Exploratory study on clustering methods to identify electricity use patterns in building sector

Selin Yilmaz¹, Jonathan Chambers¹, Stefano Cozza¹, Martin K. Patel¹

¹ Energy Efficiency Group, Institute for Environmental Sciences and Forel Institute, University of Geneva, Boulevard Carl-Vogt 66, 1206 Geneva, Switzerland

Selin.yilmaz@unige.ch

Abstract. In this paper, we perform a cluster analysis using smart meter electricity demand data from 656 households in Switzerland, collected during one year. First, we present the silhouette analysis to determine the optimum number of clusters for a k-means clustering approach. Secondly, we try different distance functions used in the k-means clustering to partition the samples into different categories. We find that the choice of distance function has no effect on the clustering performance. Finally, we investigate the “dimensionality curse” and find that low dimensions should be preferred to increase the quality of the clustering outcome.

1. Introduction

To meet the ambitious climate targets, 173 countries have established renewable energy targets at the national level, with most of them also adopting related policies [1]. The increasing penetration of renewable energy sources (RES) in power systems intensifies the need of enhancing the demand side management in order to accommodate the uncertain power output of the RES such as wind and solar generation [2], [3]. Therefore, research on electricity demand profiles has received significant interest from researchers and utilities worldwide for demand side management such as demand response programmes. Thanks to the deployment of smart meters, there is an extensive amount of electricity demand data with high resolution (from minutely to hourly). Recently, cluster analysis is increasingly applied to smart meter electricity demand data to identify patterns in electricity consumption in order to improve the demand side management. However, there is no consensus in the literature on the standardisation of clustering approaches. This paper investigates the clustering approach in two aspects. First, we investigate the performance of distance functions used in partitioning the load profiles and to show the impact on the clustering performance. Second, we focus on the dimensionality of the dataset. A major problem encountered when applying machine learning is the so-called “curse of dimensionality”, which refers to the fact that many algorithms that work well in low dimensions become intractable when the input is high-dimensional [4]. For this, we investigate the dimensionality of the dataset by clustering the profiles on different time-resolution (decreasing and increasing the dimensionality) and by quantifying its effect on the clustering performance.
2. Background

2.1. Clustering methods

Table 1 summarises a number of clustering methods which have been applied to household level data. Generally, clustering is classified as an unsupervised machine learning algorithm. Even though hierarchical clustering is widely used, it is restricted to small datasets due to its quadratic computational complexity \[5\]. Non-hierarchical methods such as k-means, self-organising maps (SOMs) are instead very efficient for clustering large datasets. Given its efficiency for large sample size and simple nature of the algorithm, the k-means clustering method is one of the widely used classification techniques for clustering daily load profiles of residential customers. The drawback of this method is that the number of clusters \(k\) needs to be known in advance. Several methods are available to optimise the \(k\) value such as Cluster Distance Performance, Davies–Bouldin index (DBI), Silhouette analysis \[6\].

Table 1. Brief review of clustering methods and reference studies

| Clustering method          | Method description                                                                                                                                                                                                 | Reference studies                      |
|----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------|
| k-means                    | The method starts by choosing \(k\) observations randomly, each of which represents a cluster centre. Then each observation is assigned to the nearest cluster based on a distance function between the cluster centre and the observation. The process is repeated until there is no triggered change in the cluster membership | \[7\]–\[10\]                           |
| Self-organising maps       | A sample vector is selected randomly, and the map of weight vectors is searched to find which weight best represents that sample. The weight that is chosen is rewarded by being able to become more like that randomly selected sample vector. From this step the number of neighbours and how much each weight can learn decreases over time. This whole process is repeated for a large number of times. | \[11\]–\[13\]                          |
| Hierarchical clustering    | It groups a given dataset of load profiles into the required number of clusters through a series of nested partitions. This results in a hierarchy of partitions leading to the final cluster                                                                 | \[6\], \[14\], \[15\]                 |
| Finite mixture models      | Describes the distributions and correlations between instances by automatically choosing the optimal weights for each of the input parameters for each cluster.                                                                 | \[16\]                                 |

2.2. Distance functions

Non-hierarchical cluster analysis partitions a set of observations into subsets (clusters) in a way that objects belonging to the same cluster have high similarity, while objects belonging to different clusters have low similarity. The clusters are established according to a “dissimilarity function” based on distances. There are several definitions of distances and they are used in the clustering process. Here, we analyse the most common ones which are Euclidean, Manhattan, Canberra and Chebyshev. The Euclidean distance is the most commonly used distance function in engineering and physical sciences. It is computed via the root of the squared difference between coordinates of pair of objects. The Manhattan distance measures distances on a rectilinear basis, namely it computes the absolute differences between coordinates of pair of objects. The Canberra distance measures the sum of absolute fractional differences between the features of a pair of data. Finally, the Chebyshev distance measures the maximum value distance and it is computed as the absolute magnitude of the differences between coordinate of objects pair. The mathematical definitions of the formulas can be found in the Scikit-learn library \[17\].
3. Dataset and methods
The dataset used for this work consists of electricity readings (in Watts) from apartments located in West Switzerland. Table 2 gives the description of the most relevant measured variables.

Table 2. Description of the dataset

| Variable             | Description                              |
|----------------------|------------------------------------------|
| Number of households | 656                                      |
| Monitoring period    | 1st January 2015 to 31st December 2015   |
| Temporal resolution  | 15 minutes                               |
| Building type        | Multi-family flats                       |

Firstly, one daily average profile is created for each household by averaging the daily profiles of each household throughout the year which resulted in 656 profiles, one for each household. Then, the shape of the load profiles is defined by normalising the average load profiles. The normalisation is obtained by dividing each measurement in a day by the sum of that day, such that the integral over the yearly-mean daily profile curve is equal to 1.

Secondly, k-means clustering is applied according to the implementation in the Scikit-learn library. The performance of the cluster model is evaluated by the Silhouette score. This metric compares the “distance functions” used based on cluster geometry in terms of compactness (instances in the same cluster have high similarity) and distinctness (instances in different clusters have low similarity) for each observation (or load profile). The silhouette score has a range of [-1, 1], where scores close to +1 indicate that the observation has similar features to the other cluster elements and dissimilar to elements assigned to other clusters. Silhouette score is calculated using formula below:

\[
\text{Silhouette score} = \frac{\max(a,b)}{b-a} \quad (\text{Eq.1})
\]

where:
- \(a\) = average intra-cluster distance,
- \(b\) = average shortest distance to another cluster.

4. Results
4.1. Determination of “number of clusters”
First, we determine the optimum number of clusters by using the Silhouette score. Figure 1 shows the silhouette analysis calculated for varying number of cluster (from \(k= 2\) to \(k=20\)). These plots show the silhouette score for every load profile grouped and coloured by the cluster label. The vertical line shows the average silhouette score for all the observations. This enables visualisation of the distribution of silhouette scores across the clusters, allowing a rapid overview of the relative sizes of each cluster, identification of particularly poorly performing clusters, and verification that overall silhouette scores are not dominated by particularly good/bad scores for a small number of observations. Ideally, each cluster should be characterized by a large positive silhouette score without any negative components. Figure 1 shows the silhouette analysis for k-means clustering for the normalised load curves averaged for number of “k” using Manhattan distance. Silhouette score is the highest when “k” is equal to 3 and the silhouette score decreases as the number of the clusters increase.
Figure 1. Silhouette analysis for k-means clustering using the daily load shape (optimum number of clusters)

4.2. Comparison of distance functions used
In this section, we investigate whether the distance function has an effect on the clustering results. Figure 2 shows the silhouette analysis for k-means clustering using difference distances explained in Section 2.2. For all the distance functions used, k =3 had the highest Silhouette score, hence the Silhouette scores for k=3 is shown in Figure 2. The analysis show that the chosen distance has no effect on the clustering performance. The average score has not changed for different distances defined.
Manhattan distance (average silhouette score = 0.186)  
Euclidean distance (average silhouette score = 0.186)  
Canberra distance (average silhouette score = 0.188)  
Chebyshev distance (average silhouette score = 0.185)

Figure 2. Silhouette analysis for k-means clustering using different distances existing in literature

4.3. Dimensionality
One approach for dimensionality reduction with regard to the temporal dimension is to average values from 15-minute data to hourly frequency (decrease from 96 dimensions to 24) and further to 6-hourly data (dimensionality = 4). Figure 3 shows the change in silhouette score of the clustering for different dimensions. It can be seen that the average silhouette score increases as the values are getting closer to the average values of the cluster (by lowering the dimensionality).

Figure 3. Average silhouette score calculated for averaged values of data by 15-minutely to 6 hourly data, for k=3 clusters.
5. Conclusion

This paper presents an analysis of clustering approaches for grouping the electricity demand profiles for households in Switzerland. The clustering analysis was applied to average household electricity profiles and daily electricity profiles of 656 multi-family flats in Switzerland. The results show that distance function such as “Euclidean”, “Manhattan”, “Canberra” and “Chebyshev” do not change the performance of the clustering outcome. On the other hand, dimensionality of the dataset can affect the clustering performance, as expected. The simulation results show that the quality of the clusters (silhouette scores) increases as the dimension of the dataset decreases. Therefore, it is recommended that a trade-off should be found between using lower dimensions to tackle the challenge of “curse of dimensionality” and sufficient number of features to define the shape of the load curve.

References

[1] IRENA, “REthinking Energy,” 2017.
[2] IEA “Implementing Agreement on Demand-Side Management Technologies and Programmes,” 2018.
[3] P. Grunewald and M. Diakonova, “Flexibility, dynamism and diversity in energy supply and demand: A critical review,” Energy Res. Soc. Sci., vol. 38, no. January, pp. 58–66, 2018.
[4] R. Bellman, Bellman Richard. Adaptive Control Processes. Princeton University Press, Princeton, NJ. 1961.
[5] J. Shieh and E. Keogh, “i SAX,” 14th ACM SIGKDD Int. Conf. Knowl. Discov. data Min., vol. KDD ’08, p. 623, 2008.
[6] S. Pan et al., “Cluster analysis for occupant-behavior based electricity load patterns in buildings: A case study in Shanghai residences,” Build. Simul., vol. 10, no. 6, pp. 889–898, 2017.
[7] I. Benítez, A. Quijano, J. L. Díez, and I. Delgado, “Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers,” Int. J. Electr. Power Energy Syst., vol. 55, pp. 437–448, 2014.
[8] J. Kwac, J. Flora, and R. Rajagopal, “Household energy consumption segmentation using hourly data,” IEEE Trans. Smart Grid, vol. 5, no. 1, pp. 420–430, 2014.
[9] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, “Clustering analysis of residential electricity demand profiles,” Appl. Energy, vol. 135, pp. 461–471, 2014.
[10] A. Al-Wakeel, J. Wu, and N. Jenkins, “K -Means Based Load Estimation of Domestic Smart Meter Measurements,” Appl. Energy, vol. 194, pp. 333–342, 2017.
[11] G. Chicco, R. Napoli, and F. Pighione, “Comparisons among clustering techniques for electricity customer classification,” IEEE Trans. Power Syst., vol. 21, no. 2, pp. 933–940, 2006.
[12] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, “An electric energy consumer characterization framework based on data mining techniques,” IEEE Trans. Power Syst., vol. 20, no. 2, pp. 596–602, 2005.
[13] L. Hernández, C. Baladrón, J. M. Aguiar, B. Carro, and A. Sánchez-Esguevillas, “Classification and clustering of electricity demand patterns in industrial parks,” Energies, vol. 5, no. 12, pp. 5215–5228, 2012.
[14] G. J. Tsekouras, N. D. Hatzigiorgiou, and E. N. Dialynas, “Two-stage pattern recognition of load curves for classification of electricity customers,” IEEE Trans. Power Syst., vol. 22, no. 3, pp. 1120–1128, 2007.
[15] R. Mena, M. Hennebel, Y. F. Li, and E. Zio, “Self-adaptable hierarchical clustering analysis and differential evolution for optimal integration of renewable distributed generation,” Appl. Energy, vol. 133, pp. 388–402, 2014.
[16] S. Haben, C. Singleton, and P. Grindrod, “Analysis and clustering of residential customers energy behavioral demand using smart meter data,” IEEE Trans. Smart Grid, vol. 7, no. 1, pp. 136–144, 2016.
[17] Scikit-learn (2019). Distance metric: Available from: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html
[18] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” vol. 12, pp. 2825–2830, 2012.