlation coefficient of the combined BCs originating from BBR and EPC data.

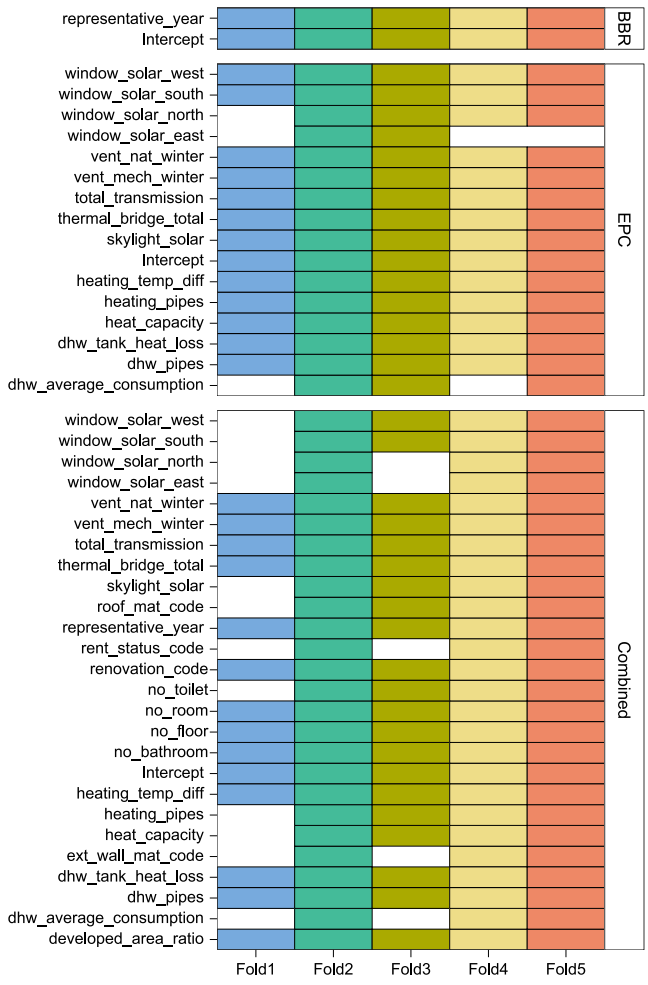


Fig. B.18. Detailed results of the variable selection of the best model of MLRGL of each fold of the outer CV including the intercept.

Consequently, they are also insufficient to understand why a building is in a particular energy use cluster with high certainty.

The approach to merge energy clusters based on domain knowledge aimed to increase the understanding of the driving differences between the clusters. This showed improvements in the MCC (0.44 to 0.50). However, overall, the same building characteristics (BCs) were selected for the non-simplified clusters. This confirms that using only the representative year and information about whether a building was renovated leads to nearly the same MCC as detailed information about a building. These results affirm those from the non-simplified clusters, more, or respectively different, information is needed to understand why a building is in a specific energy use cluster. Further, it can be concluded that overall 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

In terms of limitations, it must be considered that the sample of buildings used in this study is highly homogeneous, consisting only of single-family houses from one city in Denmark, known for its high income parity. Therefore, the obtained results may not be generalisable, and other BCs could become more important if different samples were considered. Additionally, the uncertainty in the used BCs, as highlighted by Schaffer et al. [28], could introduce noise, potentially reducing the performance of the classification methods used. Overall further research is needed to determine the additional information required to understand the determining factors for the energy clusters. However, based on previous research [69–71], it is hypothesised that occupant practices, such as heating system usage, daily routines, heating setpoints, and possible building faults, could have a significant influence.

Overall 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 with okay performance if the clusters were simplified (MCC $\approx$0.5). Furthermore, the approach does not rely on expert knowledge and is thus expected to be well suited for stakeholders commonly involved in the DH network. Additionally, for policymakers, the proposed workflow offers the opportunity to gain insights into the building stock, potentially supporting tailored policies to increase the decarbonisation of the building stock. For the research community, the results emphasise the importance of reconsidering traditional season definitions. Moreover, they highlight that, beyond available detailed building characteristics, understanding cluster membership is limited, indicating the need for future research to explore this further.

All used code is available at: https://github.com/markus-schaffer/co-clustering.

CRediT authorship contribution statement

Markus Schaffer: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **J. Eduardo Vera-Valdés:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Anna Marszal-Pomianowska:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this work has been published in Data in Brief - see [28] in the manuscript.

Acknowledgements

This work was funded by the Independent Research Fund Denmark under FOREFRONT project (0217-00340B). The authors would like to acknowledge the IEA EBC Annex84 “Demand management of buildings in thermal networks” project funded by The Energy Technology Development and Demonstration Programme – EUDP (Case no. 64020-2080) for support in the dissemination of the results.

Appendix A. Building characteristic correlation

See Fig. A.17.

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

See Fig. B.18.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.apenergy.2024.123586.

References

[1] Republic of Austria. Bundesgesetz zum Ausstieg aus der fossil betriebenen Wärmebereitstellung (Erneuerbare). 2022, p. 11, URL: https://www.parlament.gv.at/PAKT/VHG/XXVII/ME/ME_00212/fname_1451879.pdf.

[2] Maach ML. Bred aftale i Folketinget: Fra 2035 skal ingen boliger opvarmes af gas. In: DR. Copenhagen; 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] Lund H, Werner S, Wiltshire R, Svendsen S, Thorsen JE, Hvelplund F, et al. 4Th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems. Energy 2014;68:1–11. http://dx.doi.org/10.1016/j.energy.2014.02.089.

[5] Averfalk H, Benakopoulos T, Best I, Dammel F, Engel C, Geyer R, et al. Low-temperature district heating implementation guidebook: Final report of IEA DHC Annex TS2. Implementation of low-temperature district heating systems. Fraunhofer IRB Verlag; 2021, http://dx.doi.org/10.24406/publica-fhg-301176.

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

[7] European Commission. EU buildings factsheets | Energy. 2022, URL: https://ec.europa.eu/energy/eu-buildings-factsheets_en.

[8] 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.

[9] European Parliament. Directive (EU) 2018/2002 amending directive 2012/27/EU on energy efficiency. Official J Eur Union 2018;(L 328/210). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018L2002&from=EN.

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

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

[12] Tureczek AM, Nielsen PS, Madsen H, Brun A. Clustering district heat exchange stations using smart meter consumption data. Energy Build 2019;182:144–58. http://dx.doi.org/10.1016/j.enbuild.2018.10.009, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778818314725.

[13] Le Ray G, Pinson P. Online adaptive clustering algorithm for load profiling. Sustain Energy Grids Netw 2019;17:100181. http://dx.doi.org/10.1016/j.segan.2018.100181, URL: https://linkinghub.elsevier.com/retrieve/pii/S2352467718301498.

[14] Lu Y, Tian Z, Peng P, Niu J, Li W, Zhang H. GMM clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system. Energy Build 2019;190:49–60. http://dx.doi.org/10.1016/j.enbuild.2019.02.014.

[15] Johra H, Leiria D, Heiselberg P, Marszal-Pomianowska A, Tvedebrink T. Treatment and analysis of smart energy meter data from a cluster of buildings connected to district heating: A Danish case. In: Kurnitski J, Kalamees T, editors. In: E3S Web of Conferences, vol. 172, EDP Sciences; 2020, p. 12004. http://dx.doi.org/10.1051/e3sconf/202017212004, https://www.e3s-conferences.org/10.1051/e3sconf/202017212004. 12th Nordic Symposium on Building Physics, NSB 2020.

[16] do Carmo CMR, Christensen TH. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy Build 2016;125:171–80. http://dx.doi.org/10.1016/j.enbuild.2016.04.079, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778816303565.

[17] Gianniou P, Liu X, Heller A, Nielsen PS, Rode C. Clustering-based analysis for residential district heating data. Energy Convers Manage 2018;165(December 2017):840–50. http://dx.doi.org/10.1016/j.enconman.2018.03.015, URL: https://linkinghub.elsevier.com/retrieve/pii/S019689041830236X.

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

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

[20] Lumbreras M, Martin-Escudero K, Diarce G, Garay-Martinez R, Mulero R. Unsupervised clustering for pattern recognition of heating energy demand in buildings connected to district-heating network. In: 2021 6th international conference on smart and sustainable technologies (spliTech). IEEE; 2021, p. 1–5. http://dx.doi.org/10.23919/SpliTech52315.2021.9566420, URL: https://ieeexplore.ieee.org/document/9566420/.

[21] Pörtner H, Roberts D, Poloczanska E, Mintenbeck K, Tignor M, Alegría A, et al. IPCC sixth assessment report. In: Cambridge university press. Technical report, Switzerland: IPCC Geneva; 2022, p. 3–33.

[22] EPA. Seasonality and climate change: A review of observed evidence in the United States. Technical report December, U.S. Environmental Protection Agency; 2021.

[23] 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.

[24] Wang Y, Chen Q, Hong T, Kang C. Review of smart meter data analytics: Applications, methodologies, and challenges. IEEE Trans Smart Grid 2019;10(3):3125–48. http://dx.doi.org/10.1109/TSG.2018.2818167, URL: https://ieeexplore.ieee.org/document/8322199/.

[25] Kang X, An J, Yan D. A systematic review of building electricity use profile models. Energy Build 2023;281:112753. http://dx.doi.org/10.1016/j.enbuild.2022.112753, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778822009240.

[26] Tounquet F, Alaton C. Benchmarking smart metering deployment in EU-28. European commission. (December):2020, p. 142. http://dx.doi.org/10.2833/492070, URL: https://op.europa.eu/en/publication-detail/-/publication/b397ef73-698f-11ea-b735-01aa75ed71a1/language-en.

[27] Schaffer M, Tvedebrink T, Marszal-Pomianowska A. Three years of hourly data from 3021 smart heat meters installed in danish residential buildings. Sci Data 2022;9(1):420. http://dx.doi.org/10.1038/s41597-022-01502-3, URL: https://www.nature.com/articles/s41597-022-01502-3.

[28] Schaffer M, Veit M, Marszal-Pomianowska A, Frandsen M, Pomianowski MZ, Dichmann E, et al. Dataset of smart heat and water meter data with accompanying building characteristics. Data Brief 2023;52:109964. http://dx.doi.org/10.1016/j.dib.2023.109964.

[29] Warren Liao T. Clustering of time series data - A survey. Pattern Recognit 2005;38(11):1857–74. http://dx.doi.org/10.1016/J.PATCOG.2005.01.025, URL: www.elsevier.com/locate/patcog.

[30] Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering – A decade review. Inf Syst 2015;53:16–38. http://dx.doi.org/10.1016/J.IS.2015.04.007.

[31] Zhang M, Parnell A. Review of clustering methods for functional data. ACM Trans Knowl Discov Data 2023;17(7):1–34. http://dx.doi.org/10.1145/3581789, URL: https://dl.acm.org/doi/10.1145/3581789.

[32] Bouveyron C, Bozzi L, Jacques J, Jollois FX. The functional latent block model for the co-clustering of electricity consumption curves. J R Stat Soc Ser C Appl Stat 2018;67(4):897–915. http://dx.doi.org/10.1111/rssc.12260, https://rss.onlinelibrary.wiley.com/doi/10.1111/rssc.12260.

[33] Divina F, Vela F, Torres M. Biclustering of smart building electric energy consumption data. Appl Sci 2019;9(2):222. http://dx.doi.org/10.3390/app9020222, http://www.mdpi.com/2076-3417/9/2/222.

[34] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022, URL: https://www.r-project.org/.

[35] Bouveyron C, Jacques J, Schmutz A. funLBM: Model-based co-clustering of functional data. 2022, URL: https://cran.r-project.org/package=funLBM.

[36] Govaert G, Nadif M. Co-clustering. Co-clustering: models, algorithms and applications, vol. 9781848214, Hoboken, USA: John Wiley & Sons, Inc.; 2013, p. 1–200. http://dx.doi.org/10.1002/9781118649480, URL: http://doi.wiley.com/10.1002/9781118649480.

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

[38] Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, vol. 2, Springer; 2009.

[39] Vincent M, Hansen NR. Sparse group lasso and high dimensional multinomial classification. Comput Statist Data Anal 2014;71:771–86. http://dx.doi.org/10.1016/j.csda.2013.06.004, URL: https://linkinghub.elsevier.com/retrieve/pii/S0167947313002168.

[40] Wang Q, Augenbroe G, Kim JH, Gu L. Meta-modeling of occupancy variables and analysis of their impact on energy outcomes of office buildings. Appl Energy 2016;174:166–80. http://dx.doi.org/10.1016/j.apenergy.2016.04.062.

[41] Gang W, Augenbroe G, Wang S, Fan C, Xiao F. An uncertainty-based design optimization method for district cooling systems. Energy 2016;102:516–27. http://dx.doi.org/10.1016/j.energy.2016.02.107.

[42] Sun Y, Gu L, Wu CF, Augenbroe G. Exploring HVAC system sizing under uncertainty. Energy Build 2014;81:243–52. http://dx.doi.org/10.1016/j.enbuild.2014.06.026.

[43] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 2006;68(1):49–67. http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x, https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00532.x.

[44] Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. J Comput Graph Statist 2013;22(2):231–45. http://dx.doi.org/10.1080/10618600.2012.681250, http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.681250.

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

[46] Chicco D, Jurman G. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21(1):1–13. http://dx.doi.org/10.1186/s12864-019-6413-7, https://link.springer.com/article/10.1186/s12864-019-6413-7.
[47] Gelman A. Scaling regression inputs by dividing by two standard deviations. Stat Med 2008;27(15):2865–73. http://dx.doi.org/10.1002/sim.3107, URL: www.interscience.wiley.com.
[48] Vincent M, Hansen NR. msgl: Multinomial sparse group lasso. 2019.
[49] Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett 2010;31(14):2225–36. http://dx.doi.org/10.1016/j.patrec.2010.03.014.
[50] Genuer R, Poggi J-M, Tuleau-Malot C. VSURF: Variable selection using random forests. 2022, URL: https://cran.r-project.org/package=VSURF.
[51] Genuer R, Poggi JM, Tuleau-Malot C. VSURF: An R package for variable selection using random forests. R J 2015;7(2):19–33. http://dx.doi.org/10.32614/rj-2015-018, URL: http://cran.r-project.org/package=VSURF.
[52] Virdi JS, Peng W, Sata A. Feature selection with LASSO and VSURF to model mechanical properties for investment casting. In: 2019 international conference on computational intelligence in data science. IEEE; 2019, p. 1–6. http://dx.doi.org/10.1109/ICCIDS.2019.8862141, URL: https://ieeexplore.ieee.org/document/8862141/.
[53] Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl 2019;134:93–101. http://dx.doi.org/10.1016/j.eswa.2019.05.028, URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417419303574.
[54] Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov 2019;9(3):1–15. http://dx.doi.org/10.1002/widm.1301.
[55] Zhang Z, Kattan MW. Drawing nomograms with R: applications to categorical outcome and survival data. Ann Transl Med 2017;5(10):211. http://dx.doi.org/10.21037/atm.2017.04.01, https://atm.amegroups.com/article/view/14736.
[56] Barlow T, Neville P. Case study: visualization for decision tree analysis in data mining. In: IEEE symposium on information visualization, 2001. IEEE; 2001, p. 149–52. http://dx.doi.org/10.1109/INFVIS.2001.963292, URL: http://ieeexplore.ieee.org/document/963292/.
[57] Van Den Elzen S, Van Wijk JJ. BaobabView: Interactive construction and analysis of decision trees. In: VAST 2011 - IEEE conference on visual analytics science and technology 2011, proceedings. IEEE; 2011, p. 151–60. http://dx.doi.org/10.1109/VAST.2011.6102453, URL: http://ieeexplore.ieee.org/document/6102453/.
[58] Parisot O, Didry Y, Bruneau P, Otjacques B. 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, p. 80–7. http://dx.doi.org/10.5220/0004740800800087, URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0004740800800087.
[59] Therneau T, Atkinson B. rpart: Recursive partitioning and regression trees. 2022, URL: https://cran.r-project.org/package=rpart.
[60] Milborrow S. rpart.plot: Plot 'rpart' models: An enhanced version of 'plot.rpart'. 2022, URL: https://cran.r-project.org/package=rpart.plot.
[61] Schaffer M, Widén J, Vera-Valdés JE, Marszal-Pomianowska A, Larsen TS. Disaggregation of total energy use into space heating and domestic hot water: A city-scale suited approach. Energy 2024;291:130351. http://dx.doi.org/10.1016/j.energy.2024.130351, URL: https://linkinghub.elsevier.com/retrieve/pii/S0360544224001221.
[62] Schaffer M, Leiria D, Vera-Valdés JE, Marszal-Pomianowska A. Increasing the accuracy of low-resolution commercial smart heat meter data and analysing its error. In: 2023 European conference on computing in construction. 2023, http://dx.doi.org/10.35490/EC3.2023.208, URL: https://ec-3.org/publications/conference/paper/?id=EC32023_208.
[63] Danish Property Assessment Agency. Bygnings- og boligregistret. 2023, URL: https://bbr.dk/om-bbr.
[64] Fox J, Monette G. Generalized collinearity diagnostics. J Amer Statist Assoc 1992;87(417):178–83. http://dx.doi.org/10.1080/01621459.1992.10475190, URL: http://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190.
[65] Guldbrandt Brønnum M, Skriver Steffensen J, Monin C, Henriksen C. Coronarestriktioner er fortid: Se tidslinjen over pandemien. 2022, URL: https://www.tv2nord.dk/coronavirus/coronarestriktioner-er-fortid-se-tidslinjen-over-pandemien.
[66] George D, Pearre NS, Swan LG. High resolution measured domestic hot water consumption of Canadian homes. Energy Build 2015;109(July):304–15. http://dx.doi.org/10.1016/j.enbuild.2015.09.067, URL: https://linkinghub.elsevier.com/retrieve/pii/S0378778815303066.
[67] Fuentes E, Arce L, Salom J. A review of domestic hot water consumption profiles for application in systems and buildings energy performance analysis. Renew Sustain Energy Rev 2018;81(May 2017):1530–47. http://dx.doi.org/10.1016/j.rser.2017.05.229, URL: https://linkinghub.elsevier.com/retrieve/pii/S1364032117308614.
[68] Meireles I, Sousa V, Bleys B, Poncelet B. Domestic hot water consumption pattern: Relation with total water consumption and air temperature. Renew Sustain Energy Rev 2022;157(March 2021):112035. http://dx.doi.org/10.1016/j.rser.2021.112035, URL: https://linkinghub.elsevier.com/retrieve/pii/S1364032121012971.
[69] Gadd H, Werner S. Fault detection in district heating substations. Appl Energy 2015;157:51–9. http://dx.doi.org/10.1016/J.APENERGY.2015.07.061.
[70] Larsen TS, Knudsen HN, Kanstrup AM, Christiansen ET, Gram-Hanssen K, Mosgaard M, et al. Occupants influence on the energy consumption of danish domestic buildings. 2010, p. 77.
[71] Andersen R. The influence of occupants' behaviour on energy consumption investigated in 290 identical dwellings and in 35 apartments. In: 10th international conference on healthy buildings 2012, vol. 3, (May):2012, p. 2279–80.