Detection of Appliance Utilization Patterns via Dimensionality Reduction

Fernanda Villar[1], Luiz Carlos Pereira da Silva[1], Pedro Henrique Juliano Nardelli[2,3] and Hader Hazini[1]
[1] School of Electrical and Computer Engineering, State University of Campinas, Brazil
[2] School of Energy Systems, LUT University, Finland
[3] Centre for Wireless Communications, University of Oulu, Finland

***Abstract*—This paper focuses on the detection of utilization patterns in electricity residential consumption, which are closely related to the occupant characteristics (e.g. number, age, occupancy, and social class). Our goal is to identify groups of appliances that are often used together via their statistically relatedness. This relation might be obvious (as in TV and Home Theater), or not. The results can be used, for example, to guide a recommendations letter from the energy supplier to the final user, suggesting specific change of habits in order to improve the residence's energy efficiency. We propose here a methodology for identifying patterns from a large sets of system status, which is a computationally hard task defined in $\mathbb{R}^n$ with n being the number of appliances involved. The approach consist in the following steps: (i) the Principal Component Analysis method is employed to reduce the set dimensionality to $\mathbb{R}^3$ with explained variance from 68% to 90% to guarantee minimum information loses, (ii) the k-means method to clustering appliances into different groups and (iii) the elbow method was used to define the best number of clusters for each house with explained variance of at least 93% and reaching more than 99% for the best k. Numerical tests using the UK-DALE dataset are presented to show the effectiveness of the proposed solution. The main contribution of this work is a method with low computational cost that requires no other information than a large set of reliable system status (binary vectors) to reveal utilization patterns in the residence.**

***Index Terms*—data mining, dimensionality reduction, k-means, Principal Component Analysis, pattern recognition**

I. Introduction

The efficiency use of energy is an effective way to improve the utilization of limited resources but it poses different challenges. Modulating loads' consumption, for instance, is a way to achieve energy efficiency by avoiding undesired peaks, which ultimately determine the generation and transmission systems design. However, switching the daily demand peak is not enough and other strategies to improve energy conservation is desirable. With the advancements in Information and Communication Technologies (ICTs), the processing of such a large amount of data has become feasible [1].

Patterns detection in electricity consumption is an intuitive example in this context: it is now possible to collect a huge amount of measurements to serve as the basis of different big data methods. Several studies have been performed on the demand side attempting to characterize and classify the consumers to improve the efficiency [2], [3], [4] and [5]. In [8], for example, eighteen characteristics from the household and its occupants have been inferred with accuracy levels that reached over 80%.

Besides, there are several available databases with actual measurements that can be employed in research, as in [6], [7], [8]. With a detailed utilization patterns detection, the consumer could receive a periodic feedback letter attached to the energy bill with suggestions to improve the residence's energy efficiency, and this can be a motivation factor so more people would allow the residence to be monitored.

The present work follows this line by focusing on the detection of different appliances' utilization patterns using dimensionality reduction and k-means clustering method [9]. Residential energy consumption (when heating/cooling is not electrified) is closely related to occupants daily routines, which usually have utilization patterns. Except by some unusual moments (guests in the house or travelling), the electricity demand behavior shows patterns that are repeated daily, weekly and monthly. These patterns, once known, can be used to identify possible optimization in the appliances utilization. In specific terms, we will focus on a relatively large set of system states and use already established dimensionality reduction and clustering algorithms to find utilization patterns represented by groups of appliances statistically related (meaning that they used at the same time).

Our main contribution here is the proposed methodology for utilization patterns detection on residential installations by extracting groups of appliances that are often (or always) used together. The appliances inside each group does not necessarily have to be used fin combination (e.g., video-game and TV); they are rather statistically related indicating that they are frequently used at the same time. The rest of this paper is divided as follows: Section II revisits some important related work. The details of algorithms employed here are explained in Section III. The group definition is presented in Section IV together with a discussion of how this information may serve as an input for a detailed feedback regarding the energy consumption.

II. Related work

Different researchers have recently focused on improving the demand side load model to implement fast and effective energy efficiency programs. In this sense, consumers may be classified according to their individual energy-saving potential. This allows energy efficiency programs to become more targeted [10], [11]. In the European Union, the energy savings obligations program has already demonstrated positive results

[12]; demand side consumption modulation and malfunction prevention with fast detection are indeed urgent necessities.

With the rising popularity of smart meters, collecting good quality data have become easier and cheaper. Nowadays there are available several datasets containing aggregated measurements taken in different sampling periods (from hours to seconds). These datasets also have different sort of information like individual measurements of appliances (or channels), type of appliance and rated power, social information of the house occupants. We can cite the following examples: UK-Dommestic Appliance Level Energy (UK-DALE) [6], The Reference Energy Disaggregation Dataset (REDD) [7], and BLUED [8].

At the same time, the popularization of machine learning methods and data mining algorithms makes now possible to carry out studies using the aforementioned datasets to both recognize utilization patterns and classify users. In [2], the authors studied consumption patterns based on peak positions via two clustering techniques, namely k-means and self-organizing maps. In [3], a method to find the "main activity performed inside the house" was proposed by using non-intrusive load monitoring based only on the total power consumption; this method reached a remarkable accuracy of more than 80% for most of the user's activities. In [5] a classifier was trained to detect household characteristics like size of the family, age and even employment and retirement status only observing the total consumption. Their accuracy reached more than 80% in some specific cases.

III. Methodology

We propose here a methodology for detecting utilization patterns in the use of different appliances in households based on an existing dataset based on actual electricity demand. Our goal is to find patterns that represent groups of appliances that are statistically related to each other. Note that they must not need to be used for the same purpose, but they are rather often used at the same time.

The dataset selected is in this study is the UK-DALE [6]. It contains measurements from five different households in United Kingdom with 6-second granularity for periods of more than a month. The measurements contain the individual consumption of 52, 18, 4, 5 and 24 individual channels (each channel can be one single appliance or a group of them). This leads to system state domain of high dimension: $\mathbb{R}^{52}$, $\mathbb{R}^{18}$, $\mathbb{R}^4$, $\mathbb{R}^5$ and $\mathbb{R}^{24}$, respectively. Considering one sample every 6 seconds, the set of system status in one month has more than 400,000 measurements. In addition, defining the system status as a binary vector that indicates each appliance's status (i.e. on or off, or 1 or 0), the number of possible status for each house is 2^{52}, 2^{18}, 2^4, 2^5 and 2^{24}. These numbers makes traditional statistical analysis limited. To address this problem it will be performed a dimensionality reduction to make the problem more tractable to extract (useful) information. Once the system status is in $\mathbb{R}^2$ or$\mathbb{R}^3$, a clustering step will define the group of appliances whose use are related to each other.

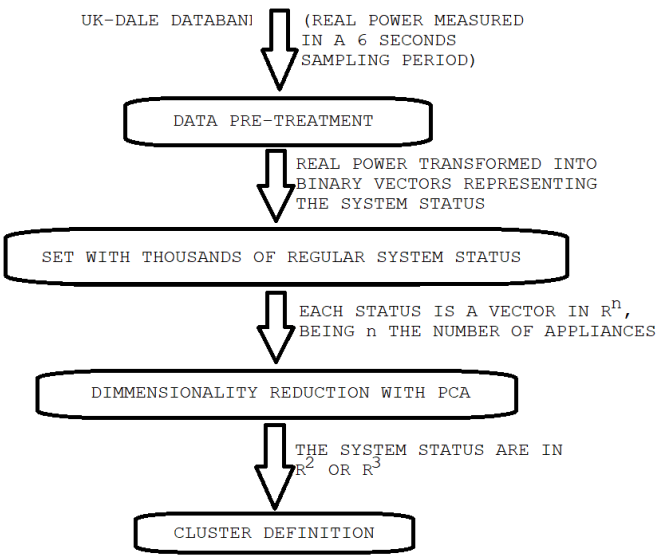

Fig. 1: Proposed method combining dimentionality reduction via PCA and clustering via k-means

To perform the dimensionality reduction, the method of Principal Component Analysis (PCA) [13] was chosen. PCA is a established linear and non-parametric algorithm that gives the directions for projecting the data to maximize the variance (quantity of information) while minimizing noise and redundancy. The method can be considered reliable and simple.

Figure 1 shows the proposed approach with its main steps. The pre-treatment step was responsible for transforming real power and timestamp measures of each individual channel into a binary vector with the status of each appliance. This includes synchronization of the individual channels by removing periods of sampling failure, removal of individual channels that contain quantity of samples too different from the others in the household and definition of the real-power consumption threshold that defines the appliance status as on/1 or off/0.

A. Principal Component Analysis

Principal Component Analysis (PCA) is a linear and non-parametric method that maps a high dimension set of data by projecting it into specific directions that lead variance maximization, and noise and redundancy minimization. It is a established and simple method for extracting relevant data from noisy or confusing data sets (which can be large or not). In the following, we will provide a brief description of PCA (for more details about the method, refer to [9]).

The central question that PCA is based: How to define a matrix P that multiplies by the samples matrix X to return a matrix Y that captures core features of a given phenomenon? Mathematically, we have: $Y = PX$, where X is the $m \times n$ matrix of samples, with m being reated to individual appliances and n the sample in time; P is the linear transformation matrix; Y is the resulting projection of data X.

The multiplication of the samples matrix X per the linear transformation matrix P results simply in a rotation and re scaling of X. Considering (a) p_i as the line vector of matrix P for $i = 1, \ldots, m$, and (b) x_j as the column vector (samples) of matrix X for $j = 1, \ldots, n$, then each line of the resulting matrix Y can be written as: acha

$$y_i = [p_i x_1 \;\; p_i x_2 \;\; \ldots \;\; p_i x_m], \tag{1}$$

which is the projection of the columns of X in the directions represented by each p_i.

The next question is: which are the directions that best represents the interesting data in X? Now assuming that the direction that contains the largest data variance is the one that keeps the better portion of information, the question can be written as: which rotation in the orthonormal basis used to express the measurements in X will result in data with largest variance? Let us define here the covariance matrix C_X as:

$$C_X = \frac{1}{(n-1)XX^T}, \tag{2}$$

where $1/(n-1)$ is a normalizing term.

The matrix C_X is symmetric and contains the variance of the measurement types of X on the diagonal, and the off-diagonal terms contains the covariances between then. As the variance expresses the amount of signal contained, and the covariance measures the redundancy between two measurement types, the objective of PCA is to find the matrix P such that the covariance matrix of $Y = PX$ is diagonal (maximize variance and minimize covariance).

A very simple algorithm for PCA is described by [13] as:

- Find a normalized direction vector in the m-dimensional space along which the variance of X is maximized. This is p_1.
- Find other direction vector in the m-dimensional space, orthonormal to p_1, that also maximizes the variance of X. This is p_2.
- Repeat the procedure, finding vectors that are orthonormal to all the previously selected, until0 m directions were found.

The main advantages of PCA are given next. First, the method does not require any parametrization; that means that no additional information of the phenomenon is required. Second, the variances associated to each direction p_i can be interpreted as a measure of "how principal the component is"; the method lists the components ordered from the largest variance to the lowest. Third, if the first 2 or 3 components together add a high percentage of the total variance (about 60%), this means that the 2D or 3D projection of the original data are enough for capturing the data dynamics, turning the PCA into a dimensionality reduction tool.

B. Appliyng PCA to UK-DALE dataset

The UK-DALE dataset [6] is composed by five measurement sets of houses in United Kingdom with 52, 18, 4, 5 and 24 individual channels. They are collected every 6 seconds for a period of more than 4 years, 193 days, 35 days , 151 days and 122 days, respectively. At each measured instant, the system status is a point in: $\mathbb{R}^{52}$, $\mathbb{R}^{18}$, $\mathbb{R}^{4}$, $\mathbb{R}^{5}$ and $\mathbb{R}^{24}$. At this sampling frequency, it is easy to see that the set of system status is very redundant since every specific status stays unchanged for several samples. As the samples are made in real houses (instead of generated from a simulator), the presence of noise in the data is almost certain. Noisy measurements may be eliminated when the real power value of the individual channels is mapped into a bit status (0 for off and 1 for on). Considering also that there are 14,400 samples a day (if the meters don't fail at any moment), we have to deal with over 504,000 samples for a 35 day set. Even for the smaller house (with 4 individual channels), it is unfeasible to identify the behavior patterns in the raw data.

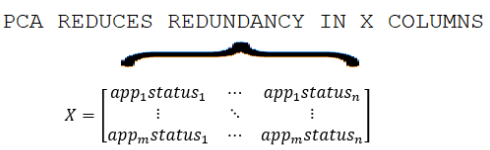

Fig. 2: Illustration of matrix X for redundancy reduction

Referring to the method explanation above, the matrix X contains the individual appliances in the columns, and the status in the lines. In this way, the dimensionality reduction will not lead to loss of reference in the appliances (see Figure 2). The PCA revealed the dynamics involved in the residential appliances operation patterns, identifying the appliances statistically related. The results are presented in Section IV.

C. Minimal Spanning Tree and k-means

After PCA performed the dimensionality reduction, the projection over the first three principal components were used to define clusters. For helping the visualization, we have used the Minimum Spanning Tree [14] for a 3D representation of the PCA results. The Minimal Spanning Tree (MSP) is a tool from graph theory that gives the shortest path between any two nodes (points) of a connected set. This representation makes possible to see a 3D result in a 2D figure (see figures 4, 6 and 8). The method used to generate the MSP was PRIM [3].

The cluster definitions were made according to k-means method [14]. Simulations were made using MATLAB R2016a. To define the number of clusters (the ideal k), it was used the elbow method that varies the number of clusters from 1 to m: 1 cluster means all the appliances together while m clusters means one cluster to each appliance. Every time one new cluster is created, the percentage of variance explained, i.e. the ratio of the between-clusters variance to the total variance, is quantified. There is a value of k from which the addition of one more cluster does not increase much of the variance explained (as illustrated by the elbow curve k vs. %variance explained presented in Figures 3, 4 and 5). The selection of k comes directly from it.

IV. Numerical results and discussions

During the pre-treatment phase, houses 3 and 4 were excluded from the present analysis because they have too little individual channels and the results were trivial. Figures 3, 4 and 5 show the minimal Spanning Tree resulting from the output of the first three principal components of the status set for House 1, 2 and 5, respectively. The appliances with distances smaller than one were represented as a single node. The polygons in each figure represent the clusters resulting from appliance of k-means algorithm (with k selected using the elbow method).

A. House 1

In House 1, the three first principal components represent a explained variance of 68,81%, much lower than the in Houses 2 and 5. Nevertheless it is still a good result that validates the dimensionality reduction. This difference was expected because the number of individual channels monitored is more than twice the quantity for the other houses.

After applying the elbow method for selecting the best number of clusters, the selection of $k = 6$ resulted in a percentage of variance explained of 93,46%. About the clusters showed in Figure 3, the larger group with all the main appliances regarding rated power as soldering iron, dishwasher and washing machine indicates a good potential for load modulation. The appliances represented as a single point are also very interesting results, like soldering iron plus kettle and can be used as guide for a detailed user's feedback. From the clustering results it can also be inferred a minimum number of people living in the house looking at the type of appliances in the same group. For example, it would be very unlike that a single person would use several kitchen appliances and the soldering iron at the same time, so there should be at least 2 people. Through the same logic, including the lights operation observation suggests that there are people in the office and living room during the busy periods.

B. House 2

The application of PCA in House 2 was very effective. The first three principal components together have a explained variance of 90,18%, confirming that the method is a reliable tool for this dimensionality reduction. Applying the elbow method for selecting the best number of clusters, the selection of $k = 3$ resulted in a percentage of variance explained of 99,04%. About the clusters composition (see Figure 4), some of them are obvious, like "modem plus server plus router", but the washing machine together with the PlayStation and the other kitchen appliances reveal some habits of the household population that can exploited. Also, it can be inferred from this result that there are probably three people in the house at that specific moment: one in the kitchen, one using the running machine and one playing video-game.

C. House 5

The application of PCA in House 5 was also very effective. The 3 first components represent a explained variance of 90,97%. The elbow method resulted in the selection of $k = 4$ with percentage of variance explained of 99,04%. About the clusters (see Figure 5), the group with the office appliances (desktop plus sky HD box plus core2 server and others) make a lot of sense; but the group with the kitchen appliances together with hairdryer, steam iron and washer dryer suggests that the house has a good load modulation potential. In this case a feedback suggesting the use of some of these appliances in a different period of the day could be interesting. Regarding the number of occupants in the house, it seems that there are at least two during the busiest moment in the house, one using the kitchen and other playing the video-game (PS4).

D. Discussions

The results presented for the three houses indicate that there are indeed a relation between which kind of appliances are used together in a household. This relation can be expected or not. It also reflects the house occupancy level, which can be also inferred from the data. Besides, the visualization in clusters used here can be useful for both understanding the users' behavior and test the efficacy of interventions to change consumption patterns at houshold level.

In the practical field, the utilization patterns recognition can be offered as a reward for user's who allow the house to me monitored. During the clusters analisys, the devices that can be programmed to work in any time of the day (dish washer and washing machine for example) are a good option for the suggestion of habits modification. Nevertheless, while this framework shows its potential, it still requires further analysis using other datasets.

V. Conclusions and future work

Dimensionality reduction and clustering proved to be capable of revealing utilization patterns for appliances in residential installations. The PCA method performed dimensionality reduction keeping high values of variance explained for the cases under consideration.

The k-means algorithm together with the elbow method for selecting the best k lead to results with corresponding explained variance of more than 93% for the three households analyzed here. The appliances groups resulting from k-means need to be analyzed individually.

This is an interesting tool to understand utilization habits and potential for load modulation or energy efficiency improvements. The results can be used to compose a feedback letter for the user, with some recommendations regarding the use of main appliances (in particular the ones that can be programmed to work at any time of the day). By adding some other information about the house ()The results can be also used to estimate a minimum number of people in the house in a certain period , and detect potential for load modulation that can be included in the users' feedback.

We plan as a future work to extend this method by using other tools that are non-linear and/or allows for some level of parametrization (e.g. Kohonen maps) and apply for a daabanks with a larger number of residences.

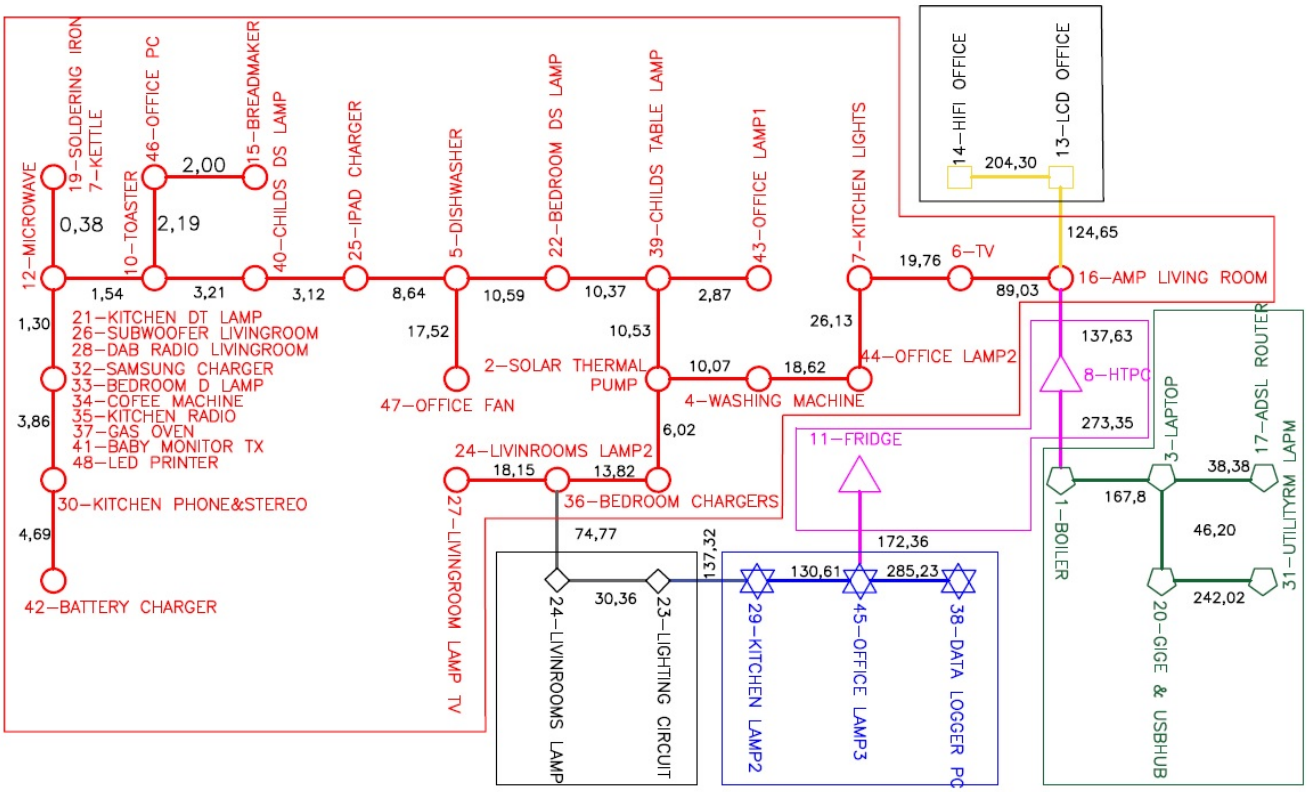

Fig. 3: Minimum Spanning Tree for House 1, with six clusters defined by k-means method.

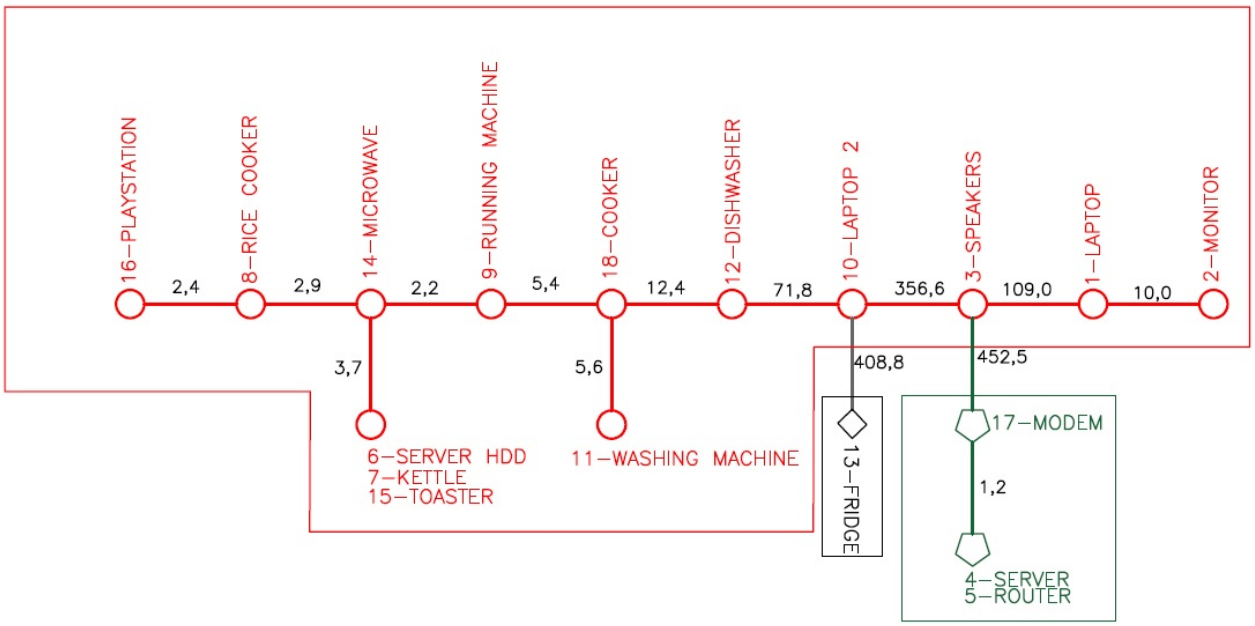

Fig. 4: Minimum Spanning Tree for House 2 with three clusters defined by k-means method.

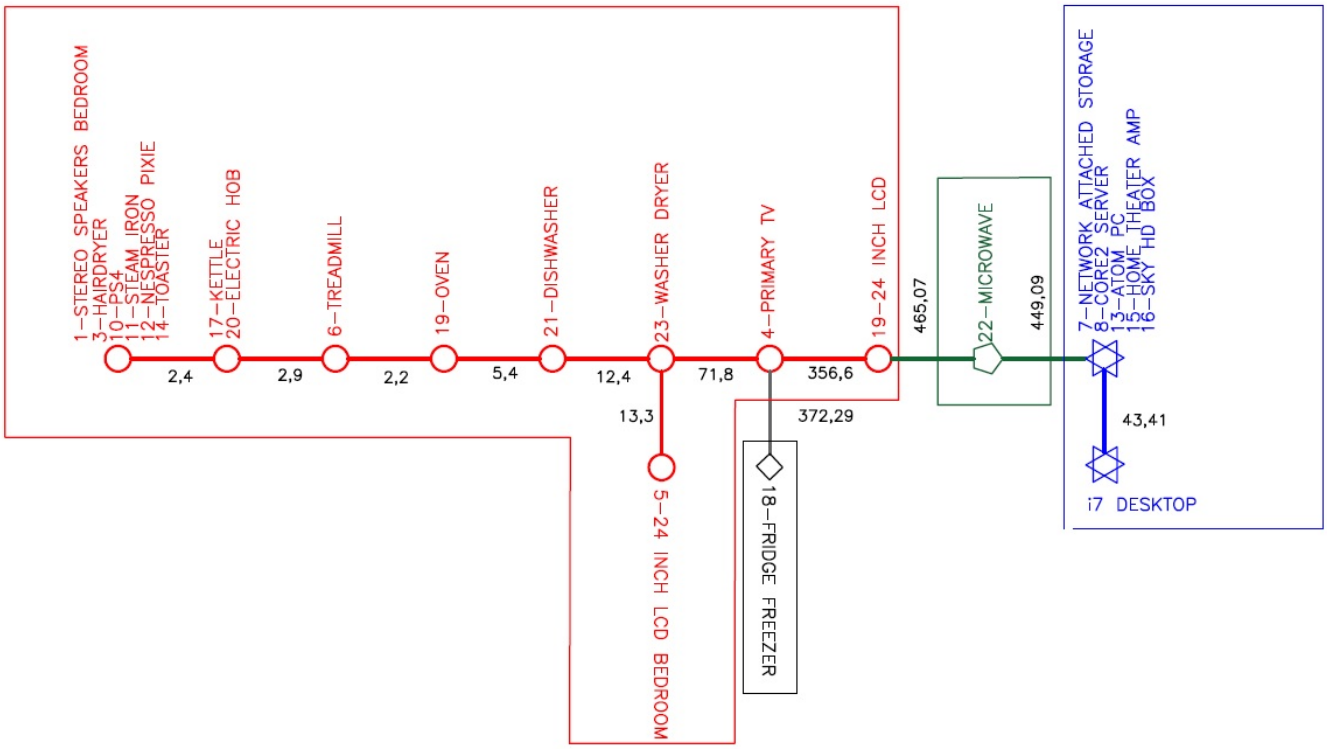

Fig. 5: Minimum Spanning Tree for House 5 with four clusters defined by k-means method.

ACKNOWLEDGEMENTS

This work is partly funded by Academy of Finland via ee-IoT (ICT2023/n.319009), BCDC Energy (SRC/n.292854), and 6Genesis Flagship (n. 318927).

REFERENCES

[1] H. Allcott, "Social norms and energy conservation," *Journal of public Economics*, vol. 95, no. 9-10, pp. 1082–1095, 2011.

[2] H.-Â. Cao, C. Beckel, and T. Staake, "Are domestic load profiles stable over time? an attempt to identify target households for demand side management campaigns," in *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2013, pp. 4733–4738.

[3] Y.-C. Chen, C.-M. Chu, S.-L. Tsao, and T.-C. Tsai, "Detecting users' behaviors based on nonintrusive load monitoring technologies," in *2013 10th IEEE International Conference on Networking, Sensing and Control (ICNSC)*. IEEE, 2013, pp. 804–809.

[4] S. Iyengar, D. Irwin, and P. Shenoy, "Non-intrusive model derivation: automated modeling of residential electrical loads," in *Proceedings of the Seventh International Conference on Future Energy Systems*. ACM, 2016, p. 2.

[5] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, 2014.

[6] J. Kelly and W. Knottenbelt, "The uk-dale dataset, domestic appliance-level electricity demand and whole-house demand from five uk homes," *Scientific data*, vol. 2, p. 150007, 2015.

[7] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," in *Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA*, vol. 25, no. Citeseer, 2011, pp. 59–62.

[8] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, and M. Berges, "Blued: A fully labeled public dataset for event-based non-intrusive load monitoring research," in *Proceedings of the 2nd KDD workshop on data mining applications in sustainability (SustKDD)*, 2012, pp. 1–5.

[9] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," 1988.

[10] I. Ayres, S. Raseman, and A. Shih, "Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage," *The Journal of Law, Economics, and Organization*, vol. 29, no. 5, pp. 992–1022, 2013.

[11] N. J. Goldstein, R. B. Cialdini, and V. Griskevicius, "A room with a viewpoint: Using social norms to motivate environmental conservation in hotels," *Journal of consumer Research*, vol. 35, no. 3, pp. 472–482, 2008.

[12] S. Rezessy and P. Bertoldi, "Energy supplier obligations and white certificate schemes: Comparative analysis of results in the european union."

[13] A. Jonathon Shlens, "Tutorial on principal component analysis http://www. brainmapping. org/nitp/pna," *Readings/pca. pdf*, vol. 12, p. 10, 2005.

[14] D. Cheriton and R. E. Tarjan, "Finding minimum spanning trees," *SIAM Journal on Computing*, vol. 5, no. 4, pp. 724–742, 1976.