Regional and Business Studies (2022) Vol 14 No 1, 57-72
Hungarian University of Agriculture and Life Sciences Kaposvár Campus
Kaposvár
doi: 10.33568/rbs.3343

K-MEANS CLUSTER ANALYSIS OF HOURLY MEASURED POWER DEMAND IN THE DISTRICT HEATING SYSTEM IN KAPOSVÁR – PUTTING THEORETICAL FIGURES INTO PRACTICAL USE

Uwe RADTKE
Hungarian University of Agriculture and Life Sciences, Doctoral School in Management and Organizational Sciences, Kaposvár Campus, 7400 Kaposvár, Guba Sándor u. 40., Hungary

ABSTRACT

In this paper the actual cluster analysis is performed to identify clusters within the Kaposvár district heating system. The data were not measured directly in households but at heat transfer stations. The smart meters were installed at the heat transfer stations for several reasons: to measure and control the required supply temperature and also to find leakages quicker and easier. K-means method was used and four different clusters were identified, which were the following: high demand of heating with long operating hours, low demand of heating with long operating hours and low demand of heating with short operating hours. The details and the determined values will be used for further research. First steps towards identifying new heat sources have also been identified.
Keywords: smart meter data analysis, carbon emissions reduction, waste heat utilization, energy efficiency, sustainable energy
JEL codes: C38, L97

INTRODUCTION

Faced with the challenges of climate change and the need to ensure sustainable economic growth and social cohesion, Europe must achieve a genuine energy revolution to reverse current unsustainable trends and meet ambitious policy expectations. To meet this challenge, district heating (and cooling) systems must become more efficient, smarter and cheaper. "In contrast to the analysis of electricity smart meter data, little research has been published on district heating (smart) meter data" (*Tureczek et al.*, 2019)

District heating (also known as heat networks or teleheating) is a system that distributes heat generated at a central location through an insulated pipe system to meet residential and commercial heating needs such as space heating and water heating. The heat is often generated from a combined heat and power plant burning fossil fuels or biomass, but also from heat-only boiler stations, geothermal heating, heat pumps and central solar heating. Waste heat from factories and nuclear power electricity generation is also used and is common. District heating plants can provide higher efficiency and better pollution reduction than local boilers. Some research has shown that district heating with combined heat and power (CHP) is the cheapest way

to cut carbon emissions, and has one of the lowest carbon footprints of all fossil generation plants. (*Andrews*, 2009).

The first aim of this paper is to analyse potential clusters of the measured heat demand data to provide an insight into possible key points for further discussions. The clusters found will be the starting points for additional measures to be taken in the future, which may include not only recommendations to homeowners for insulation, but also possible changes in the tariffs and charges to be paid. The tariff structure can play a key role in encouraging the adoption of environmentally friendly behaviour – lower demand would mean lower supply temperatures, less gas or biomass and – for the end user – fewer static parts in the bills. These issues are not discussed in this paper, they should be investigated in further research.

The second aim is to provide some additional ideas on smart energy use and energy production, in particular in the field of district heat production. The calculation of the clusters alone does not provide too much practical added value but the reduction of CO_2 emission used for heat production would be a sustainable result that can be based on the calculated figures. The reduction in heat demand and consumption for some individual users is highly dependent on general environmental factors such as external temperature, weather conditions and also political measures.

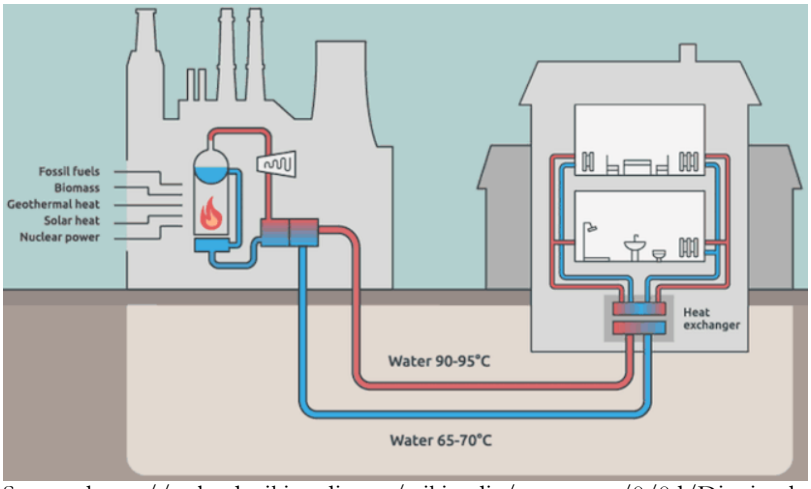
District heating (dh)

District heating is the supply of heat to buildings through a heating network that transports thermal energy. Water is heated using a power plant, solar thermal or geothermal installation or a large heat pump and then piped into an insulated (and usually buried) network of of pipes, directly to the buildings connected to the system. The water then flows through a transfer station to the building's own heat distribution system, which provides the heating energy and hot water supply. Once the water cools down, it flows back to the original heat source and the cycle starts all over again. In other words, buildings with district heating can do without their own heating systems and chimneys. A visual image of the above description can be seen in *Figure 1*. District heating systems are supply systems that consist of district heating plants, pressure and volume maintenance plants, water treatment plants, district heat transportation systems, and distribution networks, as well as customer transfer stations. As such a system must always balance district heating output and production, an additional heating center is responsible for the efficient control of the system.

District heating is therefore a valuable product that is predominantly produced by combined heat and power production (i.e., combined production of electricity and heat). The supply of the district heating is as simple as this: The heat from the district heating system is transported to the customer via a pipeline system using a transport medium (usually hot water). The heat is transferred to the building via a domestic transfer station. The cooled water is then returned from the building to the network.

District heating not only reduces the demand for resources – it also protects the climate. The requirements of the Net Zero Emissions by 2050 Scenario call for the combined share of renewable sources and electricity in global district heat supplies to increase from 8% of today to about 35% by the current decade, contributing to a reduction of carbon emissions from heat generation by more than a third. (*IEA*, 2021)

Figure 1: How the district heating system works



Source: https://upload.wikimedia.org/wikipedia/commons/0/0d/District_heating.gif

In 2021, the number of apartments and homes using district heating in Kaposvár was 6 900, which represents 30% of the city's dwellings, and provides heating for 305 other heated buildings. This is an exceptional and outstanding proportion in Hungary (*Bánkuti & Zanatyné Uitz*, 2021).

The continuous drive to improve the carbon footprint of heat production and contribute to sustainability can be seen in several areas. An example is the installation of a biogas plant in 2007. It not only reduced the dependency on natural gas and ensured the additional supply of methane gas, but it also contributed to the survival of the AGRANA subsidiary Magyar Cukor Zrt. (Hungarian Sugar Private Limited Company, hereinafter referred to as "the sugar factory"), with a capital investment of HUF 1.7 billion (about EUR 6.8 million).

In 2015, the city of Kaposvár completely replaced its outdated diesel-powered city buses with 40 compressed natural gas (CNG) vehicles, which significantly improved the city's air quality. The sugar factory power plant - with its two extremely large fermenters of a useful volume of 12000 m^3 – is still unique in the European sugar industry. Local professionals found that the production reached the value of 140.000 m^3 with an average methane content of 53% (*Csima & Szendefy*, 2009).

In 2019, the natural gas-fired heating plant operated with an installed thermal capacity of 51.7 MWt, and with a combined electricity generation of 1.9 MWt. The electrical capacity of the installed gas-fired engine was 1.35MWe. These technical facts show the intention to strive for more sustainable production throughout the process.

Literature review

The literature which clustering method should be based on was reviewed in details beforehand and will be published within the Ohio Journal of Science (*Radtke*, 2022). The conclusion shall be summarised below for a better comprehension of the subject.

Conclusion

The reviewed literature on clustering, mainly describe the methodology of clustering by K-means. The K-means algorithm is considered to be the best known and most frequently used clustering method, which divides the dataset into k clusters by minimizing the sum of all distances to the respective cluster centers (*Ramos et al.*, 2015). The use of K-means clustering algorithm is well covered in the literature and can serve as a basis for further tests on models and other clustering methods. Several alternative methods have been described and tested, but no common ground for additional clustering methods was found in the literature reviewed. Each researcher who used an alternative approach only compared results using K-means. The main differences between the articles are the database, some very small, and the data preparation. Each article used its own set of data, some very small, ranging from only two apartments up to over 500 metering points within district heating networks, while for similar electricity measurements nearly 15.000 measuring points were analysed.

However, the literature reviewed observed additional differences in those who performed a comparative analysis between electricity and heating measures: the influencing factors indicate that the outdoor temperature has a significant effect, including on the area of living. Data gathered in Genoa showed fewer peaks and less significant differences than the evaluation of data collected in Northern Europe (GB, Sweden). The conclusion was fairly similar. The most influential factors affecting volatile consumption were temperature and the type of building in which the measures were taken. The results showed the effect of modern insulation compared to non-insulated insulation. However, all researchers adjusted the data, which indicated that consumption remained stable over the period analyzed.

The use of K-means clustering algorithm is mandatory and should not be omitted. All other clustering approach should be compared to the results of the K-means method. According to the literature reviewed, the K-means algorithm (which was used in several other methods such as nbclust (GMM) showed similar results to all other methodologies used. Outliers should be respected and accepted – but this is a general rule when using K-means algorithm. According to the literature reviewed, the best documented approach is K-means; only one article used other algorithms, while two thirds of the articles contained findings and results for K-means.

Checking the geographical origin of the data, there is no coverage of Eastern European countries. Even before the functional limitation on heat loads and patterns or even district heating, no observed data were produced using data from Eastern Europe. This may be due to the lack of interest in research in or the lack of researchers' interest. But the lack of available data may also be a reason for the lack of research. A fourth explanation could be the geographical boundedness of publishing journals. Here again seems to be a research gap.

As some practical ideas should also be developed, an additional literature review on waste heat utilization was carried out. *Fang et al.*, (2013) provided a detailed investigation into the thermal grade and quantity of low-grade waste heat sources in Chifeng, China. Waste heat recovery for productive use can reduce CO_2 emissions as well as reduce the use of fossil fuels and water dissipation. The authors proposed a holistic approach to an integrated and efficient utilization of low-grade industrial waste heat.

Woolley et al., (2018) describe a framework for waste heat energy recovery that provides a four-step methodology for manufacturers. Their article provides a recommendation for facilities to assess production activities, analyzing the compatibility of waste heat sources in terms of exergy balance and temporal availability. They did this by selecting appropriate heat recovery technologies and decision support based on economic benefits. Their work was based in the UK and their research observed a lack of data. Finally, they used data that consisted of original data from a commercial company and published literature which they referred to where the information they wanted was not available from the company.

The potential for waste heat recovery data center was analyzed by *Wahlroos et al.*, (2018). They used Northern European countries, in particular Finland, as the data center operators of Finland were planning to reuse waste heat in district heating. They attempted to overcome the lack of transparency of the business models between the district heating network operator and data center operator by a life cycle assessment analysis. The authors proposed a systematic process of change to succeed in changing the priority of waste heat utilization in the data center and district heating market.

Knudsen et al., (2021) explored the key technology of thermal energy storage. It is used to enable the utilization of industrial waste heat in district heating. The authors tried to solve the problem of sizing these storages in heating plants using a variable waste-heat source. They developed a model that combined dynamic simulation and a model predictive control approach. This takes into account the dynamics and optimal control of the heating plant using a thermal energy storage. Their case study was carried out in Norway.

One of the most influential papers was written by (*Köfinger et al.*, 2018), who examined different technical and infrastructural options, and conducted economic analyses to better implement these options. Their study was based on data from the city of Linz (Austria). The existing industry showed significant unused waste heat potentials that could be integrated into the existing urban district heating system but no additional seasonal storage was available. This was due to a competition with an existing waste incineration plant in Linz. The authors found that by operating the seasonal storage strategically, the number of charging cycles could be increased, thereby the revenues of the system can significantly be increased. A combined utilization of the seasonal storage system would allow waste heat to be transferred from the summer to the winter period and it can be used as a short-term buffer.

Summarizing the reviewed literature in regard to the source of data or place of research included again a lack of easter Europa coverage. A number of studies were reviewed, particularly with regard to the source of data – no English-language studies investigating Hungary, Romania or Slovakia, or even other eastern European countries were found. China, Western and Northern Europe were investigated mainly because of their natural environment (long winters with a long heating period) and their state of industrialisation. However, the research of using alternative heat sources to replace fossil fuels is essential to sustainable and low-carbon heat supply and it is therefore a key element in demonstrating the energy systems of the future.

MATERIAL AND METHODS

As found and described in the literature review, the clustering approach should be taken using the K-means method. K-means clustering is a method of vector quantization, originally from signal processing, which aims to divide the n observations into k clusters, where each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as the prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. K-means clustering minimizes intra-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be a more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. Better Euclidean solutions can be found, for example, using k-medians and k-medoids (*MacQueen*, 1967). This is a well-known approach to clustering data (*Jain*, 2010).

The most common algorithm uses an iterative refinement technique. Due to its ubiquity, it is often called "the K-means algorithm"; it is referred to as native K-means:

Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster with the nearest mean: that with the least squared Euclidean distance[1] (this means partitioning the observations according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\} \quad (1)$$

where eachis assigned to exactly one $S_{(t)}$, even if it could be assigned to two or more of them. $S^{(t)}$ are sets $S = \{S_1, S_2, ..., S_k\}$.

Update step: compute recalculated means (centroids) for the observations assigned to each cluster:

$$m_i^{t+1} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \quad (2)$$

The algorithm converges when the assignments no longer change. The algorithm is not guaranteed to find the optimum. The algorithm is often presented as assigning objects to the nearest cluster by distance. Using a different distance function other than (squared) Euclidean distance may prevent the algorithm from converging. Various modifications of K-means such as spherical K-means and k-medoids have been proposed to allow using other distance measures (*MacKay*, 2003).

The data was provided by the district heating company of Kaposvár. The data used consists of around 300 devices, measured hourly for 365 days. With a set of 3.3 million measurements the analysis can no longer be done on a simple PC. But the

[1] In mathematics, the Euclidean distance between two points in Euclidean space is the length of a line segment between the two points. It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem, therefore occasionally being called the Pythagorean distance. These names come from the ancient Greek mathematicians Euclid and Pythagoras, although Euclid did not represent distances as numbers, and the connection from the Pythagorean theorem to distance calculation was not made until the 18th century.

research has proved our assumptions. Similar results can be achieved with random samples of 10.000 data. As only two dimensions have been available for public use - the result may still reveal 4 clusters in a 2-dimensional space in the future. No district heating company is likely to want high demand with short operating hours. Of course, the method used to analyze the data included standards such as removing unmeasured sets, controlling for data bias and transforming the data.

Data description and preliminary steps

After the initial analysis, the data was converted from the given SQL Server format into a STATA readable format. The first step was to use STATA to convert the date and time as well as the measured demand (P1) into computerized format. But subsequent steps were done using R Script, as the library and documentation were more easily accessible, and the R Script allowed computerized output. In addition, the file 'measures_corr.dta' calculated by STATA could be converted into a format readable by R Script. The full script calculated with R Script can be found in the attachment to this work.

Data

The first step to get a first impression (*Figure 2*) was to show the basic statistics for the whole dataset.

Figure 2: Description of the unprocessed dataset

Data description:

```
obs:     3,332,901
vars:    6                          24 Nov 2021 21:31
```

variable name	storage type	display format	value label	variable label
cons_id	long	%12.0g		Cons_id
Dat_Ro	str19	%19s		Dat_Ro
Op_hrs	int	%8.0g		Op_hrs
p1	str15	%15s		P1
P1_numeric	float	%9.0g		P1
datetime	double	%tc		

Summary:

Variable	Obs	Mean	Std. Dev.	Min	Max
cons_id	3332901	200140.3	43002.18	110101	241548
dat_ro	0				
op_hrs	3332752	16973.64	6312.677	0	28597
p1	0				
P1_numeric	3332748	3403154	2.35e+09	-97.9	1.62e+12
datetime	3332901	1.94e+12	9.90e+09	1.92e+12	1.95e+12

The variable cons_id describes the unique device ID, P1 and P1_numeric which show the measured consumption. Op_hrs is the parameter for the operating hours, date and time, which has already been converted to date and time values stored in dat_ro. The values of P1 and dat_ro were given as string values, which cannot be used for further analysis and had to be converted in the very first step.

By performing data cleaning, outlier handling and Winsorizing, the following results were obtained:

- Detecting unique devices: 312
- Rounding time to full hours and identifying unique moments: 9953
- Determining minimal date und maximum date: Min: 2020-10-01 01:00; Max 2021-11-21 00:00

The dataset contains more than 365 dates with measurements; therefore, the set contains more than 365 days * 24 hours unique data. 288310 duplicate rows have been removed. The next task has been to cross-reference the unique dates and IDs to complete the data frame: 60745 Date/ID Combinations are missing and are filled in with NA in the cross-referenced data table. In addition, negative values and zero values had to be set to NA. The data contain such values which fluctuate since the smart device occasionally started to measure late, or it had to be calibrated later, it was defective or simply measured wrong values due to environmental situations.

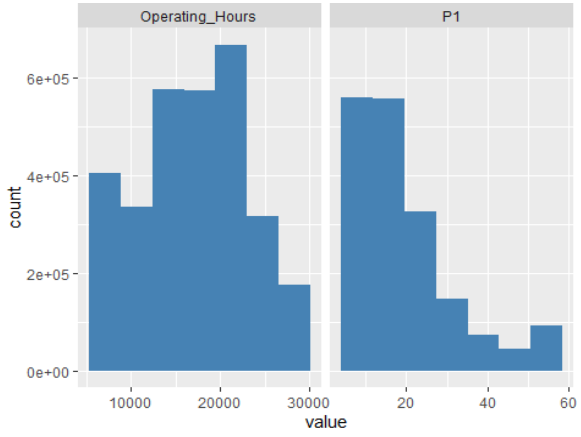
After adjusting outliers to the 95% percentile - metric statistics for non-factor features (*Table 1*).

Table 1: Percentile-metric statistics (computed by R, own representation)

Variable/measure	vars	n	mean	sd	median	trimmed
P1	1	3044437	16807.51	6057.13	17423.0	16948.06
Operating hours	2	1795745	19.24	44531	16.2	17.51
Variable/measure	**min**	**max**	**range**	**skew**	**kurtosis**	**se**
P1	5583.0	26840.0	21257.0	-0.20	-0.88	3.47
Operating hours	4.4	51.3	46.9	1.16	0.73	0.01

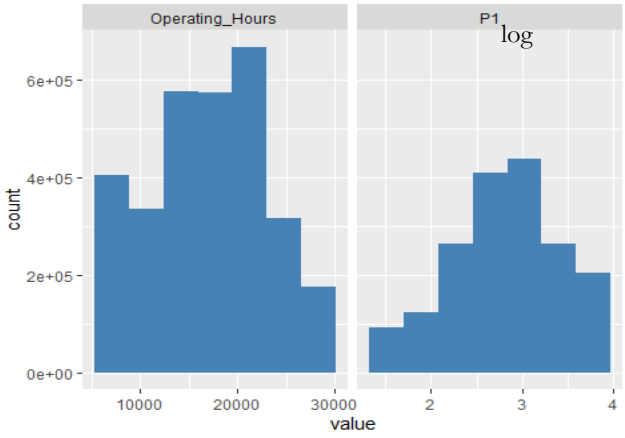
The *mean* and *sd* values of both variables are saved for later re-transformation of cluster centers. Furthermore, the distribution of these features is shown in histograms (*Figure 3* and *Figure 4*).

Figure 3: Distribution of P1 and operating hours untransformed



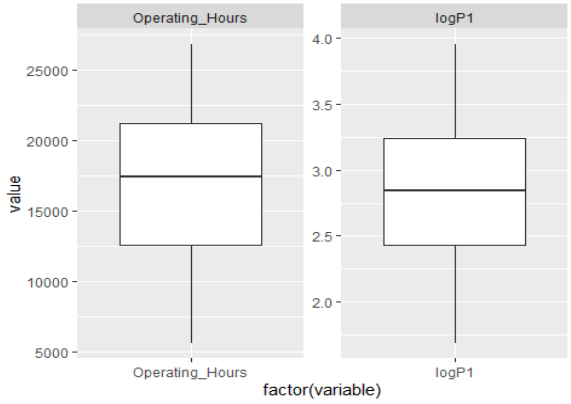
P1 heavily skewed and is therefore log-transformed in the next steps.

Figure 4: Distribution of operating hours and P1 after transformation



The P1 has become normal by now and can be used for modelling. In addition, the column has been renamed to logP1 for easier access and reading. The boxplot (*Figure 5*) also shows the normal distribution of the values used now.

Figure 5: Boxplot of operation hours and transformed P1



Finally, data preparation was almost complete. The last steps included the removal of the lines with missing values (already marked via NA) and melting the data.

RESULTS AND DISCUSSION

Sampling is required using the seed (by setting the seed for subsampling the same output at multiple runs was forced) – a sample of 10 000 values was used for subsampling df to n rows. Sampling was necessary to obtain reproducible results. The initial dataset could not be managed on a PowerBook workstation with a 6-core, 12-thread Xeon CPU at 2.7 GHz and 64 GB RAM. Clustering was performed using a virtual server cluster with 8 CPUs and 1024 GB RAM – but the system was not

available after an initial run. Therefore, we used the sub-sampling approach. The results between the full dataset and sampled dataset are shown and compared below.

To test and adjust the algorithm, a random number of k = 5 was initially used. The detailed data distribution was removed of the script below, only the sum of square was kept for later comparison.

The number of k has not been easy to identify so far, and in addition to the elbow method (*Figure 6*), gap statistics (*Figure 7*) must be used:

Figure 6: Elbow method for finding optimal k

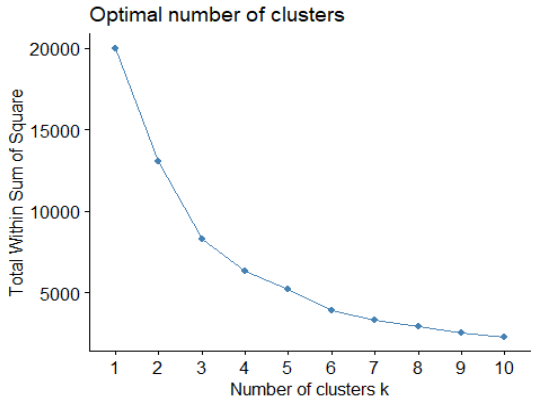
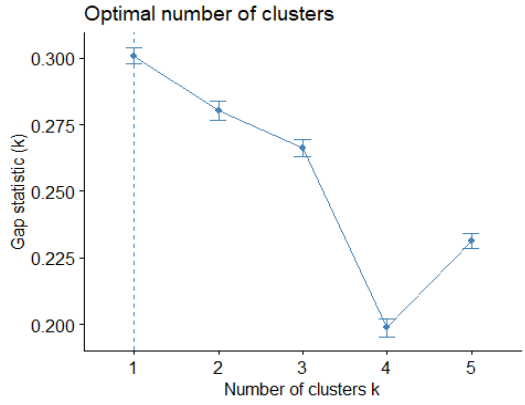


Figure 7: Optimal number of k using the gap statistics



The optimal value of k was calculated and is shown to be 4. So, in the next steps, clustering was performed using k = 4. This indicates that there are 4 clusters – the same result was obtained with the full data set. In term of the number of clusters, no difference was found between the sampled data and the full data set.

To evaluate the results, the cluster centers had to be reconfigurated. The value of the log-transformed demand alone is not sufficient, it does not allow the evaluation of the identified clusters. The cluster centers can be easily determined by using the

appropriate command. However, the previously calculated standard deviation and mean values are needed for the re-transformed centres.

$$RC_k = CV * \sigma + \mu \quad (3)$$

RC_k is the retransformed center value of each cluster k, CV is the value of the center as determined by the K-means algorithm, σ is the standard deviation and μ is the mean. The calculated center values (*Table 2*) can be determined using R.

Table 2: cluster centres (calculated by R, own representation)

cluster	Operating_Hours	logP1
1	-0.5165050	0.9126801
2	1.0258957	0.4999305
3	0.7613733	-0.9752235
4	-1.0512699	-0.6988077

The values for standard deviation and mean have already been calculated within the metric statistics and now the formula is used form the calculation of the P1 values which have already been converted into to logarithmic value, the calculation had to be reversed using the exponential function.

$\mu_{OH} = 16807.51$ $\mu_{P1} = 19.24387$ $\sigma_{OH} = 6057.133$ $\sigma_{P1} = 12.21368$

Comparison between sampled value and complete dataset

Although the sampled data (*Figure 8*) show different colors for the clusters than for the full set of data (*Figure 9*), both have revealed 4 clusters, and clusters are visually located within the same area and close to each other. The above figure is included to demonstrate that the sampled result is very close to the result of the full dataset.

Figure 8: Sampled data set result of K-means analysis

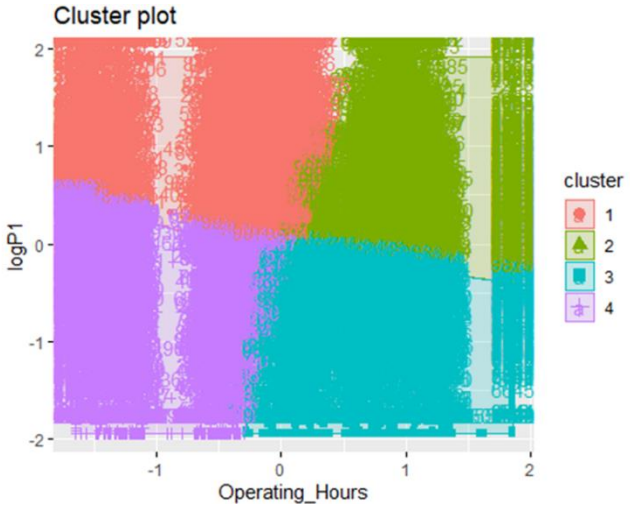
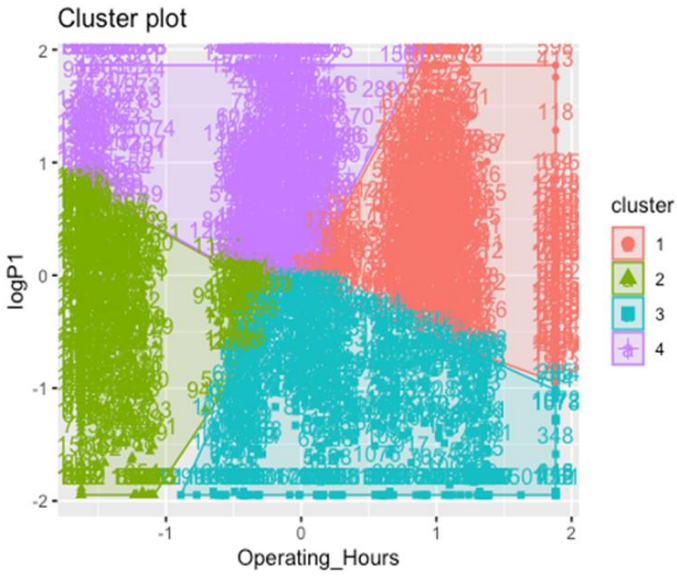


Figure 9: Result of K-means full data set analysis



The *Table 3* shows the clusters for easy comparison. Cluster 4 of the full set is set in comparison with cluster 1 of the sampled set, cluster 1 of the full set and cluster 2 of the sampled set represent the same cluster, cluster 3 is retained for the full set and the sampled set, and cluster 4 of the sampled set is cluster 2 of the full set.

Table 3: final results of transferred cluster values

Sampled data set			Full data set		
Cluster	**Operating hours**	**P1**	**Cluster**	**Operating hours**	**P1**
1	13678.97	37.45434	4	14264.07	40.21686
2	23021.50	27.16575	1	22963.40	24.17497
3	21419.25	11.63607	3	18798.65	10.98849
4	10439.83	13.10257	2	8609.02	15.10806

Without transformation, the data cannot not be used for further analysis. As the hardware demands for the entire data set were huge it is recommended to use a subset. Further tests were carried out with sampled data containing more than 10000 data points, but these showed no new results. The number of clusters and the area of the cluster remained the same. For the data provided, a total of four clusters were identified. One cluster represents rd. 25% of each measured value. The four identified clusters could be used for further analysis or as a basis for a new billing mechanism.

SUMMARY AND CONCLUSION

Of course, further details would need to be analyzed before any action could be taken to use the results discovered above. Such actions could be an adjustment of tariffs or

even the separation of individual buildings. In particular, anonymised data in quartiles should be reviewed. Since heat transfer stations were provided and measured, the houses behind these transfer stations should be identified. In case of the hospital, where higher demand is expected throughout the year – different measurements should be taken than for any single-family house or multi-tenanted building.

In summary, the results are as follows:

- K-means clustering on big data requires high performing hardware;
- Data sampling reveals the same number of clusters;
- Even with data without private information such as building type or addresses clustering can be performed;
- Four clusters were identified in a two-dimensional space;
- Under normal conditions, LOW operating hours with HIGH demand is the least preferrable option (the specific technical details and the inclusion of the Kaposvár sugar factory may positively contribute to turn this finding into an advantage);
- Longer and stable operating hours are the most favourable – with low demand. Low demand with shorter operating hours is also welcome.

Any further steps, such as identifying meters and underlying consumers for cluster #1 from sampled data or possibly performing the test for another year should be part of a follow-up research. These should be considered as shortcomings or gaps as the primary goal was to evaluate and assess the potential for clustering and to identify clusters. Follow-up research could also carry out structuring / data analysis including private data (house size, type of use (private, office)).

The result showed significant heat demand at low operating hours – given the initial time of measurements, this is early in the heating period. The use of waste heat, such as the heat and methane gas produced by the sugar factory could be increased by additional sources. That would be a measure that could be taken without further research into the details of each cluster.

Inter-company waste heat utilization, which means that waste heat that cannot be used within the company can be used by third parties, for example, in commercial or residential buildings. The main challenges of this option are access to reliable data to compare and match waste heat potentials and demands, and the fact that these do not always match. At present, the most economically viable and feasible utilization of waste heat requires spatial proximity between the waste heat source and waste heat demand. Heat recovery or heat displacement are the most efficient and simplest technological approaches to use waste heat to increase overall energy efficiency and cost-effectiveness. Heat exchangers are often used for this purpose. Heat exchangers transfer waste heat to a transport medium, which then transfers the heat to other units. However, losses are incurred in such cases. The transfer of waste heat to third parties requires additional transport infrastructure such as local and district heating pipes, buffer storage etc. The advantage of local and district heating networks is that they have the flexibility to use a large number of different heat sources, which can be both centralized and decentralized. In addition, the heat network draws on different energy sources at different levels and points, whether it is summer or winter. The waste heat generated can thus be economically recovered, extracted and profitably fed into the heating network. As a result,

the company saves on cooling water costs, generates income from the sale of heat energy, and makes significant contribution to reducing CO_2 emissions, as the heat fed in would otherwise have to be produced elsewhere.

Glass is one of the most sustainable packaging materials available, made from natural elements, largely reusable and 100% recyclable. However, its production is energy-intensive, with furnaces operating at temperatures of over 1500 °C, 24 hours a day and 7 days a week, making glass production a significant and major source of waste heat. The auxiliary heat exchanger is typically located upstream of the plant flue gas treatment, so there is no need to lower the temperature - to the temperature supported by the filter - dilute the flue gas with external air or spray it with water (quenching/melting tower). The temperature and the amount of heat that can be recovered for a single production line are often not particularly high. This usually limits the use of recovered heat for power generation with steam turbines, at least not using other fuels to ensure the steam does not overheat. Organic Rankine Cycle (ORC) is an attractive solution to generate electricity from waste heat, even for low power and discontinuous flows of hot gases with temperatures around 300 ºC or even lower. The ORC has a lower sensitivity to temperature and flow rate changes of hot gases, which allows for easier handling and eliminates the need for specialist staff. It has lower operating costs, does not require water treatment or consume water. The waste heat is initially used to produce high-pressure steam or supplied to consumers that require high temperatures. This, in turn, produces waste heat but at lower temperatures. Such waste heat is available as additional waste heat potential, which can be used, for example, for heating products, as feed water, or as boiler water. This leaves low-temperature waste heat below 100 °C, which often has no internal consumers. Instead of disposing of this energy, the best solution is to transfer it to a district or local heating network, which usually operates at fixed temperatures between 70 and 100 degrees Celsius. This is relevant as Şişecam (Şişecam is one of the most influential industrial enterprises in Turkey with a corporate history of more than 85 years. Şişecam was founded to meet Turkey's basic demand for glass products. Today, as one of the country's most powerful industrial conglomerates, Şişecam has become a global player in all key areas of the glass industry, as well as in the soda and chromium compounds businesses.), investing more than €200 million in the construction of a glass packaging plant in Kaposvár, south-west Hungary. The plant, which will be Şişecam's first glass packaging factory in Europe, will have the capacity of producing 330,000 tonnes of glass packaging material a year (*Glass online*, 2022).

The 4 identified heat demand clusters can be used to identify and integrate additional waste heat sources into sustainable heat production for the district heating in Kaposvár. One possible additional source has been mentioned above. But here again, further research and exploration would be needed, and technical feasibility should be investigated.

ACKNOWLEDGEMENT

This research was supported by the District Heating company of Kaposvár by providing the necessary data and the colleagues of Novustat GmbH, who provided insight and expertise to help the research on the details of the R programming.

REFERENCES

Andrews, D. (2009). *"Carbon footprints of various sources of heat - biomass combustion and CHPDH comes out lowest " - William Orchard*. https://claverton-energy.com/carbon-footprints-of-various-sources-of-heat-chpdh-comes-out-lowest.html

Bánkuti, Gy. & Zanatyné Uitz, Zs. (in press) (2021). A good practice in urban energetics in a Hungarian small town, Kaposvár. In M. Fathi, E. Zio, & P. M. Pardalos (Eds.), *Handbook of Smart Energy Systems*. Springer International Publishing. https://doi.org/10.1007/978-3-030-72322-4

Csima, F. & Szendefy, J. (2009). Synergic consideration of competitiveness and ecology in biogas production and use. *Regional and Business Studies*, *1*(1), 45–48.

Fang, H., Xia, J., Zhu, K., Su, Y., & Jiang, Y. (2013). Industrial waste heat utilization for low temperature district heating. *Energy Policy*, *62*, 236–246. https://doi.org/10.1016/j.enpol.2013.06.104

Glass online (2022). *Sisecam to build glass packaging plant in Kaposvár, Hungary*. https://www.glassonline.com/sisecam-to-build-glass-packaging-plant-in-kaposvar-hungary/

IEA (2021). District heat production by region, 2020, and world average carbon intensities of district heat supply in the Net Zero Scenario, 2020-2030. https://www.iea.org/reports/district-heating

Jain. A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Knudsen, B. R., Rohde, D., & Kauko, H. (2021). Thermal energy storage sizing for industrial waste-heat utilization in district heating: A model predictive control approach. *Energy*, *234*, 121200. https://doi.org/10.1016/j.energy.2021.121200

Köfinger, M., Schmidt, R. R., Basciotti, D., Terreros, O., Baldvinsson, I., Mayrhofer, J., Moser, S., Tichler, R., & Pauli, H. (2018). Simulation based evaluation of large scale waste heat utilization in urban district heating networks: Optimized integration and operation of a seasonal storage. *Energy*, *159*, 1161–1174. https://doi.org/10.1016/j.energy.2018.06.192

MacKay, D.J.C. (2003). *Information Theory, Inference, & Learning Algorithms*. Cambridge University Press

MacQueen (1967). Some methods for classification and analysis of multivariate observations. In Lucien M. Le Cam, Jerzy Neyman (Eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Statistical Laboratory of the University of California, Berkeley, 281–297. https://projecteuclid.org/Proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/proceedings-of-the-fifth-berkeley-symposium-on-mathematical-statistics-and/toc/bsmsp/1200512974

Radtke, U. (2022). A Brief Literature Review of Structuring District Heating Data Based on Measured Values. *The Ohio Journal of Science*, *122*(2) 75-84, https://doi.org/10.18061/ojs.v122i2.8845

Ramos, S., Duarte, J. M., Duarte, F. J., & Vale, Z. (2015). A data-mining-based methodology to support MV electricity customers' characterization. *Energy and Buildings*, *91*, 16–25. https://doi.org/10.1016/j.enbuild.2015.01.035

Tureczek, A. M., Nielsen, P. S., Madsen, H., & Brun, A. (2019). Clustering district heat exchange stations using smart meter consumption data. *Energy and Buildings*, *182*, 144–158. https://doi.org/10.1016/j.enbuild.2018.10.009

Wahlroos, M., Pärssinen, M., Rinne, S., Syri, S., & Manner, J. (2018). Future views on waste heat utilization – Case of data centers in Northern Europe. *Renewable and Sustainable Energy Reviews*, *82*, 1749–1764. https://doi.org/10.1016/j.rser.2017.10.058

Woolley, E., Luo, Y., & Simeone, A. (2018). Industrial waste heat recovery: A systematic approach. Sustainable *Energy Technologies and Assessments*, *29*, 50–59. https://doi.org/10.1016/j.seta.2018.07.001

Corresponding author:

Uwe RADTKE
Hungarian University of Agriculture and Life Sciences
Doctoral School in Management and Organizational Sciences
7400 Kaposvár, Guba Sándor u. 40., Hungary
e-mail: weu@gmx.de