Analysing energy use clusters of single-family houses using building and socio-economic characteristics

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

Published in:
Journal of Physics: Conference Series (Online)

DOI (link to publication from Publisher):
10.1088/1742-6596/2600/5/052004

Publication date:
2023

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

Analysing energy use clusters of single-family houses using building and socio-economic characteristics

Analysing energy use clusters of single-family houses using building and socio-economic characteristics

M. Schaffer[1*], A. R. Hansen[1], J.E. Vera Valdés[2], A. Marszal-Pomianowska[1]

[1] Department of the Built Environment, Aalborg University, Aalborg, 9220, Denmark
[2] Department of Mathematical Sciences, Aalborg University, Aalborg, 9220, Denmark

* Corresponding author msch@build.aau.dk

Abstract. Clustering has been shown to be a promising approach to reduce the large amount of data from smart heat meters to representative profiles. However, attempts to understand why a case (building including its occupants) is within a particular cluster have only been moderately accurate. Therefore, this work uses existing energy use clusters based on about 4500 single-family homes to investigate whether socio-economic characteristics (SECs) alone or in combination with building characteristics (BCs) can improve the insight into the energy use clusters. An established variable selection and classification approach based on random forests was used. The results show that the eight SECs used alone provide poor insight into the energy use clusters, achieving only a Matthew Correlation Coefficient (MCC) of around 0.1. Simplifying the energy use clusters based on similarities, which was successful in the past, only moderately increased the MCC (≈ 0.17). When combined with BCs, SECs were never selected by the algorithm used, showing that they do not lead to a (significant) increase in MCC for both unsimplified and simplified clusters. Thus, this work suggests that SECs do not provide additional insights into why a case is within its respective energy use cluster.

1. Introduction

District heating (DH) has recently come to the forefront of public and political attention as a solution to reduce Europe's dependence on gas [1,2]. While DH can be a short to medium-term solution to achieve this, most current DH systems are still heavily dependent on non-renewable energy sources. As a result, they need to undergo significant changes to become 4th generation low-temperature DH networks that can facilitate 100% renewable energy sources [3]. Only such networks can support the European Union's (EU's) long-term climate goal of a fully decarbonised building stock by 2050 [4]. The 4th generation DH networks rely on low-energy buildings. Therefore, the building stock must undergo significant renovation to facilitate the necessary transformation of DH networks. For decades the renovation campaigns were unsuccessful and with a very low rate of 1% annually, leading to nearly 75% of the EU's building stock being still energy inefficient [3,4]. However, it is well known that the energy performance gap, i.e., the mismatch between actual and predicted energy use, is particularly large in low-energy buildings [5] and can, therefore, drastically reduce the efficiency of renovations. To mitigate this issue and execute successful renovations, one must tailor renovation measures to both the building and its occupants, which requires understanding why a building has the energy use pattern it does.

The ongoing digitalisation of the built environment offers new possibilities to gain the necessary insight at a large scale. Smart heat meters (SHMs) (remotely readable heat meters) installed in buildings

connected to DH commonly transmit the total energy use of buildings at a one-hour resolution, allowing for unprecedented information on the energy use in buildings.

Given the large amount of data collected by SHM, clustering proved to be a successful approach to derive a few representative typical energy use patterns that humans can easily understand [6–10]. Such patterns can be valuable for tailoring renovation measures and for DH network operators to better understand their customers' energy use, which is crucial for optimising network operation and implementing more advanced strategies such as load shifting.

In a recent work, Schaffer et al., 2023 [10] used a co-clustering approach to derive energy use patterns based on two years of hourly energy use data of 4798 single-family houses in Aalborg Municipality, Denmark. They identified six different energy-use clusters. Next to the clustering, they aimed to understand why a building is in its respective energy use clusters based on 26 building characteristics (BCs) available for each building, ranging from the construction year to detailed information about transmission losses, window, and ventilation characteristics. To facilitate this, they used classification and variable selection techniques. Their results showed that with BCs, despite the level of detail available, only a Matthew Correlation Coefficient (MCC) of around 0.3 could be achieved for the six energy clusters. Furthermore, they simplified the six energy use clusters to three based on similarities and found that this increases the MCC to about 0.5. Thus, their results showed that BCs, even in the used level of detail, are insufficient to fully explain why a building is within its respective energy use cluster.

This work extends their analysis by investigating if socio-economic characteristics (SECs) alone or in combination with BCs can provide the needed insight to improve the understanding of why a building is in its respective energy use cluster.

2. Methodology

The objective of this work is to analyse whether SECs alone or in combination with BCs can explain why a building is in its respective energy cluster. As not all available characteristics are expected to be important, a variable selection is first performed to identify and exclude redundant information, reduce variance (noise) and obtain the simplest possible model. Such a reduction to only 'useful' variables is particularly important in the context of BCs and SECs, where more required information may mean more costly data collection at the city or country level.

Based on the findings of Schaffer et al., 2023 [10], 'Variable Selection Using Random Forest' (VSURF) [11], implemented in the R package with the same name [12], is used in this work. VSURF is a two-step procedure, the first step, *threshold,* excludes unimportant variables based on the permutation-based importance score. The second step is divided into two sub-steps. In the first sub-step, *interpretation*, nested random forests (RFs) are constructed, starting from the one that includes only the most important variable to the one that includes all variables retained from the first step. The variables of the most accurate model (based on out-of-bag (OOB) error) are retained. In the second sub-step, *prediction*, the variables are added in order of importance, based on the *interpretation* step, starting with the most important until the error is no longer significantly reduced. The interested reader is referred to the mentioned references for a more detailed description.

To avoid bias, VSRURF is run in outer fivefold cross-validation. Within VSURF, the OOB error is used as a criterion (as it could not be easily changed), while the MCC is used on the test data, as it was shown to be superior to the accuracy, especially when the classes are unbalanced [13]. The MCC ranges from -1 to $+1$ and can be interpreted similarly to the well know Pearson correlation coefficient, with $+1$ being perfect agreement.

3. Data description and preparation

3.1. Description of energy use clusters and data treatment

This paper uses the energy-use clusters obtained by co-clustering in Schaffer et al., 2023 [10]. These clusters are based on two years of hourly total heat energy use (space heating (SH) and domestic hot

water (DHW)) normalised by the building area of 4798 single-family houses in Aalborg Municipality, Denmark. An overview of the clusters' daily load profiles can be seen in Figure 1. The columns refer to the different time clusters found by the co-clustering, which are identical for all buildings. The rows represent the different energy use clusters, i.e., the different clusters of buildings with similar energy use profiles across the identified time clusters. The interested reader is referred to the reference mentioned earlier for a more detailed explanation.

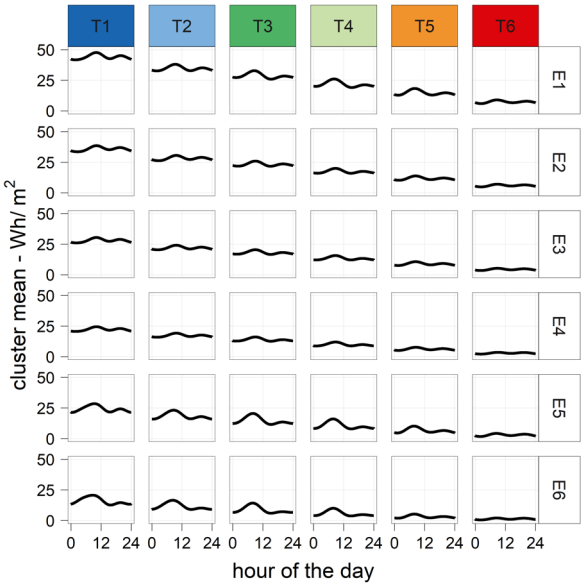

Figure 1. Mean energy use profiles of energy use clusters established by Schaffer et al., 2023 [10] using co-clustering. The columns refer to the different time clusters (T1 - T6) which are identical for all buildings; the rows are the energy use clusters (E1 – E6), i.e., the different clusters of buildings with similar energy use profiles across all time clusters found.

3.2. Description and analysis of building and social-economic characteristics

For each of the houses used in the clustering, 26 BCs originating from two sources, the Danish Building and Dwelling Register (BBR) (10BCs) and the Energy Performance Certificates (EPCs) reports (16 BCs), were retrieved. These BCs are identical to the ones used in Schaffer et al., 2023 [10]. In the remainder of the paper, the terms *BBR* and *EPC* are used to refer to respective subsets. A more detailed description is omitted for space reasons, and the reader is referred to Schaffer et al., 2023 [10]. However, it should be emphasised that all BCs used are subject to uncertainty. For this reason, a rule-based quality control framework for data derived from EPCs has also been implemented in Schaffer et al., 2023 [14], which established the dataset on which these data are based, and although this framework reduces the uncertainty, it cannot completely eliminate it.

The desired nine SECs could only be completely retrieved for 4694 (-103) buildings from Statistic Denmark. Consequently, only these 4694 buildings were considered for the analysis for the remainder of this paper. Given that this is only a reduction of about 2 % and that the reduction is relatively evenly distributed between the energy clusters (0.88% to 3.46%), this decrease is expected to have no significant effect on the validity of the established energy clusters. The obtained nine SECs are listed and described in Table 1. Based on an analysis of the Pearson correlation coefficient to identify possible collinearity, the SEC 'live alone' was removed as it was highly colinear with the number of adults and children. As the next step, the generalised variable inflation factor (GVIF) [15] as defined in Equation 1, where DF are the degrees of freedom of each variable, was calculated for all SECs and BCs combined to identify possible multicollinearity.

$$\left(GVIF^{(1/(2\times DF))}\right)^2 \qquad \mathbf{1}$$

This analysis showed that the GVIF is, for all characteristics below 5, a common rule of thumb for multicollinearity; thus, the SECs contain new information that BCs cannot explain. Consequently, no additional characteristic was removed, and the following characteristics sets were analysed separately:

- SECs: 8 SECs
- BBR + SECs: 10 BCs + 8 SECs
- EPC + SECs: 16 BCs + 8 SECs
- Combined (BBR + EPC) + SECs: 26BCs + 8 SECs

Table 1. Overview of the retrieved SECs. It is to be noted that 'live alone' was removed for all analyses as the Pearson correlation coefficient showed that it is highly colinear to number of adults and children.

Name	Description	Values
income	Household total disposable income in DKK	continuous
assets	Household total assets in DKK	continuous
age_max	Highest age in the household	continuous
live_alone	Lives alone	binary
nr_adults	Number of adults in the household	continuous
nr_children	Number of children in the household	continuous
edu4	Highest attained education in the household	4 levels: Elementary school or high school, Vocational, University College, University
technic	A technically educated person in the household	binary
unemployed	A person outside the workforce in the household	binary

4. Results

4.1. Socio-economic characteristics

Firstly, it is analysed whether SECs alone can provide insights into why a building is within its energy cluster. The *prediction* step was not performed for three of the five outer cross-validation folds, as the *interpretation* step did not exclude any SECs. Overall, the results (Figure 2) show that the MCC for all three steps is low (MCC ≈ 0.1), indicating that the SECs used contain very little information about why a building is in its respective cluster. From the number of SECs selected in each step, one can see that only one variable was not used in the *threshold* and *interpretation* steps, which was always 'technic'. In the *prediction* step, which is, as mentioned, only based on two-folds of the cross-validation, always the 'income', 'assets', and 'age_max' were selected, and one fold selected additionally 'nr_adults 'and 'unemployed'. However, given the overall low MCC, the results have little general validity.

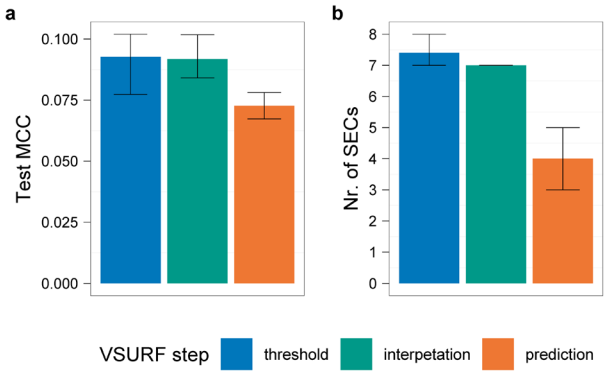

Figure 2 a) MCC for the energy use clusters for the respective set of characteristics. b) number of SECs selected in the respective steps of VSURF. It is to be noted that the *prediction* step is only based on two folds of the cross-validation.

4.1.1. Simplified energy use clusters. Following the same idea as in Schaffer et al., 2023 [10], to potentially increase the MCC and thus allow to gain at least some insight into the energy use clusters, the energy use clusters were based on their similarities merges as follows:

- E12: E1 +E2

- E34: E3 +E4
- E56: E5+E6

However, the results based on these simplified clusters indicate only an improvement in the MCC to around 0.17. Again, the *prediction* steps could only be excluded for two of the five cross-validations folds. The *interpretation* this time excluded no SEC in all folds, with two folds even using all available SECs. Thus, also for the simplified energy use clusters, the SECs seem to provide little information on why a building is within one cluster.

4.2. Social-economic and building characteristics

While the previous results focused on the SECs alone, this section focuses on whether the combination of BCs and SECs can provide more insight into why buildings are in their respective energy use clusters. First, the MCC on the test data and the number of selected characteristics were analysed (Figure 3). The *predictor* step results in the fewest selected features with a similar or, for BC+SECs, even better MCC than the other two steps, which include significantly more characteristics. Overall, however, the MCC is at most around 0.3, which does not indicate a (significant) improvement over the results of Schaffer et al., 2023 [10]. This similarity is confirmed by analysing the selected characteristics from the *predictor* step (not shown), which revealed that no SECs are selected for any of the three investigated datasets. Small deviations in the selected characteristics compared to Schaffer et al., 2023 [10] were observed, likely due to the increased variance (noise) in the predictors due to the SECs. Consequently, this means that even if only high-level statistical information about a building is available (BBR), the available SECs do not lead to a (significant) improvement in the MCC.

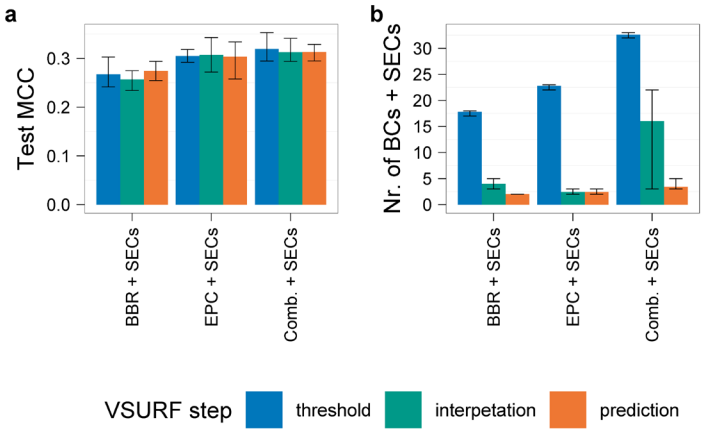

Figure 3 a) MCC for the energy use clusters for the respective set of characteristics. b) number of BCs + SECs selected in the respective steps of VSURF.

4.2.1. Simplified energy use clusters. The next step was to simplify the energy use clusters, as described above (Section 4.1.1), to see if the SECs could provide more insight for at least the simplified energy clusters. The results showed that the maximum MCC is around 0.5 and thus similar to the one previously obtained by Schaffer et al., 2023 [10]. Again, the *prediction* step led to the fewest characteristics selected at a similar MCC as the other two steps. A detailed analysis of the selected variables revealed again that no SECs were selected in the *prediction* step. These results confirm that the SECs do not provide significant additional information to the BCs alone.

5. Discussion and Conclusion

This study investigated whether clusters of thermal energy use in single-family dwellings could be understood by SECs alone or by SECs and BCs combined. A variable selection and classification approach using random forests (VSURF) was used to determine this. The results indicate that selected

SECs alone lead to poor classification performance for both original (MCC ≈ 0.1) and simplified (MCC ≈ 0.17) energy use clusters. Combined with BCs, SECS were never considered important enough to be selected by VSURF. These results indicate that SECs, at least for the data available and the SECs used, do not allow to gain (significantly) more insight into thermal energy use clusters compared to BCs alone. This work is limited to Danish single-family houses owned majoritarian by the residents. Since Denmark is known to have a high level of income equality (Gini index = 27.7) [16], it could very well be that SECs might appear more important if the sample included different regions, countries, or building types. This work suggests that SECs are not the hoped missing piece, allowing gaining more insight into why energy use pattern has a particular shape. Future work could focus on investigating actions of the occupants in the building, such as the heating habits of occupants, which is not yet easily available at a large scale, or try to process the energy use data by, e.g., disaggregating SH and DHW, before clustering. Finally, although measures have been taken to reduce the uncertainty, the BCs used are subject to uncertainty, which may introduce noise that can adversely affect the results. Therefore, future work could also further investigate and improve the quality of the BCs available today.

6. Credit-author statement

Conceptualization: M.S, A.R.H.; Methodology: M.S., J.E.V.V.; Software: M.S.; Formal analysis: all, Data Curation: M.S, A.R.H.; Writing - Original Draft: M.S.; Writing - Review & Editing: all; Visualization: M.S.; Supervision: A.R.H.; J.E.V.V., A.M.P.; Funding acquisition: A.R.H., A.M.P.

7. Acknowledgements

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

References

[1] Maach M L 2022 Bred aftale i Folketinget: Fra 2035 skal ingen boliger opvarmes af gas *DR*
[2] European Commission 2022 State aid: Commission approves €2.98 billion German scheme to promote green district heating
[3] Lund H, Werner S, Wiltshire R, Svendsen S, Thorsen J E, Hvelplund F and Mathiesen B V 2014 4th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems. *Energy* **68** 1–11
[4] European Commission 2022 Energy performance of buildings directive
[5] Hansen A R and Gram-hanssen K 2023 Over- and underconsumption of residential heating: Analyzing occupant impacts on performance gaps between calculated and actual heating demand *13TH NORDIC SYMPOSIUM ON BUILDING PHYSICS - NSB 2023*
[6] Ma Z, Yan R and Nord N 2017 A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings *Energy* **134** 90–102
[7] Wang C, Du Y, Li H, Wallin F and Min G 2019 New methods for clustering district heating users based on consumption patterns *Appl Energy* **251** 113373
[8] do Carmo C M R and Christensen T H 2016 Cluster analysis of residential heat load profiles and the role of technical and household characteristics *Energy Build* **125** 171–80
[9] Gianniou P, Liu X, Heller A, Nielsen P S and Rode C 2018 Clustering-based analysis for residential district heating data *Energy Convers Manag* **165** 840–50
[10] Schaffer M, Vera-Valdés J E and Marszal-Pomianowska A 2023 *Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses*
[11] Genuer R, Poggi J M and Tuleau-Malot C 2010 Variable selection using random forests *Pattern Recognit Lett* **31** 2225–36
[12] Genuer R, Poggi J-M and Tuleau-Malot C 2022 VSURF: Variable Selection Using Random Forests
[13] Jurman G, Riccadonna S and Furlanello C 2012 A comparison of MCC and CEN error measures in multi-class prediction *PLoS One* **7** 1–8
[14] Schaffer M, Veit M, Marszal-Pomianowska A, Frandsen M, Zbigniew Pomianowski M, Dichmann E, Grau Sørensen C and Kragh J 2023 *Dataset of smart heat and water meter data with accompanying building characteristics*
[15] Fox J and Monette G 1992 Generalized Collinearity Diagnostics *J Am Stat Assoc* **87** 178–83
[16] Wold Bank 2023 Gini index