Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
4(2): 269–281.
DOI: 10.3934/aci.2024016
Received: 18 December 2024
Revised: 19 December 2024
Accepted: 19 December 2024
Published: 26 December 2024

Research article

Clustering district heating customers based on load profiles

Vili Lavikainen[1] **and Pasi Fränti**[1,2,]*****

[1] School of Computing, University of Eastern Finland, Joensuu, Finland
[2] School of Data Science, Chinese University of Hong Kong, Shenzhen, China

* **Correspondence:** Email: pasi.franti@uef.fi.

Academic Editor: Chich-Cheng Hung

Abstract: Intelligent district heating control requires knowing the customers' past behavior and predicting their future needs. This can reduce peak energy use, optimizing energy production, accurate billing, and reducing fraud. Clustering has been used for analyzing large-scale building operational data and recognizing consumption profiles. In this work, we analyze the heat consumption profiles of district heat customers in Kuopio, Finland. We constructed two consumption profiles of their average hourly use: one for weekdays, and one for weekends. Clustering is then used to construct four consumption profiles. These profiles can be used for intelligent control, prediction of future use, and to recognize abnormal use behavior. The latter can be the first indication of a problem like heat leaking, which can prevent possible water damage.

Keywords: clustering; district heating; customer profiling

1. Introduction

District heating is the most common form of heating of buildings and their water with a 45% share in Finland [1,2]. The operating principle is that water is heated in one or more thermal power plants and transported to customers via the district heating network with the help of water or steam (Figure 1). Typical customers are residential, industrial, commercial, and public buildings. The cooled water is directed back to the heater.

Heat can be produced by solar, geothermal, wind, nuclear, and local fuels, as well as thermal power produced from combined heat and power plants. Wasted heat resources, including industrial waste heat, recycled heat from burning household waste, and biomass such as forestry waste wood, are also commonly used.

The system may include heat storage to provide flexibility of the system by reducing the demand for heat production during peak load, and by balancing the difference between demand and supply [3–6]. However, they are nowadays rare and heat consumption is allocated directly to the heating network.

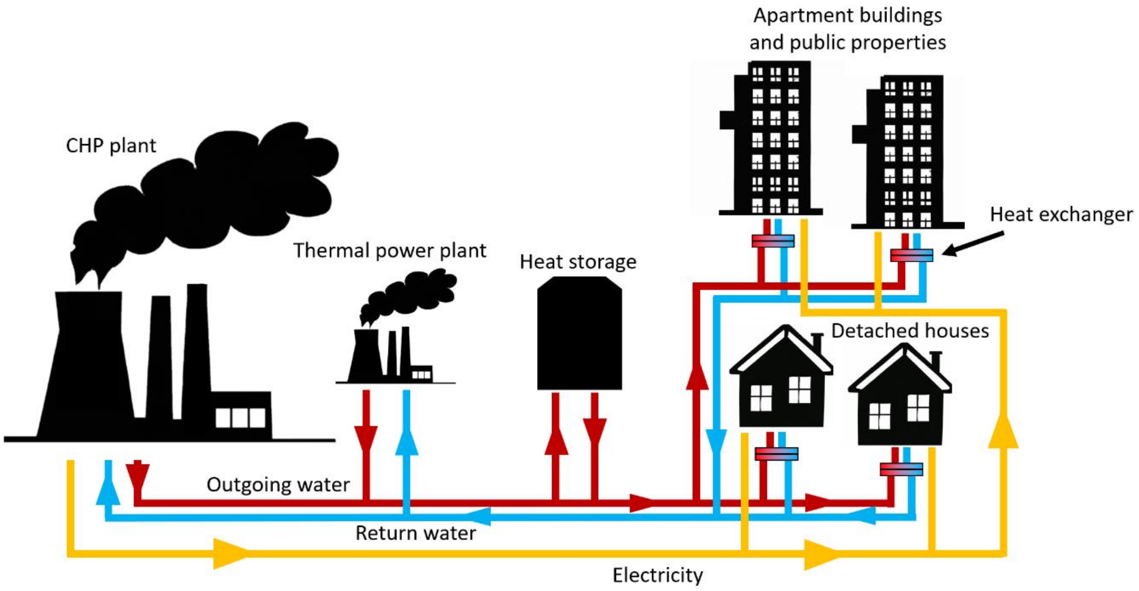


Figure 1. Principle of district heating.

The cost of district heating systems consists of production and distribution costs. The production cost depends on fuel costs, excise tax, emission rights costs, and other variable costs. The distribution cost originates from the investments, maintenance, and temperature and pressure losses in the network. To be economically feasible, the total cost must be lower than the local heat generation alternative [7].

The most important factor in the efficiency of district heating systems is to lower the distribution temperatures [8]. This requires intelligent control systems, and strategies to identify operating errors that cause high return temperatures. To plan such strategies, it is important to know the customer’s behavior and predict their future needs.

The competitiveness of district heating arises from the combination of heat production and heat distribution conditions. One important condition for heat distribution is that the heat demand must be centralized to minimize distribution costs and heat losses [9]. Low heat density in sparsely populated areas generates higher distribution costs and losses [10,11]. In a smart fourth-generation district heating system, the heating system and the distribution network must interact [12,13].

The consumption is measured by *smart meters* remotely which send data directly to the district heating company. This is an important property that allows the analysis of past and prediction of future heat use of customers. [14,15] listed several potential benefits: reducing peak energy use, optimizing indoor temperature and energy production, continuous network monitoring, accurate billing, and reducing fraud.

The data can also be used for detecting heat load patterns [16], analyzing the factors affecting energy consumption, and forecasting future use [17]. Heat load patterns are the most typical behavior patterns that reveal how different customer groups use heat. Analyzing such patterns is essential for efficient operation and management of the system [18]. A better understanding of heat use at the customer level can help to improve the efficiency of the system. District heating companies can optimize their operations, introduce new control strategies, and personalize demand management for certain customer groups [16].

The data can also be used to identify abnormal consumption because even one problematic customer can affect the performance of the network [16]. However, recognizing typical and abnormal heat uses is a complex task in a system consisting of customers with different types and characteristics.

Heat demand is determined by outside temperature, inside temperature, building materials, building structure, weather conditions, and individual behavior [19]. Individual behavior can include a secondary heating system such as a fireplace or an air source heat pump, hot water consumption, the number of people in the building, electrical devices producing heat, the building's ventilation, and the wind directed at the building. [20] estimated that 40% of heating is spent on room heating, 35% on ventilation heating, and 25% on domestic water heating in a residential building.

Clustering is one of the most popular exploratory data mining methods for analyzing heat load patterns [21]. It has been successfully used for analyzing large-scale building operational data [22], recognizing consumption profiles [23], and as a preprocessing step when using other data mining techniques to predict future energy use [24,25].

In this work, we analyze the heat consumption profiles of district heat customers in Kuopio, Finland. We constructed two consumption profiles of the users. The first is the average hourly use on weekdays, and the second one is the average hourly use on weekends. The profiles are then clustered to construct four average profiles.

These clusters provide useful information about the most common consumption profiles. They can be used for a better prediction of heat use and for intelligent heat control, which could reduce heat consumption without the customer noticing the effect. The consumption models can be used to alert when the consumption of a customer is significantly higher than the average usage in the cluster. It can be the first indication of a problem like heat leaking. Such a warning system can reduce the possibility of water damage.

The rest of the paper is organized as follows. The data is first given in Section 2. The clustering process is described in Section 3 including pre-processing to fill in missing values, removing erroneous data, feature extraction, normalization, and the clustering algorithm. The resulting clusters are presented in Section 4 and their content is analyzed. Conclusions are drawn in Section 5.

2. District heating data

We use two data sources: customer consumption measures and weather data. The consumption measures come from Kuopio district heating customers in Finland during 2021. The weather data is provided by the Finnish Meteorological Institute for Kuopio's Savilahti area in Kuopio.

The consumption data consists of 54,657,082 measurements in megawatt-hours (MWh) from 6089 customers. Each measurement has a status value that indicates the quality of the measurement. The statuses are in accordance with metered service consumption report (MSCONS) message service codes: successful reading (136), uncertain reading (Z02), and failed reading (Z03) [26]. The distribution of the readings for different statuses is shown in Table 1. The measurements are made once every hour, so there are 24 daily energy consumption measures for every customer. In total, 733,885 readings were missing from the database; they were left out of the final data. We also filtered out five customers with uncertain data or failed reading. The distribution of the 6084 used customers is summarized in Table 2 according to the building type.

Table 1. Status codes in data. The status code 136 indicates a good reading. These measurements will be used in this study.

Status	Readings
136	54,557,377
Z02	48
Z03	9,965
Missing	733,885
Total	55,301,275

Table 2. Distribution of district heating customers by building type.

Building type	Customers
Detached house	3777
Terraced house and apartment building	1602
Service building	324
Public service	213
Industrial	131
Transportation	20
Other	17
Total	6 084

3. Clustering customer profiles

To form daily consumption habits characteristic of customers, we use clustering. The clustering algorithm divides the data into groups (clusters) where customers in the same cluster are more similar than customers in other clusters. The overall process is shown in Figure 2. The details of the cluster analysis process are described in the following sub-sections.

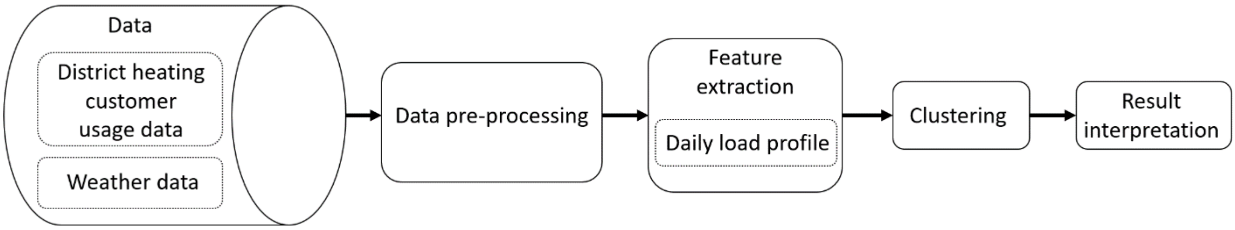


Figure 2. The cluster analysis includes four main steps: data pre-processing, generation of the customer profiles, clustering the profiles, and content analysis of the clusters.

3.1. Pre-processing

We pre-processed the data to detect and fix errors. There were a relatively large number of missing values. When more than four measures from the customer's data were missing on the same day, the entire day was deleted from the data. In other cases, missing readings are created by linear interpolation. Unrealistically large values were also filtered out. Missing outdoor temperature data were recovered using the open data service of the Finnish Meteorological Institute [27].

3.2. Consumption profiles

We extract daily consumption patterns (12 measurements per day) separately for weekdays and weekends. The result is a 24-dimensional feature vector normalized according to the outdoor temperature.

Figure 3 shows the daily heat consumption of customers of two different building types. Weekdays are marked in red and weekend days in blue. For service properties, heating peaks are most often at two and six o'clock, while for residential buildings, three significant heating moments are at 2, 5, and 17 o'clock. The morning peak in residential building customers' data is due to people using hot water in the shower (the evening peak exists but is less visible). Both customer types have slightly lower usage during the weekend, but besides this, there is very little we can observe from the averages. The consumption varies a lot between individual customers for both customer types.

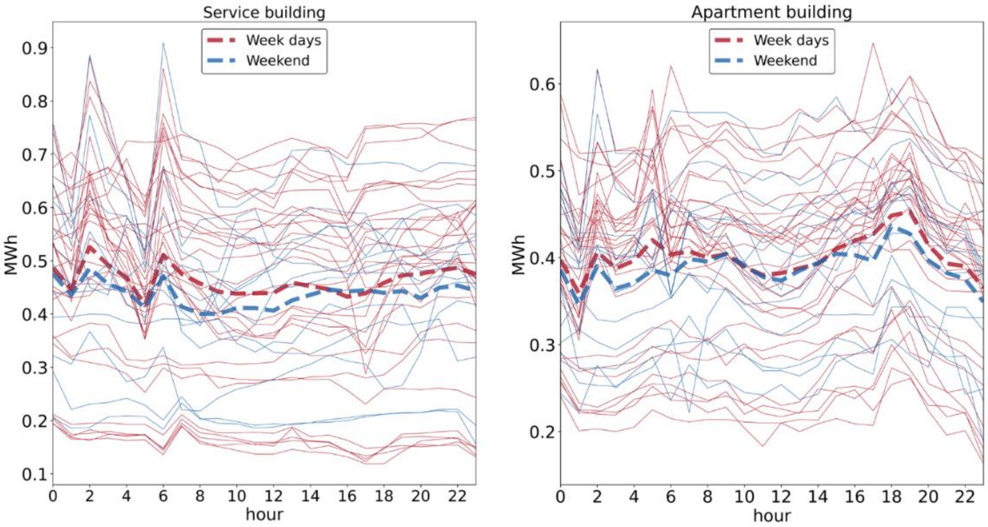


Figure 3. Plot of several daily heat uses of two sample customers. The numbers are averaged over the six months. The daily average of these days is marked with a thicker line.

The biggest common affecting factor is the outside temperature. Figure 4 shows the total consumption of district heating by month and the average temperature for the months. The consumption closely follows the outside temperature and is roughly 7 times higher in the peak winter months (December, January, and February) than in the mid-summer months (June and July). In the summer months, the heating is mainly used for hot water.

Figure 5 shows the effect of outdoor temperature. The consumption decreases linearly when the outside temperature increases until it reaches 15 °C. This is the temperature value when the heating of the building is turned off. The remaining consumption is mainly for hot water heating, and in some humid spaces, floor heating is continuous regardless of the weather. The further decrease of consumption beyond 15 °C in residential buildings can be explained by using less warm water when the weather becomes warmer.

To eliminate the effect of outdoor temperature, we construct two models of the daily consumption relative to the outside temperature by linear regression as suggested [28]. One model is for the cases when the average outdoor temperature is below 15 °C, and the other for those above 15 °C. The resulting modified daily load profile is 0/1-normalized so that normalized measurements have zero mean and unit variance.

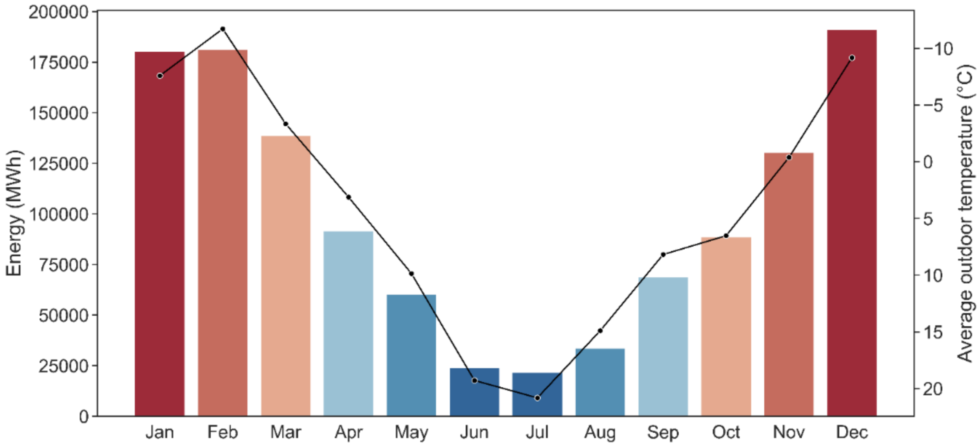


Figure 4. Energy consumption varies significantly depending on the time of year.

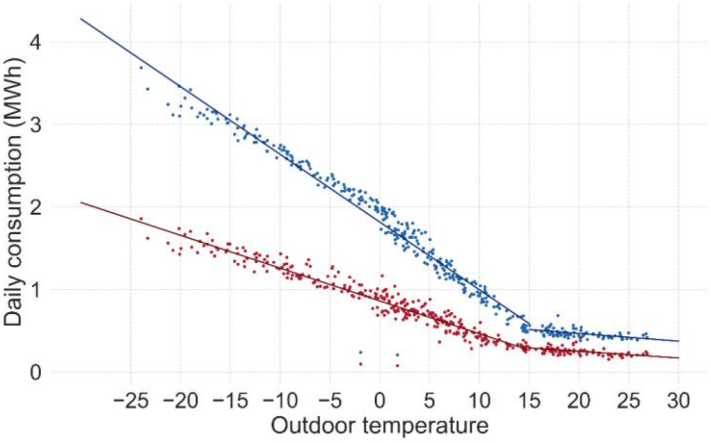


Figure 5. Example of daily consumptions of two customers (red and blue) and their two regression lines: one for outdoor temperatures below 15 °C, and another for above 15 °C.

3.3. Clustering algorithm

K-means is the most widely used clustering algorithm despite its sensitivity to random initialization. However, this deficiency can be compensated by better initialization and repeating the algorithm multiple times [29], or by using an additional random swap step as a wrapper around the k-means [30]. K-means has also been generalized to graphs [31], sets [32], and time series [33].

The customer profiles are essentially time series. For this reason, we use the k-Shape algorithm [33] which is essentially a k-means variant adopted for time series. It operates iteratively like k-means but with distance function and centroid calculation tailored for time series. Its advantages are fast speed, available implementation, and well-known properties [34]. We use the implementation available in the tslearn-library [35].

To compensate the sensitivity of k-means by repeating the algorithm 1000 times with different (random) initialization and choose the best according to the optimization function (sum-of-squared error) following the principle presented in [31]. The algorithm stops when the error decreases less than a threshold value (10^{-6}), or the number of iterations reaches a maximum value of 2000.

For selecting the number of clusters, we ran the algorithm from 2 to 40 clusters. The results were compared with five cluster validity indexes: Calinski-Harabasz-index (CHI) [36], Davies-

Bouldin index (DBI) [37], Silhouette Coefficient (SC) [38], Sum of Squared Distances (SSD), and Gap statistic (GS) [39]. For a more detailed analysis of these measures, we refer to [40].

None of the indexes indicated that there would be natural clusters in the data, see Figure 6. The number was therefore left to us, the researchers, to decide. Since our goal is to create average customer profiles that can be manually analyzed, the number of clusters was kept relatively small: four.

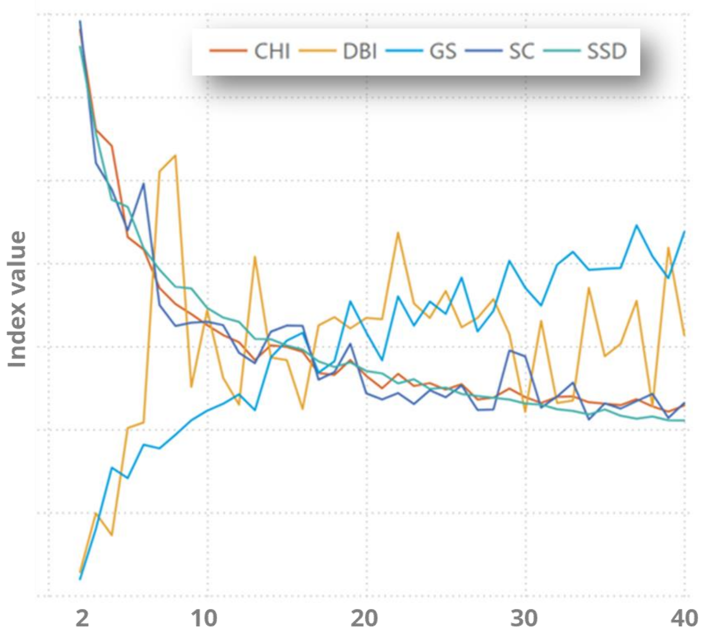


Figure 6. Comparing the number of clusters with four index values. If the data contained clusters, the minimum point of CI, SC, and SSD, and at the maximum point of DBI and GS should indicate this value. Only the least reliable index (DBI) provides visible maximum and most likely due to noise.

4. Results

After the pre-processing, five of the 6084 customers did not have enough data and were therefore dropped out. Four clusters were then created from the remaining 6079 customer profiles.

Figure 7 summarizes the average heat profiles of the four clusters. Figure 8 shows the distribution building types and consumption profiles of the customers. Small houses are the most common building type in every cluster. Despite being mostly used for living, they have different heat use profiles. Terraced and apartment buildings are in all clusters, but mostly in cluster 3, which is formed mainly of residential buildings and almost completely lacking services and industrial buildings.

Clusters 1 and 3:

All profiles have the smallest consumption at night and then increase in the morning but only clusters 1 and 3 have a clear peak in the evening around 18–20 o'clock. On the weekend, the consumption profiles are similar, but the morning peak appears about two hours later than on weekdays.

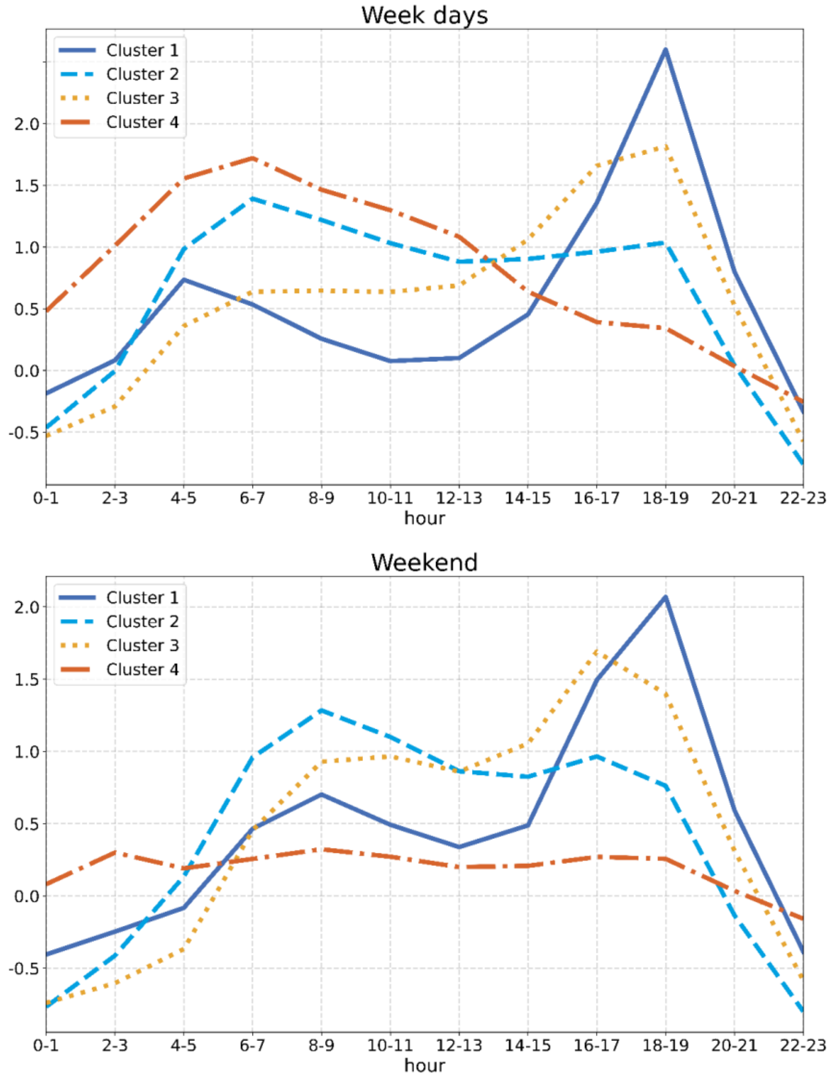
It is likely that the residents of cluster 3 work outside the home since the peaks align well with traditional working hours. There is less variation in cluster 3 than in cluster 1 having more residents in the same building, which smooths the variations.

Cluster 2:

Many service and industrial buildings belong to cluster 2, where the consumption profile follows the working hours. It is likely determined by the ventilation, and the use is flat throughout the normal opening hours.

Cluster 4:

The profile of cluster 4 decreases steadily from morning to evening and is constant on weekends. The cluster consists mostly of non-residential buildings where the most heat consumption typically takes place in the morning on weekdays and is flat on weekends. Many of the buildings are used only on weekdays.


Week days

Weekend

Figure 7. Average heat consumption profiles of the four clusters.

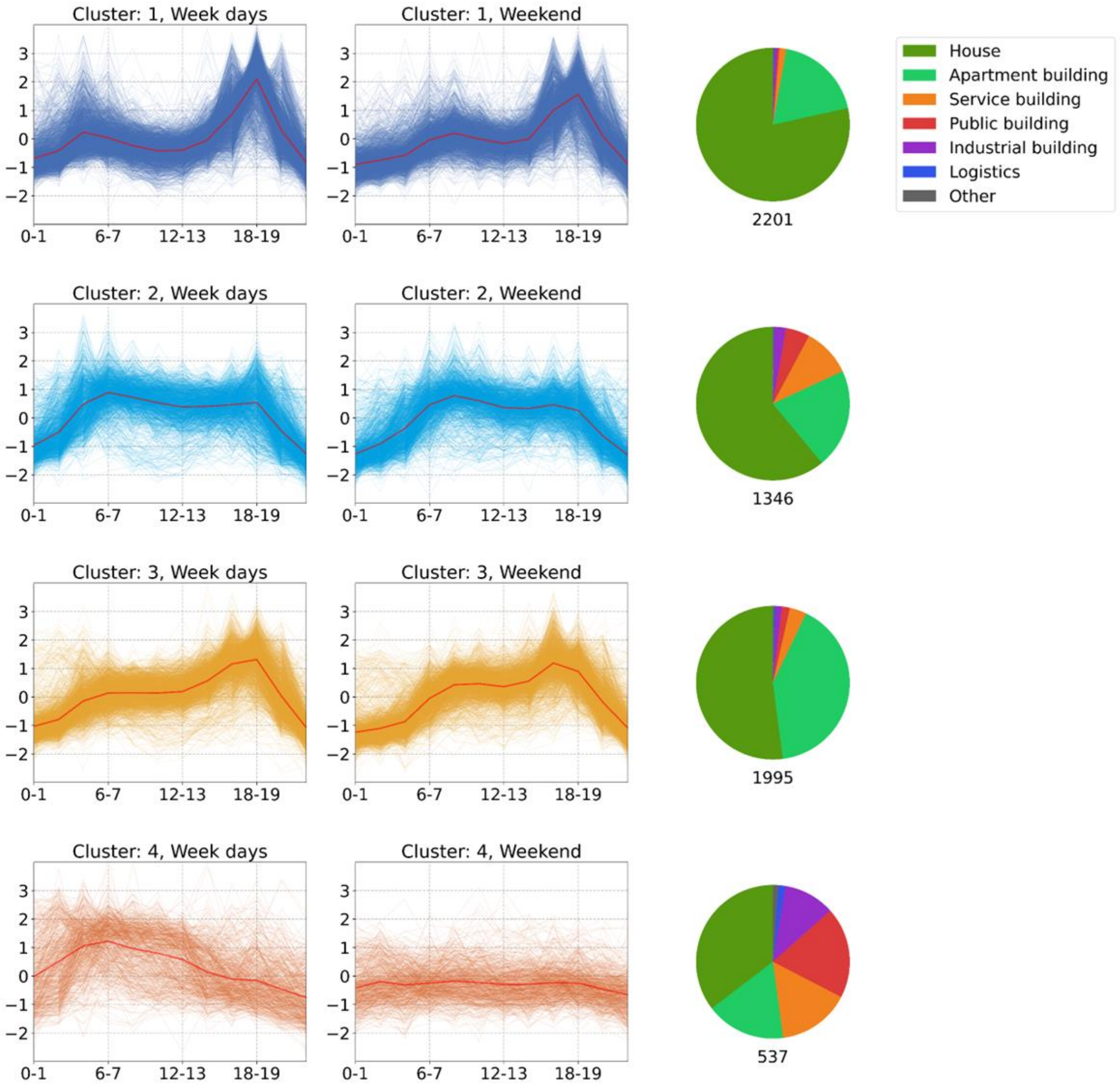


Figure 8. Consumption profiles of the customers in each cluster (left), and distribution of the building types in the four clusters (right).

The total energy use in each cluster is shown in Figure 9 according to the building type. Clusters 1 and 3 contain the most customers (2201 and 1995), yet the total consumption of customers in cluster 1 is the smallest. For the district heating company, clusters 2 and 3 are the most interesting because of the largest energy consumption. The number of service buildings is small, but they constitute a significant part of the energy consumption, especially in cluster 2. The terraced and multi-story buildings dominate the energy consumption in clusters 1 and 3. The total energy use of the small houses is insignificant in all clusters. In cluster 1, they make up more than a quarter of all building types; yet their heat consumption is less than a third of the total heat consumption in the cluster.

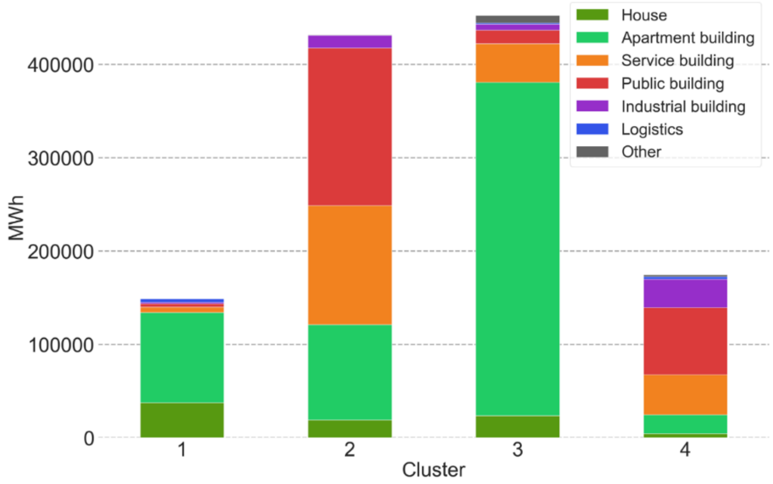


Figure 9. Total energy use of customers in each cluster according to the building type.

5. Conclusions

We constructed four average customer profiles of district heating customers in Kuopio by clustering their daily heat use profiles. These clusters provided useful information about the most common consumption profiles. They can be used for predicting future heat use and intelligent heat control, which could reduce heat consumption without the customer even noticing it. The models could also identify deviations from the normal (average) consumption to alert if the customer's consumption is significantly higher than the average usage. It can be the first indication of a heat leak or other problem. Such a warning system could reduce the extent of possible water damage.

Limitations: The shape-based measure was shown superior in [33], but the results later in [41] did not favor this measure as strongly. Our data has fixed length where all measurements have recorded at the same time of day. It might be worth to take closer look whether Euclidean distance or dynamic time warping would be more accurate.

Another limitation is that our study did not consider the energy stored in the building's structures. Solar radiation and previous heating are stored in the structures, which are later released to the interior. As a future work, this thermal inertia could be used in the model to create a more sophisticated prediction model.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools for this article.

Conflict of interest

The authors declare no conflict of interest.

References

1. *Energiateollisuus ry, Energy year 2021—electricity (Finnish)*, Finnish Energy, 2022. Available from: https://energia.fi/en/statistics/energy-year-2021-electricity/.
2. *Motiva, District heating (Finnish)*, Motiva Oy, 2022. Available from: https://www.motiva.fi/koti_ja_asuminen/rakentaminen/lammitysjarjestelman_valinta/lammitysmuodot/kaukolampo.

3. K. Skytte, O. Olsen, Regulatory barriers for flexible coupling of the Nordic power and district heating markets, *Proceedings of 13th International Conference on the European Energy Market (EEM)*, 2016, 1–5. https://doi.org/10.1109/EEM.2016.7521319
4. G. Schweiger, J. Rantzer, K. Ericsson, P. Lauenburg, The potential of power-to-heat in Swedish district heating systems, *Energy*, **137** (2017), 661–669. https://doi.org/10.1016/j.energy.2017.02.075
5. M. Razmara, G. Bharati, D. Hanover, M. Shahbakhti, S. Paudyal, R. Robinett III, Building-to-grid predictive power flow control for demand response and demand flexibility programs, *Appl. Energ.*, **203** (2017), 128–141. https://doi.org/10.1016/j.apenergy.2017.06.040
6. H. Li, S. Wang, Challenges in smart low-temperature district heating development, *Energy Procedia*, **61** (2014), 1472–1475. https://doi.org/10.1016/j.egypro.2014.12.150
7. U. Persson, S. Werner, Heat distribution and the future competitiveness of district heating, *Appl. Energ.*, **88** (2011), 568–576. https://doi.org/10.1016/j.apenergy.2010.09.020
8. H. Gadd, S. Werner, Achieving low return temperatures from district heating substations, *Appl. Energ.*, **136** (2014), 59–67. https://doi.org/10.1016/j.apenergy.2014.09.022
9. S. Werner, District heating and cooling, In: *Reference module in earth systems and environmental sciences*, 2013, 1–7. https://doi.org/10.1016/B978-0-12-409548-9.01094-0
10. S. Nilsson, C. Reidhav, K. Lygnerud, S. Werner, Sparse district-heating in Sweden, *Appl. Energ.*, **85** (2008), 555–564. https://doi.org/10.1016/j.apenergy.2007.07.011
11. C. Reidhav, S. Werner, Profitability of sparse district heating, *Appl. Energ.*, **85** (2008), 867–877. https://doi.org/10.1016/j.apenergy.2008.01.006
12. F. Levihn, CHP and heat pumps to balance renewable power production: lessons from the district heating network in Stockholm, *Energy*, **137** (2017), 670–678. https://doi.org/10.1016/j.energy.2017.01.118
13. M. Sameti, F. Haghighat, Optimization approaches in district heating and cooling thermal network, *Energ. Buildings*, **140** (2017), 121–130. https://doi.org/10.1016/j.enbuild.2017.01.062
14. G. Mbiydzenyuy, S. Nowaczyk, H. Knutsson, D. Vanhoudt, J. Brage, E. Calikus, Opportunities for machine learning in district heating, *Appl. Sci.*, **11** (2021), 6112. https://doi.org/10.3390/app11136112
15. S. Darby, Smart metering: what potential for householder engagement? *Build. Res. Inf.*, **38** (2010), 442–457. https://doi.org/10.1080/09613218.2010.492660
16. E. Calikus, S. Nowaczyk, A. Sant'Anna, H. Gadd, S. Werner, A data-driven approach for discovering heat load patterns in district heating, *Appl. Energ.*, **252** (2019), 113409. https://doi.org/10.1016/j.apenergy.2019.113409
17. A. Kipping, E. Trømborg, Modeling and disaggregating hourly electricity consumption in Norwegian dwellings based on smart meter data, *Energ. Buildings*, **118** (2016), 350–369. https://doi.org/10.1016/j.enbuild.2016.02.042
18. M. Noussan, M. Jarre, A. Poggio, Real operation data analysis on district heating load patterns, *Energy*, **129** (2017), 70–78. https://doi.org/10.1016/j.energy.2017.04.079
19. Z. Ma, H. Li, Q. Sun, C. Wang, A. Yan, F. Starfelt, Statistical analysis of energy consumption patterns on the heat demand of buildings in district heating systems, *Energ. Buildings*, **85** (2014), 464–472. https://doi.org/10.1016/j.enbuild.2014.09.048
20. L. Koskelainen, R. Saarela, K. Sipilä, *Kaukolämmön käsikirja (Finnish)*, Helsinki: Energiateollisuus ry, 2006.

21. Y. Lu, Z. Tian, P. Peng, J. Niu, W. Li, H. Zhang, GMM clustering for heating load patterns in-depth identification and prediction model accuracy improvement of district heating system, *Energ. Buildings*, **190** (2019), 49–60. https://doi.org/10.1016/j.enbuild.2019.02.014
22. Z. Yu, F. Haghighat, B. Fung, L. Zhou, A novel methodology for knowledge discovery through mining associations between building operational data, *Energ. Buildings*, **47** (2012), 430–440. https://doi.org/10.1016/j.enbuild.2011.12.018
23. F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, *Energ. Buildings*, **75** (2014), 109–118. https://doi.org/10.1016/j.enbuild.2014.02.005
24. A. Goia, C. May, G. Fusai, Functional clustering and linear regression for peak load forecasting, *Int. J. Forecasting*, **26** (2010), 700–711. https://doi.org/10.1016/j.ijforecast.2009.05.015
25. P. Duan, K. Xie, T. Guo, X. Huang, Short-term load forecasting for electric power systems using the PSO-SVR and FCM clustering techniques, *Energies*, **4** (2011), 173–184. https://doi.org/10.3390/en4010173
26. *Ediel, Message handbook for Ediel implementation guide for metered services consumption report*, Ediel forum, 2010. Available from: https://ediel.org/wp-content/uploads/2019/02/MSCONS-24E-20100215.pdf.
27. *FMI, Ilmatieteen laitoksen avoin data ja lähdekoodi*, Finnish Meteorological Institute, 2023. Available from: https://www.ilmatieteenlaitos.fi/avoin-data.
28. C. Wang, Y. Du, H. Li, F. Wallin, G. Min, New methods for clustering district heating users based on consumption patterns, *Appl. Energ.*, **251** (2019), 113373. https://doi.org/10.1016/j.apenergy.2019.113373
29. P. Fränti, S. Sieranoja, How much can k-means be improved by using better initialization and repeats? *Pattern Recogn.*, **93** (2019), 95–112. https://doi.org/10.1016/j.patcog.2019.04.014
30. P. Fränti, Efficiency of random swap clustering, *J. Big Data*, **5** (2018), 13. https://doi.org/10.1186/s40537-018-0122-y
31. S. Sieranoja, P. Fränti, Adapting k-means for graph clustering, *Knowl. Inf. Syst.*, **64** (2022), 115–142. https://doi.org/10.1007/s10115-021-01623-y
32. M. Rezaei, P. Fränti, K-sets and k-swaps algorithms for clustering sets, *Pattern Recogn.*, **139** (2023), 109454. https://doi.org/10.1016/j.patcog.2023.109454
33. J. Paparrizos, L. Gravano, Fast and accurate time-series clustering, *ACM Trans. Database Syst.*, **42** (2017), 8. https://doi.org/10.1145/3044711
34. P. Fränti, S. Sieranoja, K-means properties on six clustering benchmark datasets, *Appl. Intell.*, **48** (2018), 4743–4759. https://doi.org/10.1007/s10489-018-1238-7
35. R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, et al., Tslearn, a machine learning toolkit for time series data, *J. Mach. Learn. Res.*, **21** (2020), 1–6.
36. T. Caliński, J. Harabasz, A dendrite method for cluster analysis, *Commun. Stat.*, **3** (1974), 1–27. https://doi.org/10.1080/03610927408827101
37. D. Davies, D. Bouldin, A cluster separation measure, *IEEE Trans. Pattern Anal.*, **PAMI-1** (1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909
38. P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, **20** (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
39. R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, *J. R. Stat. Soc. B*, **63** (2001), 411–423. https://doi.org/10.1111/1467-9868.00293
40. Q. Zhao, P. Fränti, WB-index: a sum-of-squares based index for cluster validity, *Data Knowl. Eng.*, **92** (2014), 77–89. https://doi.org/10.1016/j.datak.2014.07.008

41. A. Javed, B. Lee, D. Rizzo, A benchmark study on time series clustering, *Machine Learning with Applications*, **1** (2020), 100001. https://doi.org/10.1016/j.mlwa.2020.100001

©2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)