



UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

FERNANDA SPADA VILLAR

**DETECTION OF APPLIANCES UTILIZATION PATTERNS IN RESIDENTIAL IN-
STALATIONS USING DIMMENSIONALITY REDUCTION TECHNIQUES**

**DETECÇÃO DE PADRÕES DE UTILIZAÇÃO DE EQUIPAMENTOS EM INSTALA-
ÇÕES RESIDENCIAIS VIA TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE**

CAMPINAS

2021

FERNANDA SPADA VILLAR

DETECTION OF APPLIANCES UTILIZATION PATTERNS IN RESIDENTIAL INSTALLATIONS USING DIMENSIONALITY REDUCTION TECHNIQUES

DETECÇÃO DE PADRÕES DE UTILIZAÇÃO DE EQUIPAMENTOS EM INSTALAÇÕES RESIDENCIAIS VIA TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE

Thesis presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Electrical Engineering, in the area of Electrical Energy.

Tese apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Engenharia Elétrica, na área de Energia Elétrica.

Supervisor/Orientador: Professor Dr. Luiz Carlos Pereira da Silva

Co-supervisor/Coorientador: Professor Dr. Pedro Henrique Juliano Nardelli

Este trabalho corresponde à versão final da tese defendida pelo aluno Fernanda Spada Villar, orientada pelo Prof. Dr. Luiz Carlos Pereira da Silva e coorientada pelo Prof. Dr. Pedro Henrique Juliano Nardelli.

CAMPINAS

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

V71m Villar, Fernanda Spada, 1984-
Detection of appliances utilization patterns in residential installations using
dimensionality reduction techniques / Fernanda Spada Villar. – Campinas, SP
: [s.n.], 2021.

Orientador: Luiz Carlos Pereira da Silva.
Coorientador: Pedro Henrique Juliano Nardelli.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de
Engenharia Elétrica e de Computação.

1. Reconhecimento de padrões. 2. Sistemas de reconhecimento de padrões.
3. Mineração de dados (Computação). 4. Algoritmo K-means. 5. Fusão de
classificadores. I. Silva, Luiz Carlos Pereira da, 1972-. II. Nardelli, Pedro
Henrique Juliano, 1984-. III. Universidade Estadual de Campinas. Faculdade
de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Método de detecção de padrões de utilização de equipamentos
em instalações residenciais via técnicas de redução de dimensionalidade

Palavras-chave em inglês:

Patterns recognition
Patterns recognition systems
Data mining
k-means algorithm
Classifiers fusion

Área de concentração: Energia Elétrica

Titulação: Doutora em Engenharia Elétrica

Banca examinadora:

Luiz Carlos Pereira da Silva [Orientador]
Madson Cortes de Almeida
Zita Maria Almeida do Vale
Pedro Pablo Vergara Barrios
Enes Golçalves Marra

Data de defesa: 05-11-2021

Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA –TESE DE DOUTORADO

Candidato(a): Fernanda Spada Villar RA: 016074

Data da defesa: 05 de novembro de 2021

Título da Tese: “Detecção de padrões de utilização de equipamentos em instalações residenciais via técnicas de redução de dimensionalidade”

Prof. Dr. Luiz Carlos Pereira da Silva (Presidente)

Profa. Dra. Zita Maria Almeida do Vale

Prof. Dr. Pedro Pablo Vergara Barrios

Prof. Dr. Enes Gonçalves Marra

Prof. Dr. Madson Cortes de Almeida

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Ao meu marido André e aos meus filhos Gabriel e Sofia.

ACKNOWLEDGEMENTS

I express my deep sense of gratitude to professor Dr. Luiz Carlos Pereira da Silva, for all the moments that he accepted me at Unicamp, and for the inspirations and orientations during the realization of this work. Also, to professor, and most of all close friend for so many years, Dr. Pedro Henrique Juliano Nardelli for his intense participation in papers reviews, Latex lessons, and lots of patience.

This thesis would not be possible without the participation of professor Dr. Renan Cipriano Moiola, my close friend for almost 20 years, and his google Colab and Python lessons, and for his distinguished participation in the publications we made.

I acknowledge with thanks to Dr. Arun Narayanan, for all the inspiration and participation in all the endless paper review meeting we made. Also, for the contributions for the main organization of the work, having his experience as a collaboration was very valuable.

I am very much thankful to my son and daughter, that tolerated my absence in so many moments while I was working in this thesis, and for my husband for being supportive all the time.

My most sincere thanks to my parents, for their proud looking eyes gave me the energy to keep moving forward.

Finally, I thank everyone who has crossed my path over these almost 6 years, supporting me on this or other fronts so that I could always take one more step forward.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

RESUMO

A utilização de medidores inteligentes de energia com funcionalidades além da simples medição de consumo está se tornando cada dia mais comum pelo mundo todo. Como resultado, medições de consumo no lado da carga estão disponíveis em diferentes frequências de amostragem. Diversos métodos foram propostos para inferir características de uso de equipamentos eletroeletrônicos domésticos a partir de medições de potência. Entretanto, muitas técnicas são baseadas em métodos que requerem alto custo computacional. Ainda, frequentemente necessitam de informações fornecidas por parte dos habitantes da residência. Neste trabalho é proposta uma técnica para detecção de padrões de utilização de equipamentos utilizando algoritmos de baixo custo computacional e que não requerem nenhuma informação dos moradores. Os padrões de utilização são identificados a partir do comportamento do status do sistema, representados por um grande conjunto de vetores binários contendo o status de cada um dos dispositivos monitorados, através de algoritmos de redução de dimensionalidade e clusterização. Os algoritmos de Análise de Componentes Principais, k-means e a determinação do número ótimo de clusters pelo método “*elbow*” são utilizados para definição dos clusters, e o conceito de árvore geradora mínima é utilizado para visualização dos resultados. Em paralelo, Mapas Auto-organizáveis são utilizados para criar um classificador de status. A metodologia foi aplicada em dois bancos de dados públicos com medições reais de residências de dois países diferentes: Reino Unido (UK-DALE) e Estados Unidos (REDD), mostrando diferentes padrões de utilização. As técnicas de clusterização possibilitam a gestão pelo lado da demanda, enquanto o classificador pode ser utilizado como detector de mal funcionamento de equipamentos apenas pela análise do status do Sistema.

Palavras-chave: reconhecimento de padrões de utilização; data mining; redução de dimensionalidade; k-means; Análise de Componentes Principais; Mapas Auto Organizáveis; classificadores.

ABSTRACT

Smart meters with automatic meter reading functionalities are becoming popular across the world. As a result, load measurements at various sampling frequencies are now available. Several methods have been proposed to infer device usage characteristics from household load measurements. However, many techniques are based on highly intensive computations that incur heavy computational costs; moreover, they often rely on private household information. In this work, we propose a technique for the detection of appliance utilization patterns using low-computational cost algorithms that do not require any information about households. Appliance utilization patterns are identified only from the system status behavior, represented by large system status datasets, by using dimensionality reduction and clustering algorithms. Principal component analysis, k-means, and the elbow method are used to define the clusters, and the minimum spanning tree is used to visualize the results that show the appearance of utilization patterns. Self-organizing maps are used to create a system status classifier. We applied our techniques to two public datasets from two different countries, the United Kingdom (UK-DALE) and the US (REDD), with different usage patterns. The proposed clustering techniques enable effective demand-side management, while the system status classifier can detect appliance malfunctions only through system status analyses.

Keywords: usage patterns recognition; data mining; dimensionality reduction; k-means; Principal Component Analysis; Self Organizing Maps; classifiers.

LIST OF FIGURES

Figure 1.1: Micro grid example. Adapted from 1-myelectricavenue-i2ev-projectssummaryreport.pdf (eatechnology.com)	23
Figure 1.2: Stages of Consumer engagement in Demand Response - Hui, Ding, Shi, Li, Song and Yan (2020).	24
Figure 1.3: smart city and its challenges representation. Adapted from Iqbal, Malik, Muhammad, Qureshi, Abbassi and Christi (2021).	25
Figure 1.4: building energy consumption in the United States. Numbers after 2006 are projections. Adapted from Vassileva and Campillo (2014).	26
Figure 1.5: citations for both real and synthetic datasets. Adapted from Batra, Parson, Berges, Singh and Rogers (2014).	27
Figure 2.1: Finite state machines models: (a) 500W two states appliance, (b) three states appliance (defrost refrigerator), (c) four state appliance (electrical heater) and (d) three states appliance (clothes dryer). Adapted from Hart (1992).	32
Figure 2.2: How many times Hart (1992) was referenced in other scientific works from 1993 to 2020.	33
Figure 2.3: Energy saving potentials according to the type of user's feedback. Adapted from Batra, Dutta and Singh (2013).	34
Figure 2.4: Both real and synthetic NILM datasets, organized by year and sampling frequency rate. Adapted from Batra, Parson, Berges, Singh and Rogers(2014).	37
Figure 2.5: House and Household characteristics estimation. Adapted from Beckel, Sadamori, Staake and Santini (2014).	40
Figure 2.6: Example of questionnaire questions used by Vassileva and Campillo (2014), and its answers. Adapted from Vassileva and Campillo (2014).	41
Figure 2.7: flowchart explaining the idea of this work.....	43
Figure 3.1: Representation of how the information is organized and the work main steps	45
Figure 3.2: Combination of disaggregation algorithm with utilization patterns detection.....	46
Figure 3.3: correspondence between kW sampling and system status for gas boiler, House 1, UK-DALE.	49

Figure 3.4: correspondence between kW sampling and system status for freezer, House 1, UK-DALE.	49
Figure 3.5: the pre-processing step transforms the real power measurements from individual channels into a binary vector containing each appliance status. The output file also keeps the time stamp and the system aggregated real power instantaneous demand.	50
Figure 3.6: Project flowchart.	50
Figure 3.7: Variance contribution for each Principal Component, HOUSE 2, UK-DALE	53
Figure 3.8: Illustration of matrix X for redundancy reduction	54
Figure 3.9: Variance Explained according to the number of clusters (k), for REDD, House 1. The best k is 4.	55
Figure 3.10: MST as output from Matlab R2016a. To make visualization easier, the figures shown in Chapter 4 were made using software Autocad 2021.	57
Figure 3.11: 2D hexagonal geometry and the winner neuron concept. During each training iteration, the winning neuron moves toward the sample and moves to the first neighborhood. The dimensionality reduction occurs because each neuron, a vector in R_n , is connected t	59
Figure 3.12: Initial hexagonal SOM after training. When looking at the distances between the neurons, the agglomerations became visible.	60
Figure 4.1: Example of output file after preprocessing step. The files contain: time stamp, total house's real power, and the binary status of each individual channel.	63
Figure 4.2: code used in Matlab R2016a to process PCA.	64
Figure 4.3: Variance for each PC and cumulated for UK-DALE, House 4.	65
Figure 4.4: Data Projection over First Principal Component - UK-DALE, House 4.	66
Figure 4.5: Data Projection over First and Second Principal Components - UK-DALE, House 466	
Figure 4.6: Data Projection over First, Second and Third Principal Components - UK-DALE, House 4.	67
Figure 4.7: Matlab R2016a code for generating the MST for UK-DALE, House 1	68
Figure 4.8: MST for REDD, House 1, including the distances between nodes. The appliances with distances smaller than 1 were considered as a single node.	69
Figure 4.9: Code used in Matlab for k-means.	70
Figure 4.10: MST with k-means results represented over the MST for UK-DALE, House 1, with k=6.	71

Figure 4.11: MST with k-means results represented over the MST for UK-DALE, House 2, with k=3.....	72
Figure 4.12: MST with k-means results represented over the MST for UK-DALE, House 3, with k=3.....	73
Figure 4.13: MST with k-means results represented over the MST for UK-DALE, House 4, with k=3.....	73
Figure 4.14: MST with k-means results represented over the MST for UK-DALE, House 5, with k=4.....	74
Figure 4.15: MST with k-means results represented over the MST for REDD, House 1, with k=4.....	75
Figure 4.16: MST with k-means results represented over the MST for REDD, House 2, with k=5.....	75
Figure 4.17: MST with k-means results represented over the MST for REDD, House 3, with k=5.....	76
Figure 4.18: MST with k-means results represented over the MST for REDD, House 4, with k=5.....	76
Figure 4.19: MST with k-means results represented over the MST for REDD, House 5, with k=4.....	77
Figure 4.20: MST with k-means results represented over the MST for REDD, House 6, with k=3.....	77
Figure 5.1: Code used to train the 40x40 hexagonal grid to UK-DALE, House 5.....	79
Figure 5.2: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 1.	80
Figure 5.3: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 2.	81
Figure 5.4: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 3.	82
Figure 5.5: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 4.	82
Figure 5.6: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 5.	83

Figure 5.7: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 1.....	84
Figure 5.8: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 2.....	84
Figure 5.9: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 3.....	85
Figure 5.10: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 4.....	86
Figure 5.11: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 5.....	86
Figure 5.12: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 6.....	87
Figure 9.1: Lattent (individual variance of each PC) and total variance accumulated for UK-DALE, House 1.....	108
Figure 9.2: Data Projection over First Principal Component - UK-DALE, House 1.....	108
Figure 9.3: Data Projection over First and Second Principal Components - UK-DALE, House 1.....	109
Figure 9.4: Data Projection over First, Second and Third Principal Components - UK-DALE, House 12.....	109
Figure 9.5: Variance for each PC and cummulated for UK-DALE, House 2.....	110
Figure 9.6: Data Projection over First Principal Component - UK-DALE, House 2.....	110
Figure 9.7: Data Projection over First and Second Principal Components - UK-DALE, House 2.....	111
Figure 9.8: Data Projection over First, Second and Third Principal Components - UK-DALE, House 2.....	111
Figure 9.9: Variance for each PC and cummulated for UK-DALE, House 3.....	112
Figure 9.10: Data Projection over First Principal Component - UK-DALE, House 3.....	112
Figure 9.11: Data Projection over First and Second Principal Components - UK-DALE, House 3.....	113
Figure 9.12: Data Projection over First, Second and Third Principal Components - UK-DALE, House 3.....	113

Figure 9.13: Variance for each PC and cummulated for UK-DALE, House 4	114
Figure 9.14: Data Projection over First Principal Component - UK-DALE, House 4.....	114
Figure 9.15: Data Projection over First and Second Principal Components - UK-DALE, House 4	115
Figure 9.16: Data Projection over First, Second and Third Principal Components - UK-DALE, House	115
Figure 9.17: Variance for each PC and cummulated for UK-DALE, House 5	116
Figure 9.18: Data Projection over First Principal Component - UK-DALE, House 5.....	116
Figure 9.19: Data Projection over First and Second Principal Components - UK-DALE, House 5	117
Figure 9.20: Data Projection over First, Second and Third Principal Components - UK-DALE, House 5	117
Figure 9.21: Variance for each PC and cummulated for REDD, House 1	118
Figure 9.22: Data Projection over First Principal Component - REDD, House 1.....	118
Figure 9.23: Data Projection over First and Second Principal Components - REDD, House 1 .	119
Figure 9.24: Data Projection over First, Second and Third Principal Components - REDD, House 1	119
Figure 9.25: Variance for each PC and cummulated for REDD, House 2	120
Figure 9.26: Data Projection over First Principal Component - REDD, House 2.....	120
Figure 9.27: Data Projection over First and Second Principal Components - REDD, House 2 .	121
Figure 9.28: Data Projection over First, Second and Third Principal Components - REDD, House 2	121
Figure 9.29: Variance for each PC and cummulated for REDD, House 3	122
Figure 9.30: Data Projection over First Principal Component - REDD, House 3.....	122
Figure 9.31: Data Projection over First and Second Principal Components - REDD, House 3 .	123
Figure 9.32: Data Projection over First, Second and Third Principal Components - REDD, House 3	123
Figure 9.33: Variance for each PC and acumulated for REDD, House 4	124
Figure 9.34: Data Projection over First Principal Component - REDD, House 4.....	124
Figure 9.35: Data Projection over First and Second Principal Components - REDD, House 4 .	125

Figure 9.36: Data Projection over First, Second and Third Principal Components - REDD, House 4	125
Figure 9.37: Variance for each PC and cummulated for REDD, House 5	126
Figure 9.38: Data Projection over First Principal Component - REDD, House 5.....	126
Figure 9.39: Data Projection over First and Second Principal Components - REDD, House 5 .	127
Figure 9.40: Data Projection over First, Second and Third Principal Components - REDD, House 5	127
Figure 9.41: Variance for each PC and cummulated for REDD, House 6	128
Figure 9.42: Data Projection over First Principal Component - REDD, House 6.....	128
Figure 9.43: Data Projection over First and Second Principal Components - REDD, House 6 .	129
Figure 9.44: Data Projection over First, Second and Third Principal Components - REDD, House 6	129
Figure 10.1: Variance Explained according to the number of clusters (k), for UK-DALE, House 1. The best k is 6.....	130
Figure 10.2: Variance Explained according to the number of clusters (k), for UK-DALE, House 2. The best k is 3.....	132
Figure 10.3: Variance Explained according to the number of clusters (k), for UK-DALE, House 3. The best k is 3.....	133
Figure 10.4: Variance Explained according to the number of clusters (k), for UK-DALE, House 4. The best k is 5.....	134
Figure 10.5: Variance Explained according to the number of clusters (k), for UK-DALE, House 5. The best k is 4.....	135
Figure 10.6: Variance Explained according to the number of clusters (k), for REDD, House 1. The best k is 4.	136
Figure 10.7: Variance Explained according to the number of clusters (k), for REDD, House 2. The best k is 5.	137
Figure 10.8: Variance Explained according to the number of clusters (k), for REDD, House 3. The best k is 5.	138
Figure 10.9: Variance Explained according to the number of clusters (k), for REDD, House 4. The best k is 5.	139

Figure 10.10: Variance Explained according to the number of clusters (k), for REDD, House 5. The best k is 4.....	140
Figure 10.11: Variance Explained according to the number of clusters (k), for REDD, House 6. The best k is 3.....	142
Figure 11.1: MST for UK-DALE, House 1, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.....	144
Figure 11.2: MST for UK-DALE, House 2, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.....	144
Figure 11.3: MST for UK-DALE, House 3, including the distances between nodes.	144
Figure 11.4: MST for UK-DALE, House 4, including the distances between nodes.	145
Figure 11.5: MST for UK-DALE, House 5, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.....	145
Figure 11.6: MST for REDD, House 1, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.....	146
Figure 11.7: MST for REDD, House 2, including the distances between nodes.	146
Figure 11.8: MST for REDD, House 3, including the distances between nodes.	147
Figure 11.9: MST for REDD, House 4, including the distances between nodes.	147
Figure 11.10: MST for REDD, House 5, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.....	148
Figure 11.11: MST for REDD, House 6, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.....	148
Figure 12.1: MST with k-means results represented over the MST for UK-DALE, House 1, with k=6.....	149
Figure 12.2: MST with k-means results represented over the MST for UK-DALE, House 2, with k=3.....	150
Figure 12.3: MST with k-means results represented over the MST for UK-DALE, House 3, with k=3.....	150
Figure 12.4: MST with k-means results represented over the MST for UK-DALE, House 5, with k=3.....	151
Figure 12.5: MST with k-means results represented over the MST for UK-DALE, House 5, with k=4.....	151

Figure 12.6: MST with k-means results represented over the MST for REDD, House 1, with k=4	152
Figure 12.7: MST with k-means results represented over the MST for REDD, House 2, with k=5	152
Figure 12.8: MST with k-means results represented over the MST for REDD, House 3, with k=5	153
Figure 12.9: MST with k-means results represented over the MST for REDD, House 4, with k=5	153
Figure 12.10: MST with k-means results represented over the MST for REDD, House 5, with k=4	154
Figure 12.11: MST with k-means results represented over the MST for REDD, House 6, with k=3	154

LIST OF TABLES

Table 3.1: Number of individual channels and time monitored for every house in UK-DALE dataset	47
Table 3.2: Clusters for REDD House 1, $k=4$	56
Table 4.1: Number of samples (continuous days) considered for each House in REDD dataset.	62
Table 8.1: Number of samples in each individual channel, and the correspondent label for House 1, UK-DALE dataset. In this house, channels 22, 39, 40 and 41 were eliminated for having too few samples compared to the other channels.	102
Table 8.2: Number of samples in each individual channel, and the correspondent label for House 2, UK-DALE dataset. In this house, it is not necessary to eliminate any channel.	103
Table 8.3: Number of samples in each individual channel, and the correspondent label for House 3, UK-DALE dataset. In this house, it is not necessary to eliminate any channel.	103
Table 8.4: Number of samples in each individual channel, and the correspondent label for House 4, UK-DALE dataset. In this house, it is not necessary to eliminate any channel.	103
Table 8.5: Number of samples in each individual channel, and the correspondent label for House 5, UK-DALE dataset. In this house, channels 11 and, 25 are not included for having too few samples compared to the other channels.	104
Table 8.6: Number of samples in each individual channel, and the correspondent label for House 1, REDD dataset.	104
Table 8.7: Number of samples in each individual channel, and the correspondent label for House 2, REDD dataset.	105
Table 8.8: Number of samples in each individual channel, and the correspondent label for House 3, REDD dataset.	105
Table 8.9: Number of samples in each individual channel, and the correspondent label for House 4, REDD dataset.	106
Table 8.10: Number of samples in each individual channel, and the correspondent label for House 5, REDD dataset.	106
Table 8.11: Number of samples in each individual channel, and the correspondent label for House 6, REDD dataset.	106

Table 10.1: Clusters for UK-DALE House 1, $k=6$. The groups with “+” are considered as a single point.....	131
Table 10.2: Clusters for UK-DALE House 2, $k=3$. The groups with “+” are considered as a single point.....	133
Table 10.3: Clusters for UK-DALE House 3, $k=3$	134
Table 10.4: Clusters for UK-DALE House 4, $k=3$	134
Table 10.5: Clusters for UK-DALE House 5, $k=4$. The groups with “+” are considered as a single point.....	135
Table 10.6: Clusters for REDD House 1, $k=4$	136
Table 10.7: Clusters for REDD House 2, $k=5$	137
Table 10.8: Clusters for REDD House 3, $k=5$	138
Table 10.9: Clusters for REDD House 4, $k=5$. The groups with “+” are considered as a single point.	139
Table 10.10: Clusters for REDD House 5, $k=4$. . The groups with “+” are considered as a single point.....	141
Table 10.11: Clusters for REDD House 6, $k=3$	142

LIST OF ABBREVIATIONS

2D	2 Dimension
3D	3 Dimension
ALM	Appliances Load Monitoring
ANN	Artificial Neural Networks
CER	Commission for Energy Regulation
COMBED	Commercial Building Energy Dataset
COVID-19	Corona Virus Disease-2019
DeA	Detection Accuracy
DER	Detection Error Rate
DiA	Disaggregation Accuracy
DSR	Demand Side Response
EEUD	Electrical-End-User Dataset
EMF	Electromagnetic Field Detectors
FD	Failed Detections
FHMM	Factorial Hidden Markov Models
HEMS	House Energy Management System
HMM	Hidden Markov Models
ICT	Information and Communication Technologies
ILM	Intrusive Load Monitoring
IoT	Internet of Things (IoT)
MG	Microgrid
MST	Minimum Spanning Tree
NILM	Non-Intrusive Load Monitoring
OA	Overall Accuracy
PC	Principal Component
PCA	Principal Component Analysis
REDD	Reference Energy Disaggregation Data Set
SOM	Self Organizing Maps

UK-DALE United Kingdom Domestic Appliance-Level Electricity
US United States
W Watt

SUMMARY

1. INTRODUCTION	22
2. THEORETICAL BACKGROUND	31
2.1 INTRUSIVE AND NONINTRUSIVE LOAD MONITORING (ILM AND NILM)	31
2.2 PUBLIC DATASETS WITH LOAD MONITORING INFORMATION	36
2.3 ENERGY MANAGEMENT SYSTEM AND DEMAND RESPONSE	38
2.4 TARGETED STUDIES	41
3. METHODOLOGY	44
3.1 DATA DESCRIPTION	46
3.2 PRE-PROCESSING	48
3.3 PRINCIPAL COMPONENT ANALYSIS (PCA)	51
3.4 K-MEANS	55
3.5 3D VISUALIZATION AND THE MINIMUM SPANNING TREE	56
3.6 SELF ORGANIZING MAPS	57
4. RESULTS – PRE PROCESSING, PCA AND K-MEANS	61
1.1 PRINCIPAL COMPONENT ANALYSIS	63
1.2 MINIMUM SPANNING TREE AND K-MEANS	67
1.3 K-MEANS RESULTS FOR PCA 3D	69
5. SELF ORGANIZING MAPS RESULTS	78
5.1 SOM RESULTS FOR UK-DALE	79
5.2 REDD RESULTS	84
6. DISCUSSION	87
6.1 PCA AND K-MEANS FOR UK-DALE AND REDD DATASETS	87
6.2 SELF ORGANIZING MAPS FOR THE UK-DALE AND REDD DATASETS	89

7.	CONCLUSION AND FUTURE WORK	91
7.1	CONCLUSION	91
7.2	FUTURE WORK	92
8.	APENDIX A – DATASETS	102
9.	APENDIX B – PRINCIPAL COMPONENT ANALISYS RESULTS.....	108
9.1	UK-DALE, HOUSE 1	108
9.2	UK-DALE, HOUSE 2	110
9.3	UK-DALE, HOUSE 3	112
9.4	UK-DALE, HOUSE 4	114
9.5	UK-DALE, HOUSE 5	116
9.6	REDD, HOUSE 1	118
9.7	REDD, HOUSE 2	120
9.8	REDD, HOUSE 3	122
9.9	REDD, HOUSE 4	124
9.10	REDD, HOUSE 5	126
9.11	REDD, HOUSE 6	128
10.	APENDIX C – K-MEANS RESULTS.....	130
11.	APENDIX D – MINIMUM SPANNING TREE RESULTS	143
12.	APENDIX E – K-MEANS VISUALIZATION OVER THE MST	149

1. INTRODUCTION

Today, the infrastructure of generation, transmission, distribution and consumption of electricity is facing a remarkable change. The transitions to low carbon, decentralized and heavily electrified energy, which requires (i) smarter grids that can deal with the intermittency of renewable energy sources (especially solar and wind), (ii) a reliable and robust transmission system that can deliver the energy at the main load centers, and (iii) mechanism capable of supplying the demand peaks (see My Electric Avenue Project - 1-myelectricavenue-i2ev-projectssummaryreport.pdf (eatechnology.com)). Besides, the distributed energy resources are now available in low voltage distribution networks (because it is now feasible to connect to the grid sources as small as a rooftop solar panel even for small-scale consumers). At the same time, advances in demand-side management, automation technologies and the Internet of Things (IoT) opens the possibility for the consumers to make decisions about their own energy use.

In this sense, the emergence of the neologism “prosumer”, related to an agent that can both produce and consume energy (see [EcoGrid EU - http://www.ecogrid.dk/src/EcoGridEU%20%20A%20prototype%20for%20euro-pean%20smart%20grids%20160121.pdf?dl=0](http://www.ecogrid.dk/src/EcoGridEU%20%20A%20prototype%20for%20euro-pean%20smart%20grids%20160121.pdf?dl=0)), is another evidence of structural changes in energy markets. The low voltage prosumers concept comes with the microgrids reality. The control devices (smart grid) can deal with the system constraints and demand to combine comfort (mainly related to air conditioning and heating, but not limited to them) and the benefits of microgrids/nanogrids (which may lead to a potential reduction in energy costs).

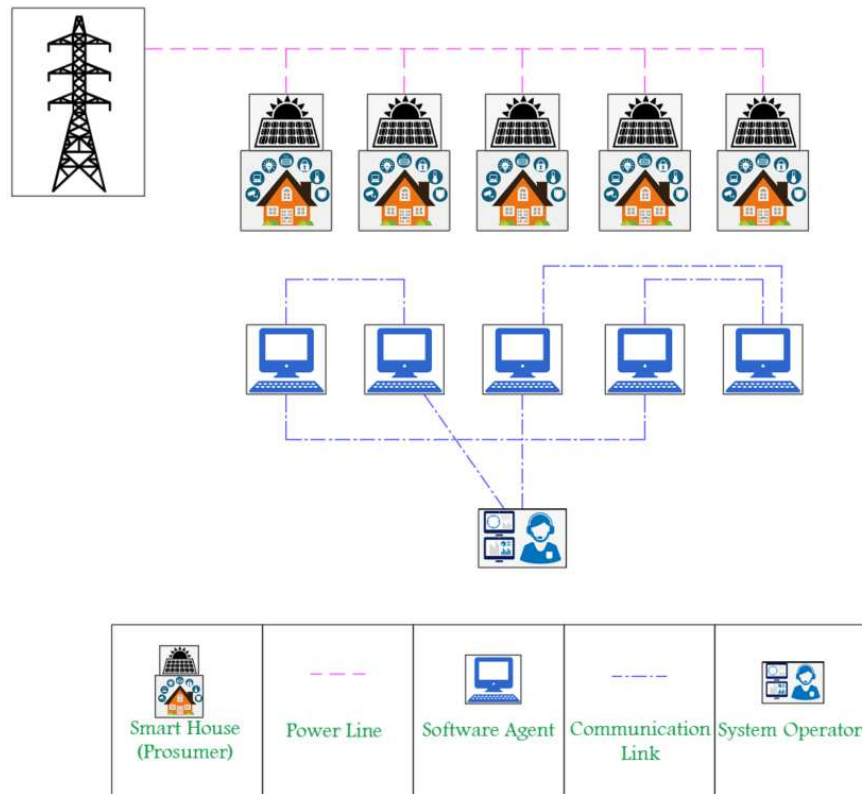


Figure 1.1: Micro grid example. Adapted from 1-myelectricavenue-i2ev-projectsummaryreport.pdf (eatechnology.com)

The use of microgrids (MG) is remarkably expanding because it allows for exploiting a wider range of renewable energy sources. The MG system can share heat and electrical energy with a higher reliability, cost, and green-house emission reduction. As a result, the system flexibility is increased, together with peak shaving possibilities, as described in Hosseini, Agbossou, Kelouwani and Cardenas (2017). But the uncertainties of this kind of energy source is operationally challenging, and thus, proper management is now of greater importance. The control system must deal with the constraints and uncertainties of this kind of energy source, while managing loads to guarantee the voltage stability and load supply, either if the microgrid is connected to the main grid or if it operates as an islanded system. In this scenario, the existence of storage devices and a demand response program is very important to achieve a technically feasible solution and the expected cost reduction.

Another important point is related to the demand response motivation. Industrial installation has a little flexibility. A specific industrial process defined by itself which load should be

turned on at each time, leaving limited room for load modulation according to the energy tariff and the peak periods. The possibility for reducing energy consumption comes from equipment retrofit with better energy efficiency. Residential users represent a great flexibility potential that can result in a better use of wind and solar energy, contributing to energy decarbonization. The load shift potential also can reduce the load peak and relieve the pressure for increasing generation, transmission and distribution capacity as load demand continues to increase, as described in Rajamand (2020).

When the energy policy focuses on residential installations, an important factor to consider is related to the user's motivation and engagement to cooperate with its potential flexibility, as aborded in Parrish, Heptonstall, Gross ans Sovacool (2020) and EPRI, 2011. A better understanding of the people's engagement with demand response can be of great help to propose more effective measures. The advertisement programs can be more targeted to customers that have greater flexibility potential, and the results can be achieved faster if the consumers are open to participate in demand response programs, as shown in Rajamand (2020). In Hui, Ding, Shi, Li, Song and Yan (2020), the consumer engagement is studied, and the three main points represented in Figure 1.2 shall be observed when developing a demand response program more focused on specific groups of users.

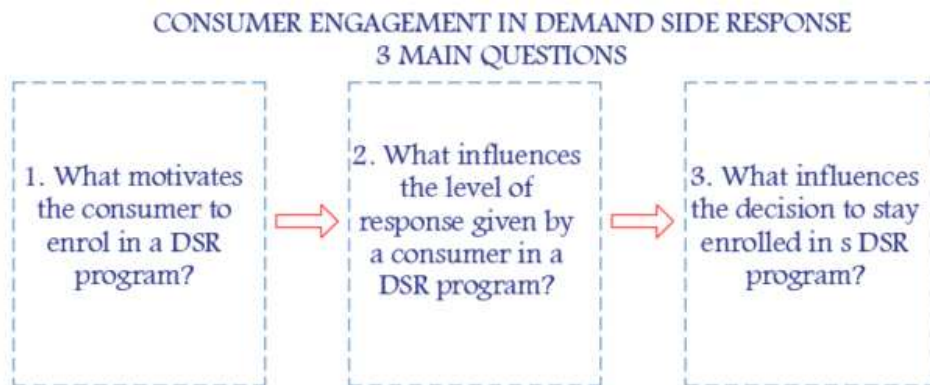


Figure 1.2: Stages of Consumer engagement in Demand Response - Hui, Ding, Shi, Li, Song and Yan (2020).

The demand response can be performed by a residential energy management system, as shown in US Department of Energy (2015), but the definition of constraints is a task closely related not only to the microgrid, but also to the household needs and priorities. This requires input information to an optimization algorithm such as load classification, daily tariff information, and

several use and shift constraints. An efficient House Energy Management System (HEMS) is an interesting demand response service to be offered to the users.

Still in the context of residential loads, the concept of smart city is a good example of the reality in the energy field today. The number of electric devices in everyday use is constantly increasing from cloud storage that requires a fast and constant internet connection to electrical vehicles that need to be charged without impairments. In Iqbal, Malik, Muhammad, Qureshi, Abbassi and Christi (2021), the authors say that the advance of internet technology together with the Internet of Things (IoT) enabled by 5G technology opens a wide range of possibilities and challenges that must be addressed. The intermittency of renewable energy sources must be incorporated to the design of future smart cities and smart grids so that the system flexibility could be increased as much as possible by employing such cutting-edge information and communication technologies (ICT).

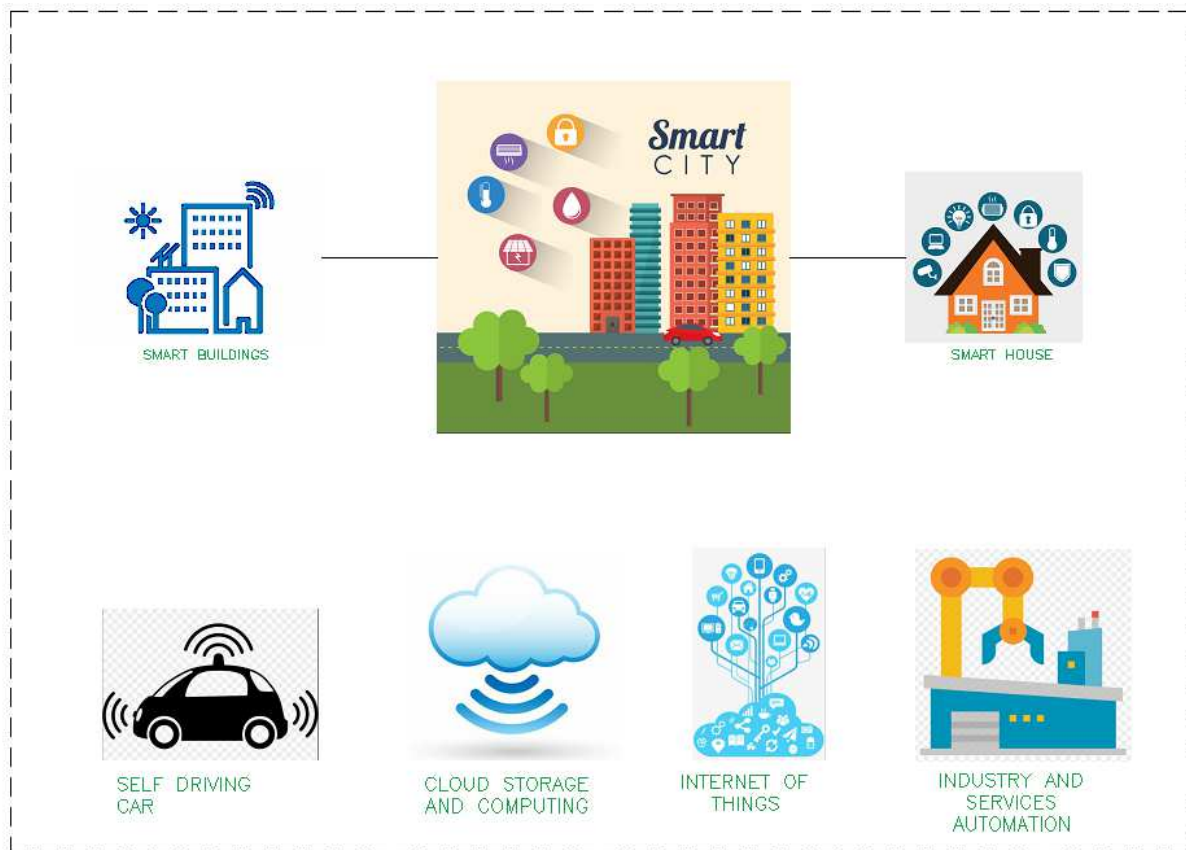


Figure 1.3: smart city and its challenges representation. Adapted from Iqbal, Malik, Muhammad, Qureshi, Abbassi and Christi (2021).

Load monitoring and demand side load management play an important role in this scenario. According to Greening, Greene and Difiglio (2000), in 2015 buildings in the United States were responsible for 38% of primary energy consumption and 76% of electricity use, and this number can be reduced to less than 50% with a building energy management system. Figure 1.4 shows the historical and projection of buildings energy use in the US.

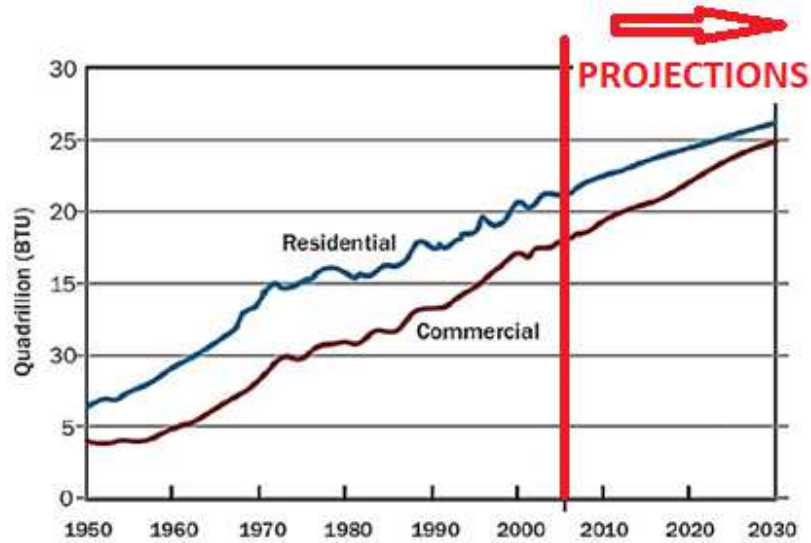


Figure 1.4: building energy consumption in the United States. Numbers after 2006 are projections. Adapted from Vassileva and Campillo (2014).

However, to develop an energy management system with good performance it is first necessary to understand the loads' behavior and the energy profile of households. The concept of Non-Intrusive Load Monitoring (NILM) was introduced in Hart (1992), and since then it has been extensively studied resulting in more effective algorithms. Several load characteristics and computational resources have been used to identify load usage without the expensive and invasive plug monitoring system. In Batra, Parson, Berges, Singh and Rogers (2014), the authors provide a historical review of NILM evolution. Despite the different goals and computational methods used to identify residential loads, which makes it difficult to measure (and thus monitor) their efficiency, a better understanding of the buildings' (either commercial or residential) load use allows the diagnosis and control of different loads connected to the grid. This can help the customers to be aware of their individual appliances' energy consumption and provide an important information source to define public policies and tariffs, among other advantages.

The high interest for demand side studies was supported by the availability of several public datasets with monitoring information of real households. The information available is used to develop and measure the efficiency of several NILM algorithms. In Batra, Parson, Berges, Singh and Rogers (2014), the authors present a comprehensive review of 42 NILM datasets through comparison tables. Figure 1.5 shows the number of citations for all datasets compared (some of them are real, others are synthetic, that means that the data is simulated).

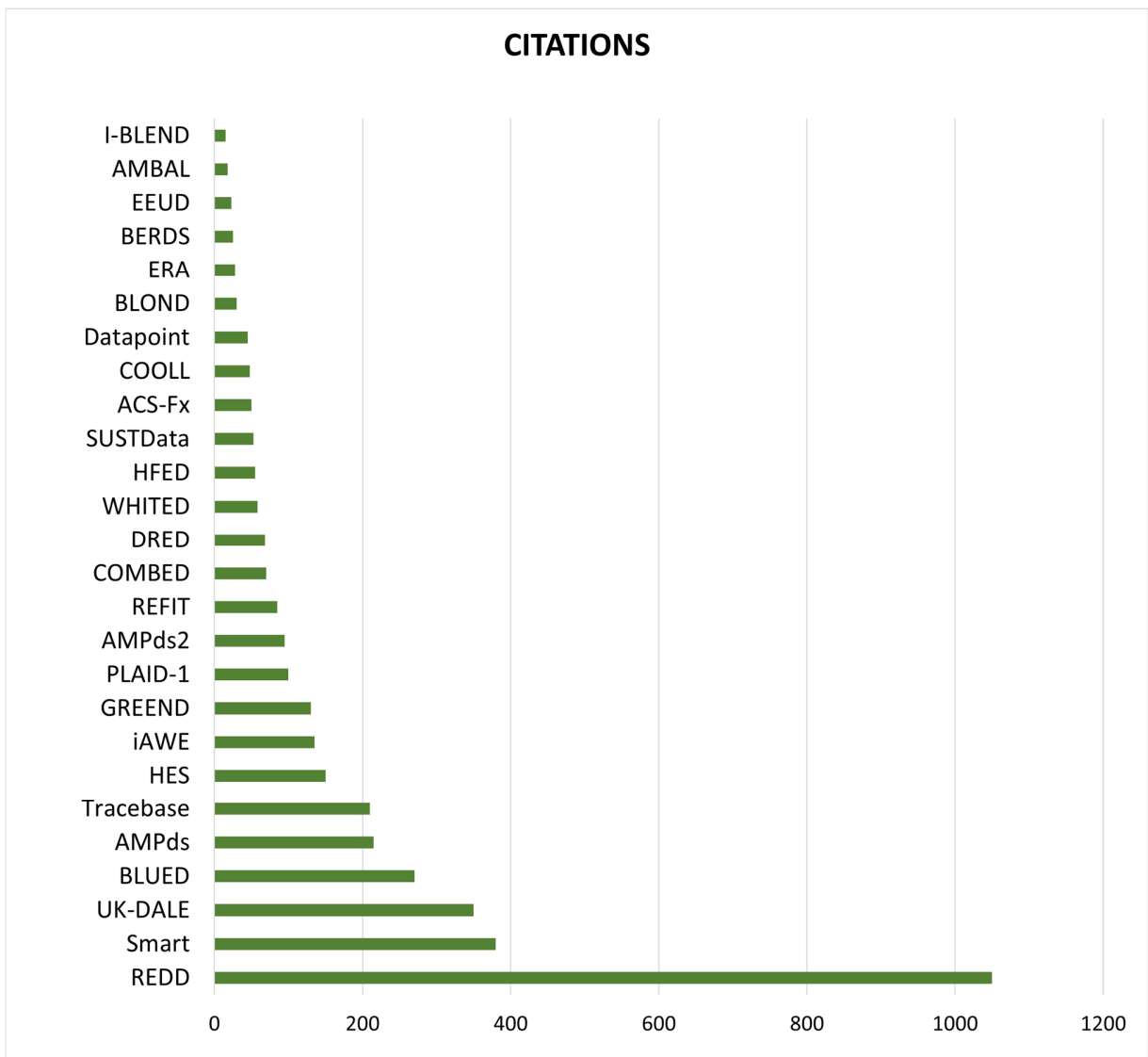


Figure 1.5: citations for both real and synthetic datasets. Adapted from Batra, Parson, Berges, Singh and Rogers (2014).

Regarding residential installations (either buildings or houses), the use of the appliances is directly affected by the daily and weekly routine of people who live in the house. This

way, school schedule, eating habits, laundry necessities, and other personal preferences contribute to the appliances' use pattern. With movement restrictions imposed in most countries due to the COVID-19 pandemic, many of these patterns have changed drastically. Many households that did not have a printer or computer screen before now have them among the most used appliances. Most meals are now prepared at home, changing the kitchen utilities profile. The appliances' use can also play an important role in psychological and physical health, as many physical exercise programs are guided through the internet, for example, and the friend's meetings also demand a set of appliances.

Regarding residential demand control or even load shifting, instead of the remarkable concern about the impact of energy consumption, most do not want to change the laundry routine or cooking habits to shift the daily demand peak, for example. Personal routines in general are not easy to change, especially when it refers to what people do inside their own houses. Of course, new appliances are of great help in the energy consumption reduction, but it is not enough. The increased quantity of electronic devices and the importance that the constant availability pushes the consumption faster than the new technologies can reduce it (this phenomenon is called in the literature as "rebound effect"). On the other hand, old concepts can be deceiving. With the current appliances' profile, finding the appliance that has the higher load modulating potential, or that better contributes to the energy efficiency improvement, is not an easy task. An individualized codesign of effective strategies is important to adapt personal habits.

From this perspective, use patterns studies are very important. Also, it must be fast and dynamic. With the advance of disaggregation algorithms, as shown in Batra, Parson, Berges, Singh and Rogers (2014), and Internet of Things (IoT), the system status can be obtained in a non-intrusive way, only using as input the information of aggregate power demand taken by the smart meter. System status behavior feedback could be offered either from the energy service providers itself, or from a consulting company, all with the customers' agreement.

With a detailed pattern utilization detection, it is possible to have a better understanding of the consumers' habits, and thus, the service provider can compose a specific feedback letter (attached to the energy bill), or a mobile phone application, with suggestions to improve the residence's energy efficiency if agreed beforehand. This might be an additional motivation factor so more people would agree to have the residence monitored. With better information, the methods can be more effective, and the programs can reach more houses, and so on.

The present work follows this line by focusing on the detection of different appliances' utilization patterns using two steps strategies: first, using only linear and non-parametric algorithms (PCA, and k-means clustering and Minimum Spanning Tree methods), the existence of appliance clusters will be investigated, and a methodology for defining the groups of appliances inside each group is established. Second, using nonlinear methods, specifically Self Organizing Maps (SOM), the possibility of using the trained map as a status classifier is explored.

In specific terms, we start from a relatively large set of system states (the larger, the better) and use an established dimensionality reduction method, followed by clustering algorithms to find utilization patterns. These patterns are represented by groups of appliances statistically related (meaning that they used at the same time). The appliances inside each group do not necessarily have to be used in combination (e.g., video game and TV); they are rather statistically related indicating that they are frequently used at the same time.

It is important to say that the formulations chosen are well established and require a small quantity of parameters (in the first step, no parametrization at all), and none of them is related to the household's habits, profession, or other personal information. This way. Our main contribution here is the proposed methodology for utilization patterns detection on residential installations with a low computational cost and that does not require personal habits information of the occupants.

The main contributions of this thesis are:

- A proposed methodology for utilization patterns detection using well-established formulations that require a small quantity of parameters at low computational cost.
- The method is noninvasive because it does not require any personal information about the households.
- The Self Organizing Maps opens a wide range of analysis opportunities for the system status, starting from malfunction and fault detection.

The contributions described above were already published in the following works:

- (1) Villar, F.; da Silva, L.C.P.; Nardelli, P.H.J.; Hazini, H. Detection of Appliance Utilization Patterns via Dimensionality Reduction. In Proceedings of the 2019 IEEE PES Innovative Smart Grid Technologies Conference-Latin America (ISGT Latin America), Gramado, Brazil, 15–18 September 2019; pp. 1–6.

- (2) VILLAR, Fernanda Spada et al. Noninvasive Detection of Appliance Utilization Patterns in Residential Electricity Demand. *Energies*, v. 14, n. 6, p. 1563, 2021.

The thesis is organized as follows:

- Chapter 1 (Introduction) explains the scenario and the central question.
- Chapter 2 (Related Work) lists methods and results already achieved by other authors.
- Chapter 3 (Methodology) details the methods used.
- Chapter 4 (PCA Results) shows the results for the Principal Component Analysis (PCA) method.
- Chapter 5 (SOM results) shows the results for the Self-Organizing Maps (SOM) method.
- Chapter 6 (Discussion) discusses the results obtained in Chapters 4, 5 and 6.
- Chapter 7 (Conclusion) gives the conclusion to the central question proposed in Chapter 1.
- Chapter 8 (References) lists the references.

2. THEORETICAL BACKGROUND

Since the 2000s, different researchers have focused on methods to better understand residential demand of the electricity grid to manage it accordingly. People's behavior and the factors that motivate them to act in one or other direction directly influences the electricity used. The interest in user behavior is not exclusive to the energy field. Goldstein, Cialdini and Griskevicius (2008) proposed two experiments to study guest's motivation in cooperating with an environmental program of water use reduction by reuse of towels. The guests were asked to participate in an environmental program by reusing the bath towel for 2 days or more during their stay. Two experiments were made, each of them using 2 types of signs to present the program, with different motivational sentences. The first 2 sentences were: "Help save the environment" and "join you fellow guests in helping to save the environment", the second one also informing that 75% of the hotel guests reused the towels during the stay. The result was: 33% adhered to the program with the first sign, and 44% with the second sign. The second experiment used the sentences: "Help save the environment", "Join your fellow guests in helping to save the environment", "Join your fellow citizens in helping to save the environment" and "join the men and women who are helping to save the environment". The idea was to evaluate how the identification with similar groups influences the motivation to participate in the program. This work shows: a) how users' behavior attracts the interest of researchers, and b) that good feedback is very important if you want to influence someone's behavior.

2.1 INTRUSIVE AND NONINTRUSIVE LOAD MONITORING (ILM AND NILM)

When researchers start to be interested in user's behavior, the concept of privacy also becomes very important. On one hand, a good prediction of resources' necessity (among them energy) requires a deep understanding of users' behavior, and to make a more intelligent use of energy, some changes in the behavior must occur. But on the other side, the sensation of being observed can make people uncomfortable, especially when they are observed in their private moments. In this context, Hart (1992) introduces the concept of Nonintrusive Load Monitoring

(NILM), with the proposal of replacing the several individual plug monitors installed in the houses monitored by a computational routine that can identify the individual appliances using the total aggregated power measurements. The work uses the representation of the residential appliances as finite-state machines with 2, 3 or 4 states (see Figure 2.1).

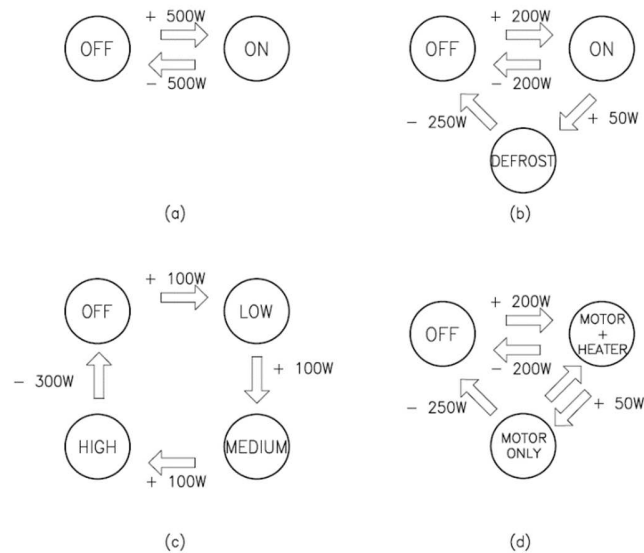


Figure 2.1: Finite state machines models: (a) 500W two states appliance, (b) three states appliance (defrost refrigerator), (c) four state appliance (electrical heater) and (d) three states appliance (clothes dryer). Adapted from Hart (1992).

The concept of nonintrusive signatures is also used to help in the appliance identification process, like the steady state real and reactive power and power factor. By the time this work was published, the computational effort to perform the NILM was a limiting factor, as was lower data transmission rates and hardware limitations to make the intrusive load monitoring (ILM) first to then use it as ground truth to test NILM performance. But with the advance of computational capacities, either data storing, transmitting, and processing, put NILM as a relevant topic in energy research. Figure 2.2 shows how many times Hart (1992) was used as a reference for scientific works from 1993 to 2020.

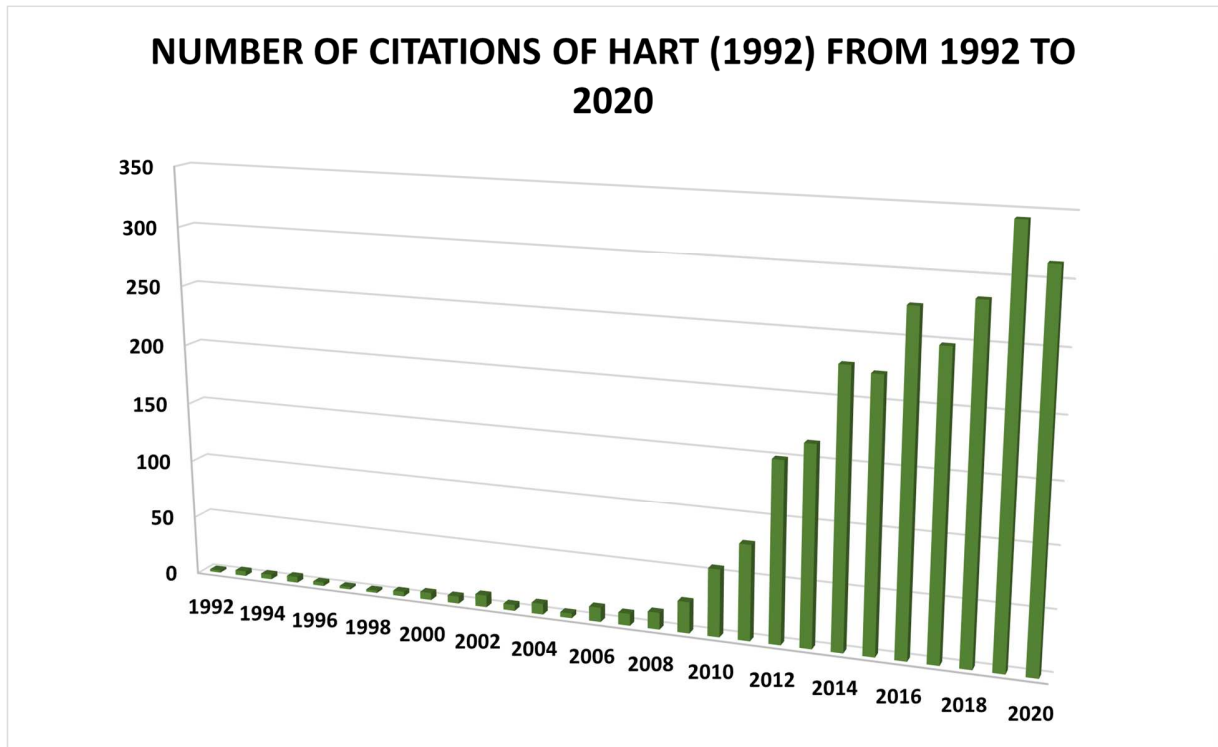


Figure 2.2: How many times Hart (1992) was referenced in other scientific works from 1993 to 2020.

In Armel, Gupta, Shrimali and Albert (2013), the energy saving potential from residences is the focus, and which actions can be made to help consumers to achieve these savings. According to it, information about the use of specific appliances can bring benefits not only to the final user, who would have a smaller energy bill, but also to the utility and policy sector, as well as research and development. Figure 2.3, adapted from Batra, Dutta and Singh (2013), shows the energy saving potential related to different types of user feedback. The conclusion is that the more detailed, fast, and personalized the feedback, the higher the energy saving potential. But to make such a feedback feasible, a large research on specific appliance information and user's behavior must be performed.

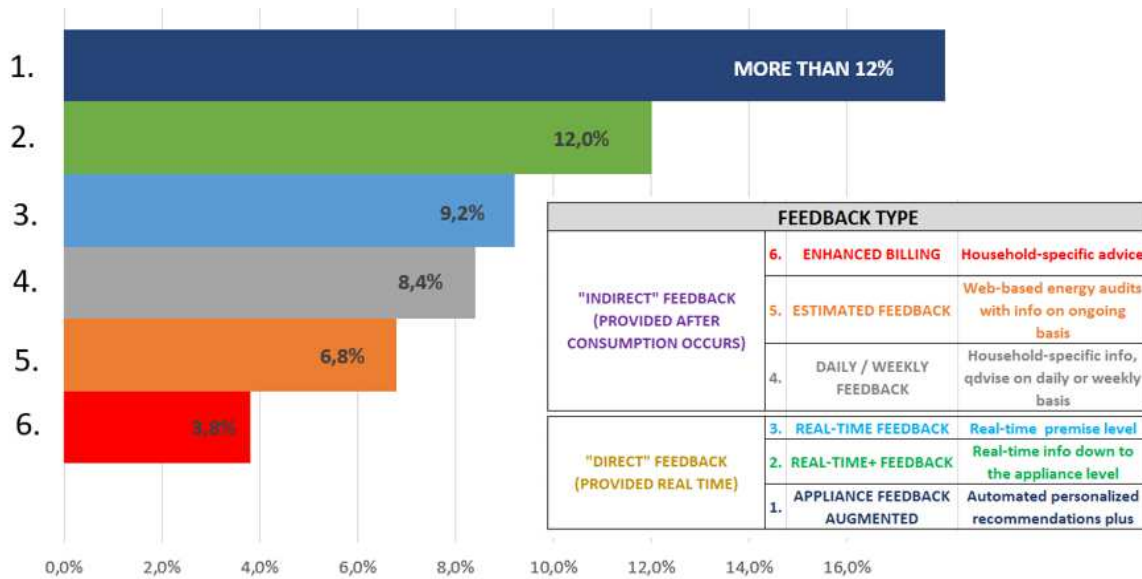


Figure 2.3: Energy saving potentials according to the type of user's feedback. Adapted from Batra, Dutta and Singh (2013).

Studies in these two fields were published since then, and the measurements of several load monitoring programs were published as public datasets, with information that could be used to improve the NILM algorithms.

In Klemenjak and Goldsborough (2016), an overview of the importance of NILM is given, starting from a vocabulary review (types of load modeling, signatures, smart meters). The term “appliance signature” is reviewed, and defined as “measurable parameters, which provide device-specific information extracted from physical quantities”. The signatures are divided into two main groups:

1. *Steady-state signatures*: features extracted from appliances during its steady state use (means that the features cannot be seen during the transient moment between one state and another - “on” to “off” in a “on-off” appliance model). Typical steady state signatures are a) real and reactive power, b) V-I features; c) V-I trajectory; d) harmonics.

Real and reactive power is the feature that requires less computational effort to be detected (requires only measurements of the fundamental frequency) but is the most deceiving signature. The overlapping of more than one appliance can result in a wrong detection. Following this line, harmonics signature requires a high sampling frequency (the higher, the larger the harmonics

profile can be observed), and thus dealing with large data sets storage and transmission but results in a more assertive detection.

2. *Transient-state signatures*: the transient signature of an appliance is strongly related to the components that it is composed of. For example, the turn-on current peak of an electrical motor is very different from a heater because of the capacitive and inductive components of the first one. Not only the shape, but also the duration of the transient in voltage and current can make an appliance detection much more assertive.
3. *Ambient appliance features*: more than power, voltage or current, some other characteristics can help to identify an appliance. Some external information can be of great help for NILM, as for example a light sensor to help differentiate a light from another appliance that has the same real power. Even Electromagnetic field detectors (EMF) were already used to contribute to the NILM algorithm. This additional information can improve the algorithm performance but requires more computational capacity for dealing with more sensors and types of variables. Also, the installation of some of these sensors can make the load monitoring more intrusive.

An overview of learning approaches is also given. The algorithms are divided in 2 main types: supervised and unsupervised learning. Supervised methods start from a dataset with a set of possible load signatures, and then an optimization or pattern recognition step performs the disaggregation to define the system status (the status of every appliance monitored). With them, it is possible to use a large set of signatures but requires a large storage and transmission capacity and a good processing unit to identify the system status changes. Another disadvantage is that if there is a new appliance in the system, the process must be reprogrammed to include a new signature.

Unsupervised methods can operate without previous information about the system and the loads. They are a promising alternative, because supervised routines require a long period of supervised learning, which can make its implementation in many houses unfeasible. Some examples of unsupervised learning are Hidden Markov Models – HMM -, Factorial Hidden Markov Models – FHMM, both using probabilistic analysis, and some artificial neural networks (ANN) methods, such as single or multilayer perceptron and Self Organizing Maps (SOM).

2.2 PUBLIC DATASETS WITH LOAD MONITORING INFORMATION

To make a high performance NILM method, it is necessary a dataset with intrusive monitoring information to be used as reference (called “ground truth”) for the tests. With the development of cloud storage, several projects made their datasets public to be used as reference for disaggregation research. In Iqbal et. Al., (2020) a comparison between 42 of these datasets can be found. Some of the datasets analyzed are described below.

- The first and widely used public dataset is REDD, published in Kolter and Johnson (2011) with aggregated and disaggregated information of six real houses in the United States. Data collected includes high frequency (15kHz) sampling of mains current and voltage, and low frequency (0.5Hz) data of the individual appliances.
- Smart* dataset, published in Barker, Mishra, Irwin, Cecchet, Shenoy and Albrecht (2012), contains electricity data from 3 houses for a period of 3 to 4 months, and Smart* (pronounced as smart star) recorded non continuous information from 7 houses for a period of 3 years.
- United Kingdom domestic appliance-level electricity dataset (UK-DALE), published in Kelly and Knottenbelt (2014) monitors 5 real houses in the United Kingdom at 10 Hz sampling frequency for periods of more than 1 month.
- Commercial building energy dataset (COMBED), published in Batra, Parson, Berges, Singh and Rogers(2014), is the first dataset related to commercial buildings. The samplings are made in a single building in India with a sampling period of more than 1 minute.
- Electrical-end-user dataset (EEUD), published by Anand Krishnan, Tyler Byers, Vincent Smart in 2021 by New Zeland Energy Efficiency and Conservation Authority, provides annual measurements for 23 Canadian houses at one-minute resolution.

The different characteristics of each dataset indicate which is the most suitable option for a specific type of algorithm. For instance, the larger the sampling frequency, the wider the range of transient state signatures that can be used in the NILM algorithm. But in every case, the scholars must deal with common problems, such as sampling “blackouts” (large periods with no measure

due to a smart meter or smart plug malfunction) and noise. Figure 2.4 shows a scatter plot of both real or synthetic NILM public datasets with releasing year and type of information (high frequency sampling, low frequency sampling or both). The availability of these large number of public datasets, with a large range of characteristics, such as type of building monitored (residential, commercial, university), number of houses and individual appliances monitored, sampling frequency and period of samplings recorded, is fundamental for the intense improvement in the disaggregation algorithms in the past years. Also, they make the methodologies replicable for other datasets.

YEAR	DATASETS						
2011	REDD						
2012	HES	Tracebase	Smart	BLUED			
2013	IHEPCDS	ACS-Fx	AMPds	iAWE	Data Port	BERDS	
2014	RBSA	COMBED	GREEND	UK-DALE	ECO	PLAID-1	SustData
2015	HEFED	DRED	REFIT				
2016	COOLL	WHITED	OPLD	SustDataED	AMPds2	SmartSim	
2017	EEUD	PLAID-II	AMBAL	ESHL			
2018	ERA	BLOND	HELD1	SHED			
2019	ENERTALK	I-BLEND	HUE				
2020	SynD	IDEAL	CU-BEMS	PLAID-III			

LOW FREQUENCY SAMPLING
HIGH FREQUENCY SAMPLING
BOTH (HIGH AND LOW) FREQUENCY SAMPLING

Figure 2.4: Both real and synthetic NILM datasets, organized by year and sampling frequency rate. Adapted from Batra, Parson, Berges, Singh and Rogers(2014).

In Pereira and Nunes (2018) a performance evaluation of NILM regarding datasets, metrics and tools is described. The comparisons are made over a set of 26 public datasets, and characteristics such as year of release, country, number of monitored households, data continuity (or not), type of smart meter used, time resolution, and others are listed in tables. The work also summarizes the performance metrics used during the more than 20 years of NILM development. The metrics are mostly based on event detection capacity. Some metrics examples are failed detections – FD, and detection error rate – DER, detection Accuracy – DeA, disaggregation accuracy – DiA, and overall accuracy – OA. Despite the lack of uniform performance metrics to measure the algorithm’s efficiency evaluation, energy disaggregation is becoming more and more feasible. The association of improving methods, tested and replicable using the public datasets, easy access to

data sampling, transmitting, and storing devices, and the development of IoT for residential appliances places energy disaggregation as an important tool that can contribute significantly to make the energy system more reliable, competitive, flexible, and sustainable, as indicated in Chapter 1.

The availability of this number and variety of load monitoring datasets makes some actual discussions about the best way to investigate one or other phenomenon possible, as for example in Tome et. al. (2021), where the sampling parameters are defined based on event detection. The proposal is very actual, and the methodology was tested in several public datasets, which makes the contributions even more interesting and opens the possibility of several future works.

2.3 ENERGY MANAGEMENT SYSTEM AND DEMAND RESPONSE

The combination of Appliances Load Monitoring (ALM) with the possibility of integrating small renewable energy sources to the grid (for example solar panels as small as a residential rooftop) makes the smart grid concept a key for smart energy consumption in the near future. To achieve the energy saving strategies, the residential installations saving potential cannot be ignored. Abubakar et al. (2017) give an overview of NILM associated with a home energy management system (HEMS). HEMS can indeed provide mutual satisfaction between customers by realizing their comfort preferences and the utility by assisting energy saving strategies. A deep understanding of appliances usage, given by NILM, can improve the HEMS capacity of both attend the households comfort preferences and make a smarter use of energy, even allowing a better integration of fluctuating energy resources. Load modulation together with the demand response projects can add flexibility to operate the smart grid.

If the energy service operators can have enough information to classify the customers and apply more targeted demand side response programs, the chance of getting a better result is expected to be higher.

In an empirical study, Ayres, Raseman and Shih (2013) analyzed field experiments that took place with approximately 75,000 households. The energy company randomly assigned a subset of these households to periodically receive mailed reports comparing their energy use with

neighbors. The households that received the letters showed significant energy use reduction compared to the ones that did not receive the feedback letter.

Cao, Beckel and Staake (2013) took a different approach and used the Irish CER Dataset, with readings for more than 4000 residential customers for 18 months over 30-minute intervals to classify customers based on demand peak position using k-means algorithm. The results allow the identification of the more demanding profiles, making the feedback programs more targeted. The methodology starts with a pre-processing step that verifies the inputs and discharges the invalid entries. Then the data is split in two subsets: summer and winter. After that, the authors used three clustering methods (hierarchical clustering, k-means and Self Organizing Maps associated with k-means) with different input parameters (mainly different distances definition), and at least the clusters quality is compared. The preprocessing step used here is very close to the one used in the work that is the object of this thesis. Both methods start with a cleaning step, with the objective of verifying the quality of the input data. The authors also used Principal Component Analysis (PCA) to reduce the data dimensionality, and k-means and Self Organizing Maps (SOM) as clustering methodologies. The main difference between this work and the one presented in this thesis is that the first one is clustering load curves, this is, consumers. The second one is clustering system status, or individual appliances.

Beckel, Sadamori, Staake and Santini (2014) reported an improvement of the previous work by inferring the household characteristics such as number of occupants and information related to the occupancy using smart meter data by adding a step for classification and regression. Some of the characteristics inferred were the number of appliances, number of bedrooms, type of cooking facility, floor area, and even the yearly household income. The main input data is the aggregated kW measurements from the smart meter. After a feature extraction step, a list of several input information is given to the classifier. The classifier results in assigning eighteen households characteristics that can be used to make the demand response programs more targeted. This work is very interesting because it can perform a real classification of the residences, and the information can be used to make specific demand response programs to several groups. The inconvenience of the method relies on the fact that the features selected for the classifier training were extracted from previous interviews performed with the households. In this way, its success depends on the people answering the form correctly. Also, it is possible that many may find the questionnaire invasive,

and do not agree to participate, or they can simply lie because they do not want their personal information revealed. Figure 2.5 shows the steps sequence of the methodology proposed.

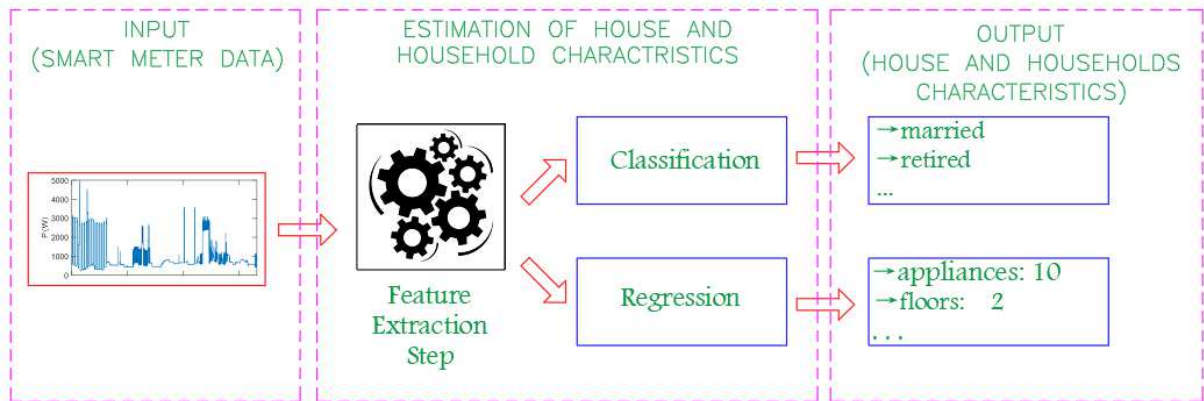


Figure 2.5: House and Household characteristics estimation. Adapted from Beckel, Sadamori, Staake and Santini (2014).

Vassileva and Campillo (2014) studied the impact of targeted user feedback. The focus was a specific class of users, namely low-income households, located in Sweden. The objective is to identify specific shared characteristics, interests, and preferences for energy visualization that can be used for keeping consumers interested in energy efficiency for a long period. The group studied approximately 2500 households, divided into two groups. After a questionnaire-based survey, the program evaluated aspects like:

- Preferred methods of electricity visualization (letter, web, mail, SMS, Apps or IHD).
- Main reasons for saving energy (Environment, money, or both).
- Factors influencing the consumer's willingness to purchase specific appliances (energy efficiency or price).

The questionnaires also asked questions regarding the household daily habits, like filling the dishwasher and washing machine before using it and trying to avoid standby mode. Despite the inconvenience of asking a lot of people to answer a questionnaire, in this work it was possible to show in numbers that being motivated to save energy, either because of the environment or to save money in the energy bill, is not enough. The average population is not specialist in energy efficiency, and thus needs help to make the correct decision in respect to such an aim. One good

way to provide this help is personalized feedback regarding the use of electricity and some suggestions to make a more intelligent use of it. Figure 2.6 shows the results of one of the questionnaires distributed.

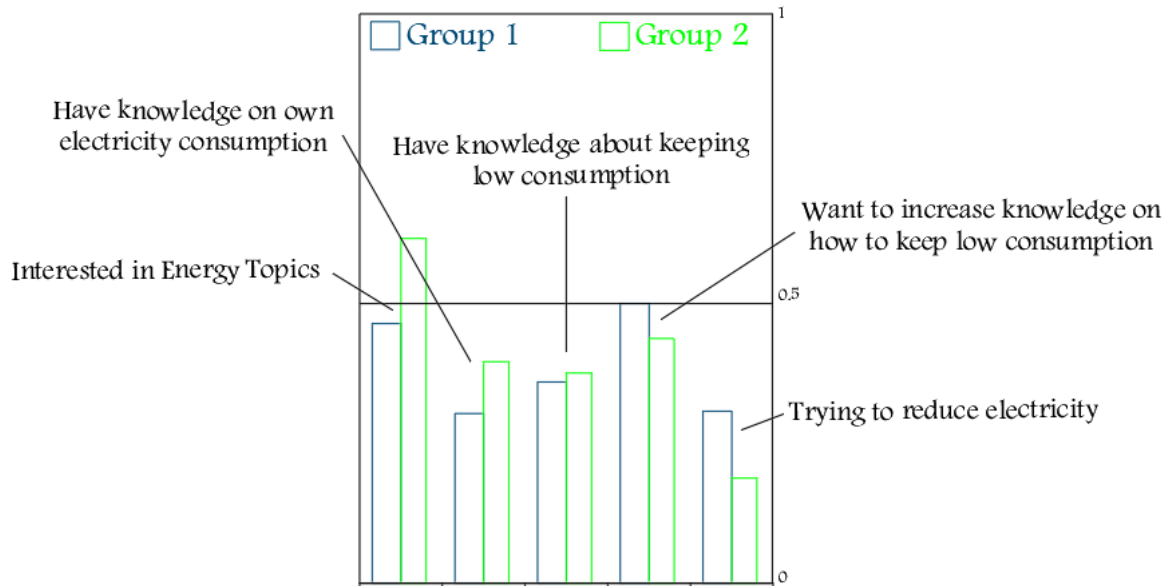


Figure 2.6: Example of questionnaire questions used by Vassileva and Campillo (2014), and its answers. Adapted from Vassileva and Campillo (2014).

2.4 TARGETED STUDIES

The availability of several public datasets with load management information, either aggregated, disaggregated, high and low frequency sampling, fundamental frequency and harmonics, individual appliance information, behavior questionnaires and others open the possibility to use data mining techniques to extract a huge set of information from the data collected. Comparing the actual scenario with the one in 1992 when Hart et.al introduces the NILM concept, the fast improvement of data storage and transmission capacities, the spread of Internet of Things (IoT), faster and cheaper data processing and the consolidation of machine learning algorithms places the challenges in different places. Now the scientific community faces a huge set of real world load monitoring samples, and we know that there is a lot of useful and interesting information in it. The question is how to extract it.

For example, several studies refer to smart campus facilities improvement using IoT and data mining tools. Yorio et. al (2018) uses IoT to improve the campus transit system with focus also in data security. The actual scenario in smart campus projects is described in Yang et. al (2018) and Muhamad et. al (2017), with projects that use IoT and data security to simulate smart cities facilities.

For residential installations, Chen, Chu, Tsao and Tsai (2013) propose a nonintrusive method to find the main activity performed inside the house based on the total power consumption. A classifier was trained to detect household characteristics like family size (number of people), age and employment (or retirement) status, among other characteristics only by monitoring the total real power consumption. The method achieved accuracy of more than 80%. Cao, Beckel and Staake (2013) two clustering techniques were used to study usage patterns and classify customers based on the demand daily peak position. The techniques used were k-means and Self Organizing Maps (SOM).

Pätäri and Sinkkonen (2014) brings a discussion regarding the assertiveness of the energy saving goals proposed from the Energy Service Companies. The need for energy efficiency improvement brings good business opportunities, but the work shows that the energy service companies have failed to implement effective energy performance contracts. Some of the reasons, according to the authors, are lack of knowledge about energy saving programs from the customer side, and a load model that is not as good as it could be, and the conclusions include a closer relation between the energy supplier and consumer, which can be achieved by a detailed and targeted feedback letter.

The question that comes at this moment is: what is the best way to prepare a good feedback letter? First, it is necessary to have a deep understanding of the consumer's profile, but this is not a simple task. For example, a program for reduce the use of energy by optimizing the use of electrical heating needs a good model of electrical heating usage, or to make a peak shaving/load modulation with targets on the most consuming loads (air conditioning/ space heater, ironing, electrical stove/oven), the usage patterns of this equipment must be studied. Also, the researchers must use methods as non-intrusive as possible. To add a novel contribution in this field requires an approach that involves not only the most consuming appliances, but able to observe all the residential appliances together.

The work presented in this thesis focuses on the observation of the dynamic behavior of all the residential appliances together, through the system status monitoring. Despite of the simple concept of “system status”, that is defined as “a binary vector containing the status of each individual appliance monitored (0 for “off” and 1 for “on”), the curse of dimensionality makes hard to visualize even a small system status dynamic in its raw state. At this point the methodologies of dimensionality reduction become important, as it will be described in the following chapters.

The methodology goal is to identify groups of appliances that are often used at the same moment, but not necessarily because they are used for the same (or similar) purpose (for example TV and video game, or oven and cooktop). In other words, we look for groups of appliances that are statistically related. Figure 2.7 explains the main steps of the work.

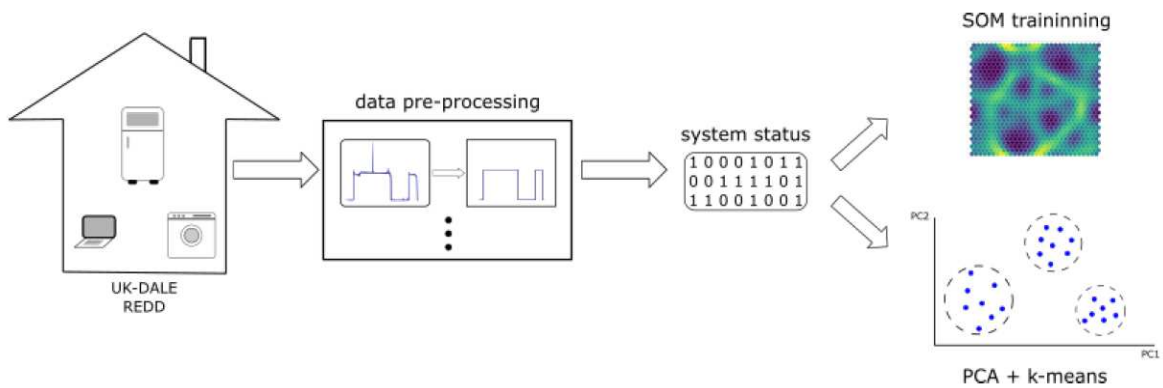


Figure 2.7: flowchart explaining the idea of this work.

This work is part of the state of art in the demand side studies of electrical energy use and brings a novel contribution with a method for noninvasive load patterns recognition, using algorithms of clustering and dimensionality reduction either method that has been consolidated for decades, such as Principal Component Analysis and k-means clustering, and actual algorithms for clustering and classification such as Self Organizing Maps.

3. METHODOLOGY

The main goal of this work is the definition of an algorithm that identifies utilization patterns in a fixed group of appliances. The focus is residential installations, and the definition of the appliances that are at the same group will not be guided by the type of appliance (for example toaster and mixer), but in a statistical analysis of a large set of system status. The information contained in these utilization patterns can be of great value for the development of more target and effective energy saving strategies. For example, the energy supplier companies can classify the customers according to the energy saving potential and suggest some small behavior changes (according to the groups) resulting in a more intelligent use of electricity.

The main analysis starts from a large set of vectors representing the status of each monitored appliance (system status). This way, each position of the binary vector states if the specific device is on (“1”) or off (“0”).

During the development of this work particularly, the system status vectors were obtained from an intrusive monitoring system, this means that one monitoring device was installed to follow each specific appliance (or plug). This is a limitation factor for the system size: in a house with 20 devices, it is necessary to have 20 monitors, including installing, synchronizing, collecting data and dealing with malfunctioning.

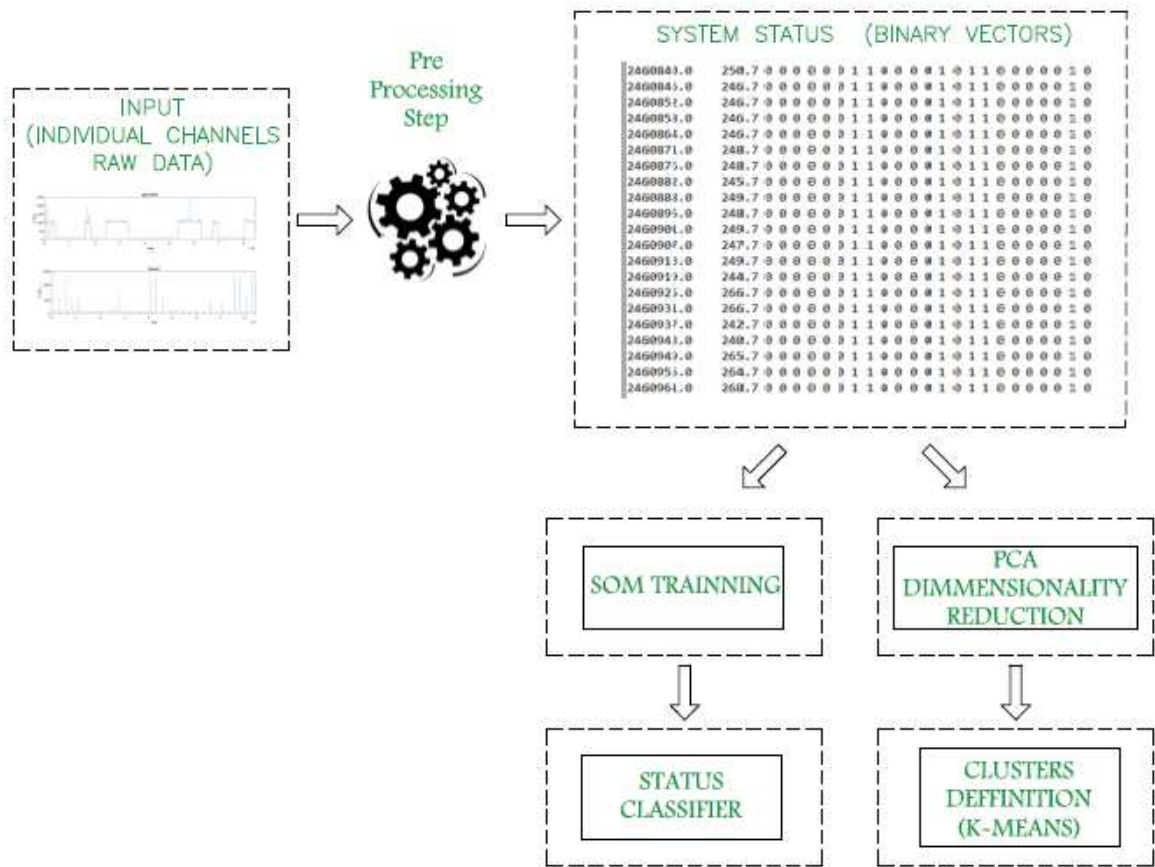


Figure 3.1: Representation of how the information is organized and the work main steps

Note that the intrusive scenario is important to test the proposed solution. It is expected that this framework will be used in practical scenarios as a combination of the pattern recognition, and a fast and effective disaggregation algorithm. The patterns recognition methodology is well developed in this work, but the association with a fast and assertive disaggregation algorithm is mandatory for the application to become feasible in a daily basis. Only this way it will be possible to collect and process data from residences in a non-intrusive way, this is, without the inconvenience of having a person inside the house to install the monitoring devices and avoiding the hardware costs.

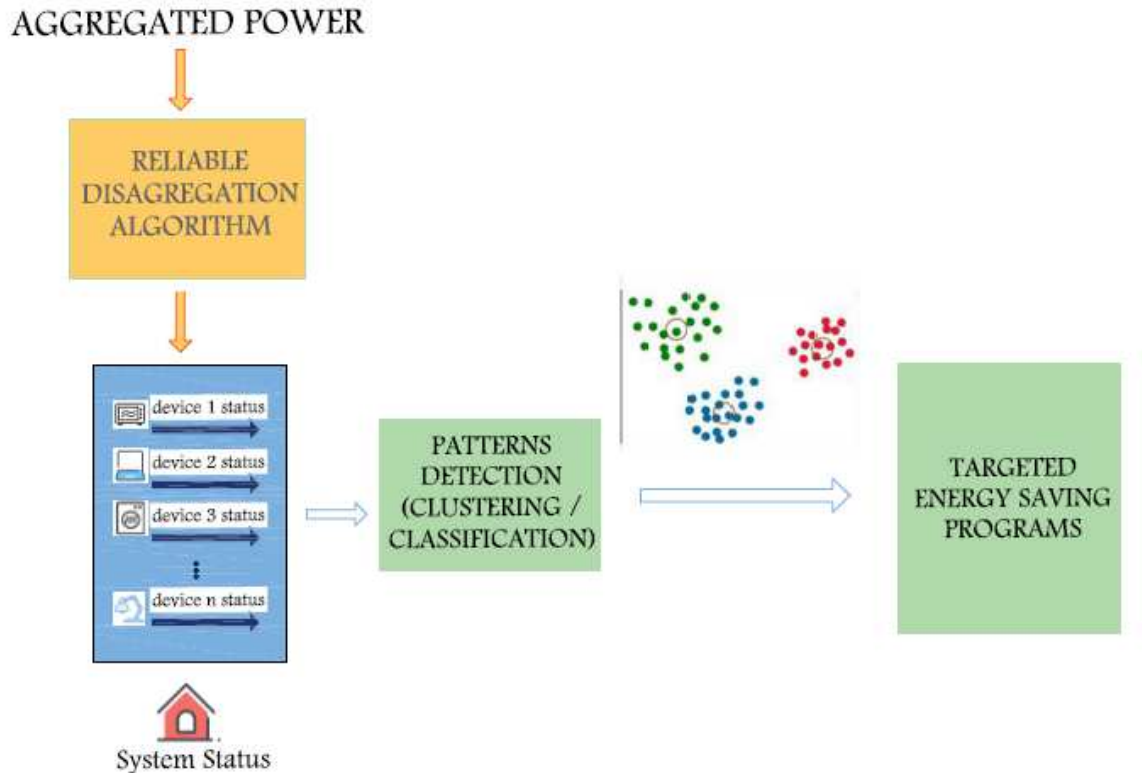


Figure 3.2: Combination of disaggregation algorithm with utilization patterns detection

3.1 DATA DESCRIPTION

The datasets selected in this study are the UK-DALE (Kelly and Knottenbelt, 2014) and REDD (Kolter and Johnson, 2011). The UK-DALE dataset contains measurements from five different households in the 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).

The REDD dataset contains measurements from 6 different households in the United States with 3-second granularity, with monitoring periods from 2.7 to 25 days. The monitoring periods are not the best for detecting utilization patterns (2.7 days can measure unfortunately some holiday, for example, and even 25 days does not take different seasons in the year), but on the other hand this dataset records a good quantity of individual channels (18, 9, 20, 18, 24, 15), each of them representing one individual appliance.

Table 3.1: Number of individual channels and time monitored for every house in UK-DALE dataset

HOUSE	INDIVIDUAL CHANNELS	TIME MONITORED
UK-DALE HOUSE 1	52	more than 4 years
UK-DALE HOUSE 2	18	193 days
UK-DALE HOUSE 3	4	35 days
UK-DALE HOUSE 4	5	151 days
UK-DALE HOUSE 5	24	122 days
REDD HOUSE 1	18	25 days
REDD HOUSE 2	9	11 days
REDD HOUSE 3	20	14 days
REDD HOUSE 4	18	19 days
REDD HOUSE 5	24	2.7 days
REDD HOUSE 6	15	13 days

The results recorded from these two datasets contain aggregated information (total real power measurement from the entire house) and disaggregated (real power from each individual channel). UK-DALE dataset has also some information about the appliances (called metadata), including appliances description, manufacturer, model, and a suggestion of “on power”. For individual channels there are also some files called “button press”, that indicate the moment when the specific appliance was turned on. However, this information is not complete, and the preprocessing step was necessary to create the binary vectors representing the system status. REDD dataset has also some additional sampling in a higher sampling rate for houses 3 and 5, but this information was not used for this work.

The information in Table 1 leads to a system state domain of high dimension: R^{52} , R^{18} , R^4 , R^5 and R^{24} and R^{18} , R^9 , R^{20} , R^{24} and R^{15} , respectively for UK-DALE and REDD. Considering one sample every 6 seconds for UK-DALE, the set of system status in one month has more than 400,000 measurements. For REDD, with one sample every 3 seconds, the set of system status has 28,800 measurements for each monitored day. In addition, if each system status is represented by a binary vector that indicates each appliance’s status (i.e., on or off, or 1 or 0), the number of possible statuses for each house is 252, 218, 24, 25 and 224 for UK-DALE and 218, 29, 220, 224

and 215. These numbers make traditional statistical analysis limited. To address this problem, performing the dimensionality reduction is a key step to make the problem more tractable to extract information. Once the system status is in \mathbb{R}^2 or \mathbb{R}^3 , the distances between the points (appliances) become visible in a chart, and a clustering step (linear with k-means and nonlinear with SOM) will define the groups of appliances whose use are related to each other.

3.2 PRE-PROCESSING

As described in the previous section, the results recorded from these two datasets contain aggregated information and real power sampling from each individual channel. In this work, the first task is to transform the real power values from the individual channels into a binary vector representing the whole system status, one for each sample. The pre-processing step was responsible for performing this task.

Figures 3 and 4 show the intermediate results for 2 appliances in UK-DALE House 1. For every individual channel it was determined a limit of real power above which the device status was considered “on”. The possibility of improving the recognition algorithm using disaggregation techniques, such as appliance signatures (transients during switch on and off) and machine cycles is known, but at this moment the focus is on dimensionality reduction and clustering. Figure 3.3 shows the proposed approach with its main steps.

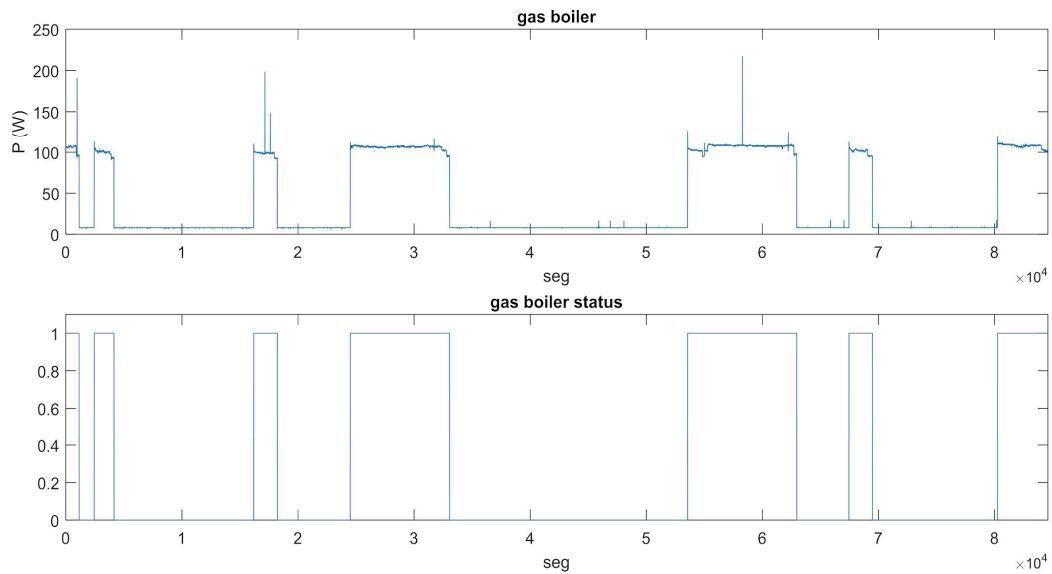


Figure 3.3: correspondence between kW sampling and system status for gas boiler, House 1, UK-DALE.

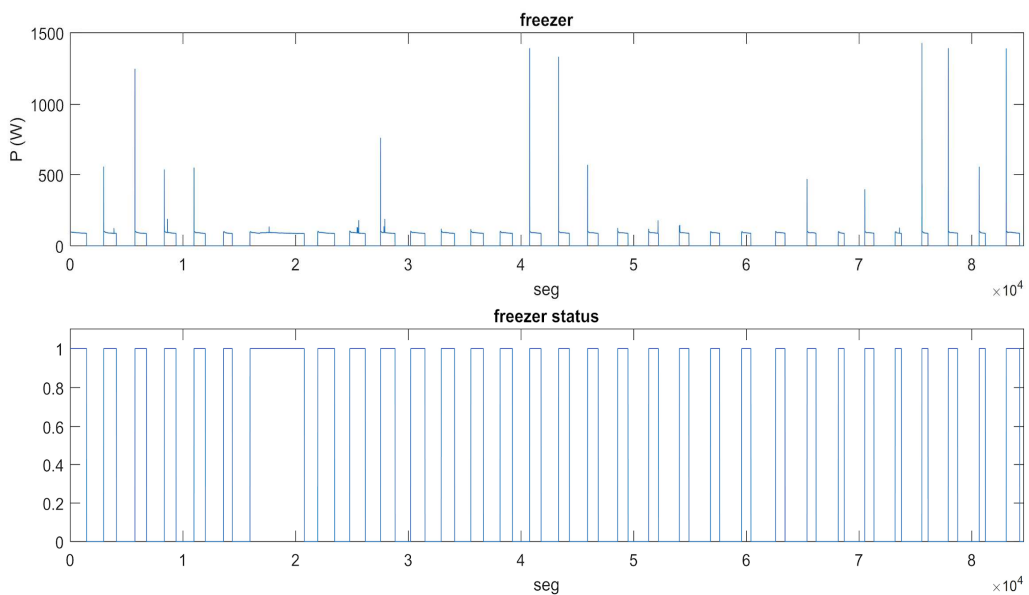


Figure 3.4: correspondence between kW sampling and system status for freezer, House 1, UK-DALE.

The next step is to compile the individual channels status into a single system status. This is a very simple task if the timestamp of every sample for all channels are the same. Unfortunately, this does not happen in every house. Often it encountered a lapse of measurement in some channel, from 1 missing sample to several minutes. It is also possible that some channels have no

missing sample, but a delay of 1 second during some periods. If some individual channels contained a quantity of samples too different from the others in the household (less than 25% of average for the installation), the channel was removed. Finally, the system status was recorded as binary vectors in a text file. The time stamp and the total power in kW are still present in the output file generated by the pre-processing algorithm (see Figure 3.5 for an example).

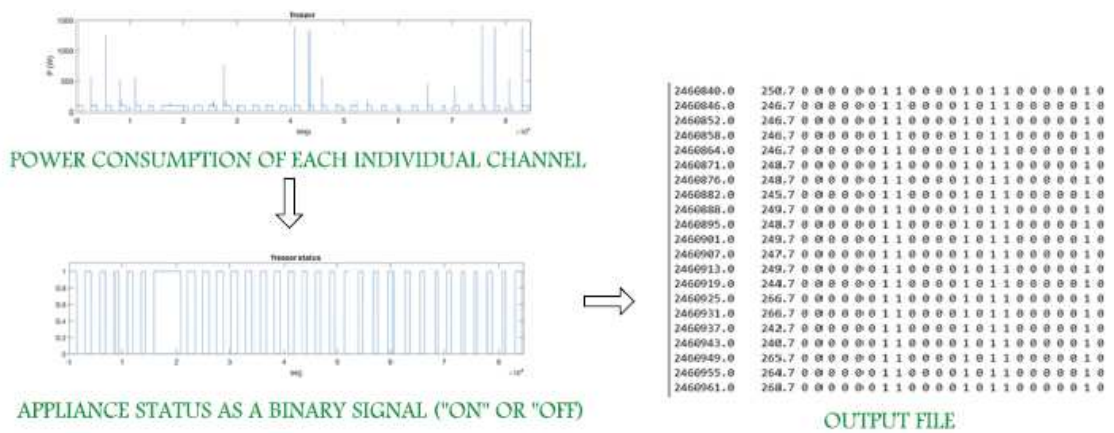


Figure 3.5: the pre-processing step transforms the real power measurements from individual channels into a binary vector containing each appliance status. The output file also keeps the time stamp and the system aggregated real power instantaneous demand.

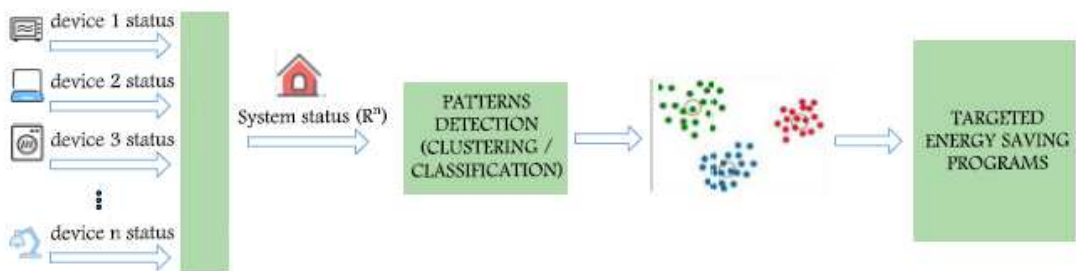


Figure 3.6: Project flowchart.

3.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis (PCA) is a linear and non-parametric method that projects high dimension data into specific directions that lead to variance maximization, and noise and redundancy minimization.

The central question of PCA is: How to define a matrix P that multiplies by the sample's 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 related 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 $i = 1, \dots, m$, and (b) x_j as the column vector (samples) of matrix X for $j = 1, \dots, n$, then each line of the resulting matrix Y can be written as:

$$y_i = [p_{iX_1} \ p_{iX_2} \ \dots \ p_{iX_m}]$$

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 represent the interesting data in X ? Now if 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:

$$C_X = [(n - 1)XX^T]^{-1},$$

where $(n-1)^{-1}$ 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 contain the covariances between them. 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 Jonathon Shlens (2015) 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, until m directions were found.

The main method outputs are the projection direction p_i (as many directions as the original dataset dimension), called Principal Component (PC), and the variance associated with each PC. It is an established and simple method for extracting relevant data from noisy or confusing data sets (which can be large or not). The main contribution of PCA in this work is the capacity of resume redundancy (with the sampling frequency considered, it is expected that each appliance remains in the same status for several samples) in a simple, linear, and fast algorithm.

The main advantages of PCA are:

- The method does not require any parametrization; that means that no additional information of the phenomenon is required.
- Requires a small computational effort.
- The variances associated to each projection 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.
- As the method is nonparametric, it is impossible to set it in the wrong way and miss an important result. The method does not show any hidden information nor suggests any conclusions, instead it reduces redundancy and shows the data from an angle where the information is easier to visualize. In our case specifically, we are not looking for any specific cluster, only for a way of visualizing the system status set in a 2D or 3D projection, so it will be possible to analyze the data in a simple plot.

The right quantity of principal components necessary to represent the phenomena without significant loss of information is not the same from one problem to another, but it is always related to the total variance accumulated from the PCs. One way to choose how many PCs are enough to represent the dataset is to plot the total variance accumulated for every set of PCs. The method, called “the elbow method”, applied to this perspective, states that the point where the curve has its elbow (if it has one) is the ideal number of PCs that represent the dataset without significant loss of information (variance). If, for example, the curve has no elbow, it means that all the PCs are equally important and the PCA is not a good tool to perform dimensionality reduction.

Figure 7 shows the variance curve for the UK DALE dataset, HOUSE 2. From the 6th PC on, the PCs have an insignificant contribution to the dataset's variance. This means that the information they carry, if ignored, has a small impact on the information captured.

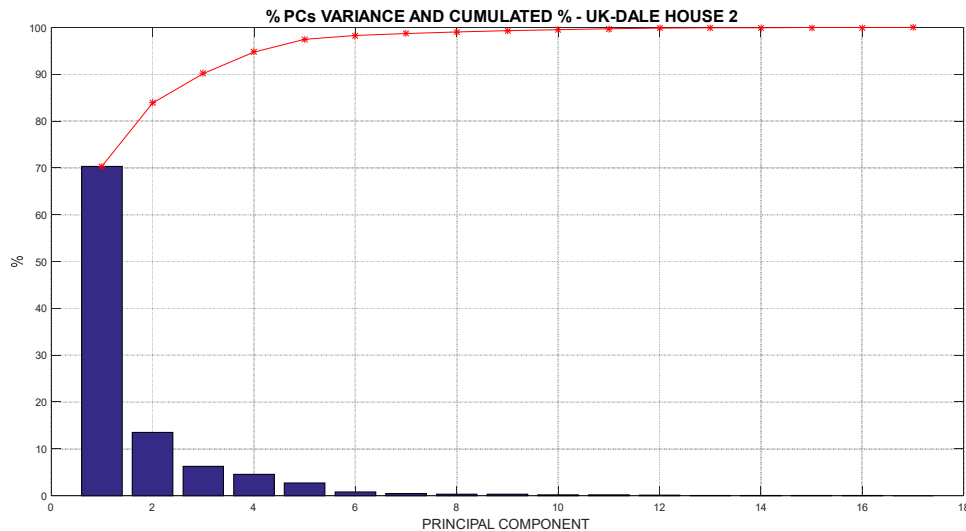


Figure 3.7: Variance contribution for each Principal Component, HOUSE 2, UK-DALE

If the first 3 PCs together add a high percentage of the total variance, and the total variance curve shows one inflection point at the 3rd PC, this means that the 3D projection of the original data is enough for capturing the data behavior.

For more details about the method and its mathematical formulation, refer to Jonathon Shlens (2005).

About how to apply PCA to perform the dimensionality reduction in our problem: the UK-DALE dataset is composed of five measurement sets of houses in the United Kingdom with 52, 18, 4, 5 and 24 individual channels. The data is 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} . For the REDD, the consumption information of 6 houses is taken in a 3-second frequency, with 18, 9, 20, 18, 24 and 15 appliances monitored. After the pre-processing, the system status dataset is recorded in a text file, with the system status in the lines. 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 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.

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 8).

Simulations were made using MATLAB R2016a over the preprocessing output files, and the results are described in Chapter 4, and discussed in Chapter 6.

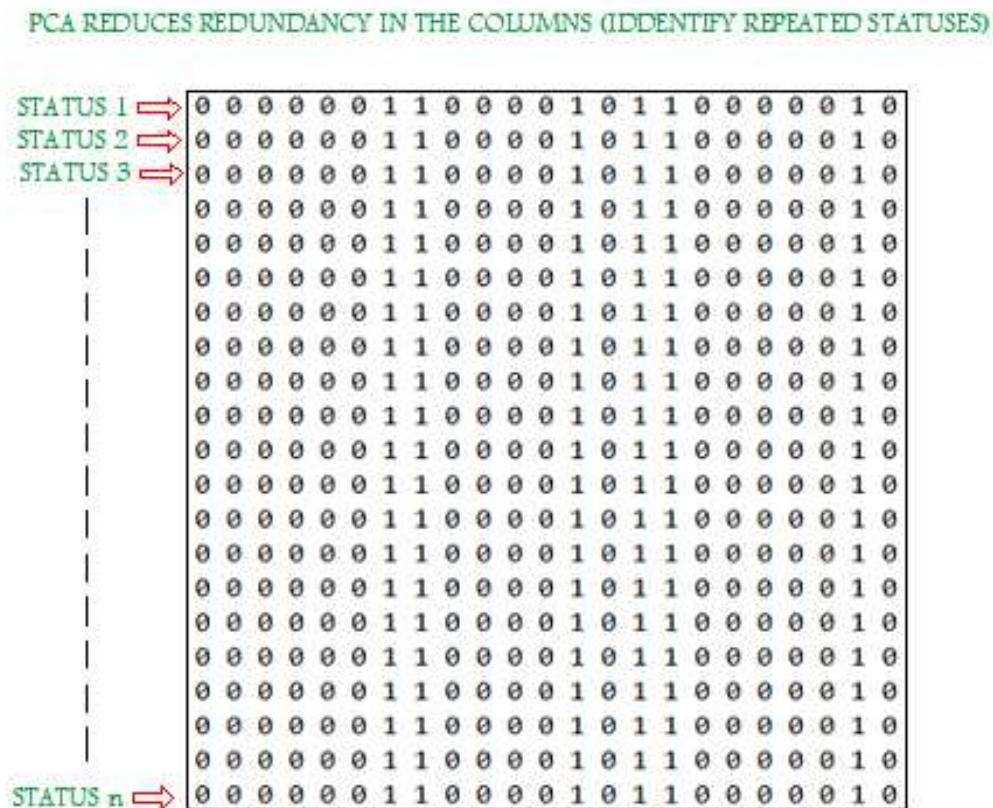


Figure 3.8: Illustration of matrix X for redundancy reduction

3.4 K-MEANS

After PCA is performed the dimensionality reduction, the projection over the first three principal components (or more, according to the elbow method and the total variance curve) are used to define clusters. For the simulations it was used MATLAB R2016a, and the cluster definitions were made according to k-means method Jain and Dubes (1988). There are several clustering algorithms available, and there isn't any restrictions to use other technique to perform this task. The exploration of other clustering algorithms is described as a possibility of future work.

To define the best configuration of clusters (the ideal number of clusters - k), it was also used the elbow method. Varying the number of clusters from 1 to m (1 cluster means all the appliances together while m clusters mean 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. The selection of k comes directly from it.

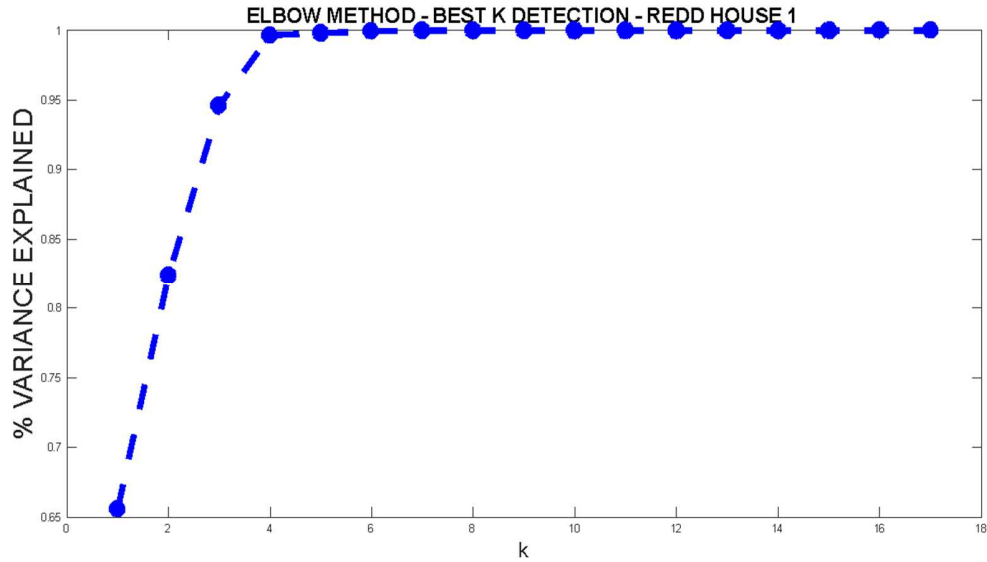


Figure 3.9: Variance Explained according to the number of clusters (k), for REDD, House 1. The best k is 4.

Table 3.2: Clusters for REDD House 1, $k=4$.

CLUSTER INDEX	CHANNELS
1	5-KITCHEN OUTLETS 1 6-KITCHEN OUTLETS 2
2	1-OVEN 1 2-OVEN 2 3-REFRIGERATOR 4-DISHWASHER 8-WASHER DRYER 1 9-MICROWAVE 10-BATHROOM GFI 11-ELECTRIC HEAT 12-STOVE 13-KITCHEN OUTLETS 1 14-KITCHEN OUTLETS 2 17-WASHER DRYER 2 18-WASHER DRYER 3
3	7-LIGHTING
4	15-LIGHTING 2 16-LIGHTING 3

3.5 3D VISUALIZATION AND THE MINIMUM SPANNING TREE

The result for this part of the work is to show graphically the agglomerations between the appliances. This visualization is impossible from the original data because of the high dimension of the system status.

One way of visualizing the clusters is to plot the data projection over the Principal Components. This work can be done with one, two or three Principal Components and still use a simple plot chart to make the results visible (a 3D plot). However, at this point the nature of the variance curve is not known, and it is possible that the elbow method suggests more than 3 PCs to

visualize the system status dynamics with minimum loss of information. In this case, another visualization tool is used: the Minimum Spanning Tree.

Considering each one of the appliances as a node from a graph, independently from the data dimension, the Euclidian distance between the nodes can be calculated. The Minimal Spanning Tree (MST) is a tool that gives the shortest path between any two nodes (appliances) of a connected set. This representation allows one to see a result from any dimension in a 2D figure without any loss of information. Also, the plotting scale can be manipulated to fit all the data in one single figure. The method used to generate the MSP was PRIM. Figures 3.10 shows an example of a MST generated from Matlab R2016a. The figures shown in Chapter 4 were made using software Autocad 2021.

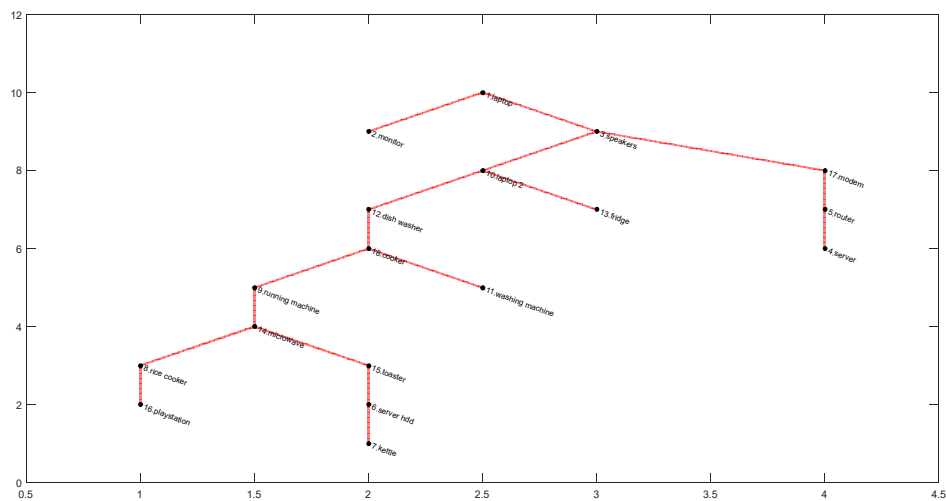


Figure 3.10: MST as output from Matlab R2016a. To make visualization easier, the figures shown in Chapter 4 were made using software Autocad 2021.

3.6 SELF ORGANIZING MAPS

A Self Organizing Map (SOM) is a tool from the set of unsupervised learning algorithms in machine learning, described in details by Haykin (2008). It consists basically in a grid of neurons that are connected to each other by a specific topography, usually in \mathbb{R}^1 (linear), \mathbb{R}^2 or \mathbb{R}^3 .

The training of a SOM starts with a predefined number of neurons (in this case the neuron model is simply a vector in \mathbb{R}^n , being n the dimension of the system status).

The dimensionality reduction occurs in the neurons' connections: they are arranged in a linear or bidimensional way. During the training, the neurons move toward the samples, reflecting agglomerations of similar data that can be observed after the training is finished through the connection's length.

The training of a SOM uses the concepts of *competitive learning* and *winner neuron* – see Haykin (2008). This way, when some sample is presented to map, the neurons will compete to decide which one best represents that specific data. In our case, the neuron closest (according to Euclidean distance) to the sample will be the winner and will move towards the sample (see figure 3.10). Previous parametrization of the method states if the winner neurons move alone, or if it also takes the neighbors. In our case, both the winning neurons and the neighbors moved.

One iteration of the training is completed when every sample of system status is presented to the map, and the neurons have moved according to them. After many iterations (hundreds), the neurons' structure will show agglomerations that reflect sample correlations.

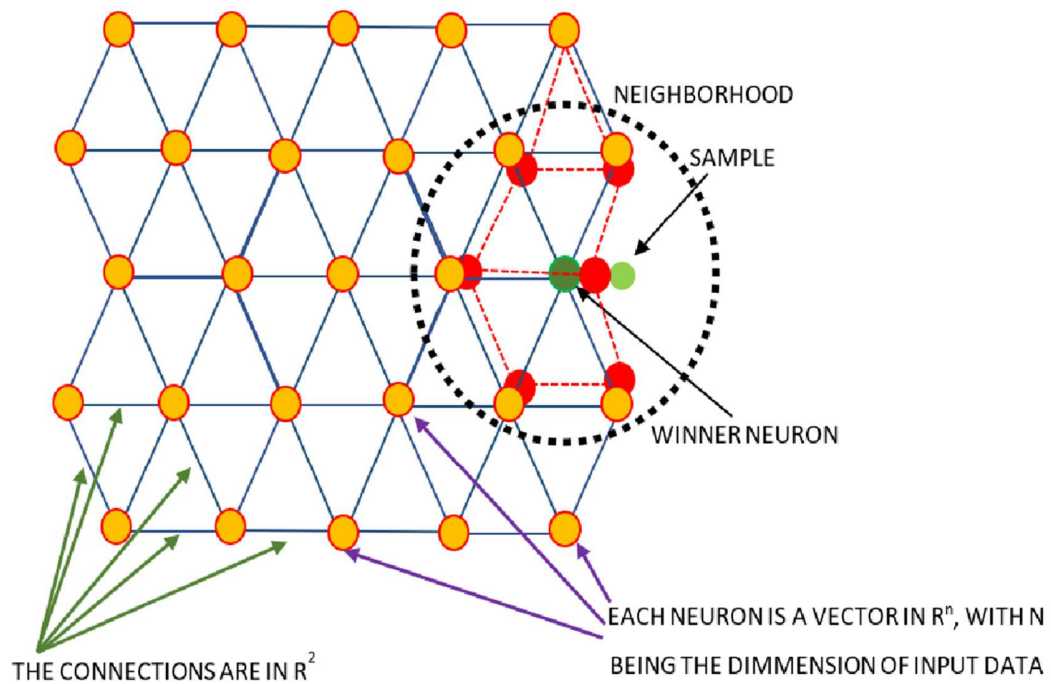


Figure 3.11: 2D hexagonal geometry and the winner neuron concept. During each training iteration, the winning neuron moves toward the sample and moves to the first neighborhood. The dimensionality reduction occurs because each neuron, a vector in R^n , is connected to

After training, the initially uniformly distributed neuron's structure presents agglomerations that indicate quantitatively a natural classification of the samples set. Referring to Figure 3.10, some neurons moved and became very close to each other, other connections became larger. The agglomerations exist and suggest some classification between the samples, but in most cases, a detailed analysis of the distances distribution to define the groups is needed.

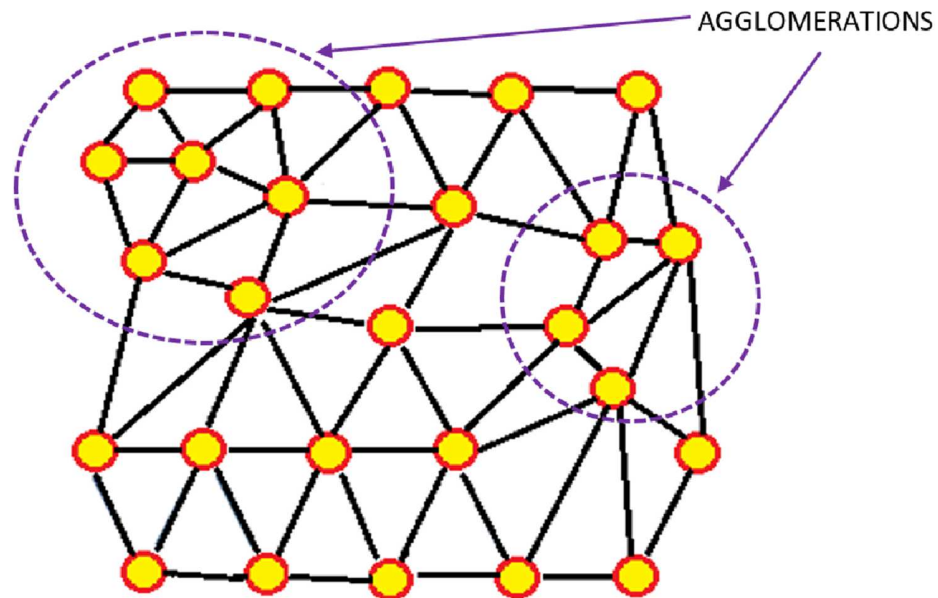


Figure 3.12: Initial hexagonal SOM after training. When looking at the distances between the neurons, the agglomerations became visible.

In this case, Python was employed to apply the SOM algorithm through the library SimpSOM. As a machine learning tool, the results depend on the size of the map, the shape (connections between the neurons) and the number of training epochs. Other factors also can influence the result, but these ones are the most important.

To be able to compare the results, the training parameters were the same for all houses and datasets:

- 3D grid with 40x40 neurons.
- hexagonal connections.
- 30.000 epochs of training.

4. RESULTS – PRE PROCESSING, PCA AND K-MEANS

The methodology described in Chapter 3 was applied to two public datasets selected: UK-DALE and REDD. In UK-DALE there are two houses with a small number of individual channels (Houses 3 and 4 with 4 and 5 individual channels respectively), but the period monitored is longer than in the REDD Houses. This means that we can expect as a result utilization patterns closer to reality than in REDD. Another difference between the datasets is the sampling frequency: UK-DALE has one sample every 6 seconds, while REDD has one sample every 3 seconds.

The first step of the preprocessing routine was to analyze if the channels in each house are compatible in sample quantity. During the sampling period, some individual channels can experience one or more outage periods due to sampling device malfunction or any other reason that leads to loss of data. The main goal of this first step was to identify channels that have a sampling quantity too different from the other channels in the same house and discharge it. If the outage period is not significant, the last appliance status (on or off) is assumed to be unchanged during the period.

The results are:

- In UK-DALE House 1, channels 22 (hoover), 39 (hair dryer), 40 (straighteners) and 41(iron) were excluded.
- In UK-DALE House 5, channels 11 (PS4) and 25 (vacuum cleaner) were excluded.
- In REDD, it was not necessary to exclude any individual channel due to sampling blackouts.

The exact quantity of samples in each individual channel for both datasets is summarized in Tables A.1 to A.11 in Appendix A.

But even after these eliminations, the sample quantity is still very different from one channel to another inside each house, and from one house to another. To make the results comparable, it was considered for every house in UK-DALE only first 504,000 samples (corresponding to 35 days of continuous sampling). For the REDD dataset, the number of days considered in each house is listed in Table 4.1.

Table 4.1: Number of samples (continuous days) considered for each House in REDD dataset.

HOUSE FROM REDD DATASET	SAMPLING DAYS	SAMPLES
HOUSE 1	25	720,000
HOUSE 2	11	316,800
HOUSE 3	14	403,200
HOUSE 4	18.5	532,800
HOUSE 5	2	57,600
HOUSE 6	13	374,400

The next two steps of the processing were to transform real power demand into a binary status, and then make the synchronization of the channels. The binary status of each individual channel was determined based on a threshold value of real power consumption above which the device is considered on. Each channel has its own “on/off” limit, that was determined graphically.

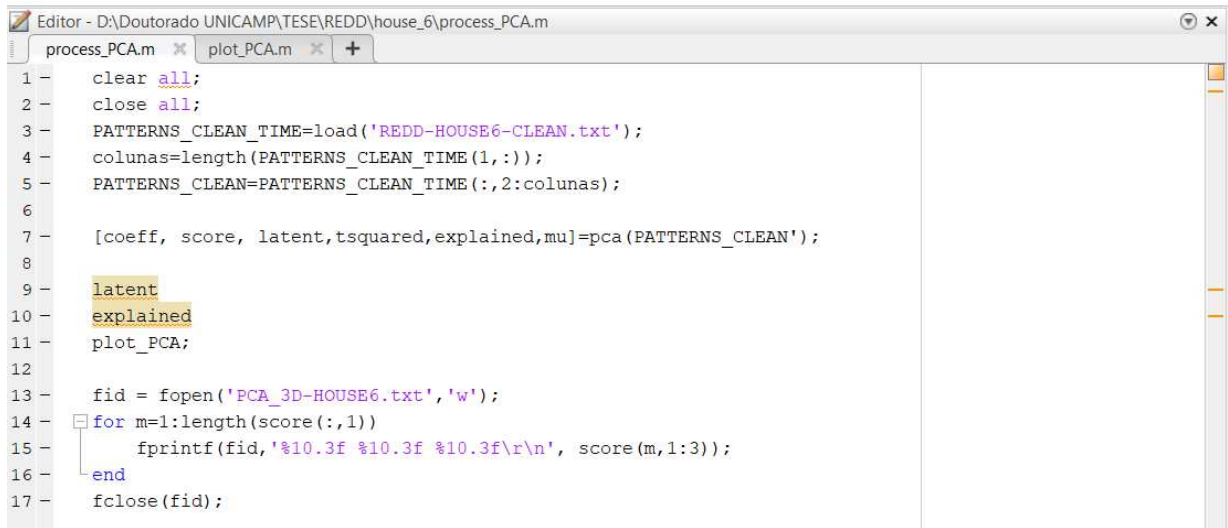
The synchronization step was necessary to deal with periods of sampling gaps. For that, the main channel time stamp was the reference, and in case of a gap, the last status (“on” or “off”) was maintained until the real status was back. Finally, the system status was recorded in a txt file, containing timestamp, House’s aggregate power, and the status for every individual channel, as shown in Figure 4.1

Timestamp	Real Power	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5	Channel 6	Channel 7	Channel 8	Channel 9	Channel 10	Channel 11
0.0	715.7	0	1	0	1	0	0	1	1	1	0	0
6.0	748.7	0	1	0	1	0	0	1	1	1	0	0
12.0	711.7	0	1	0	1	0	0	1	1	1	0	0
18.0	728.7	0	1	0	1	0	0	1	1	1	0	0
25.0	718.7	0	1	0	1	0	0	1	1	1	0	0
31.0	712.7	0	1	0	1	0	0	1	1	1	0	0
37.0	521.7	0	1	0	1	0	0	1	1	1	0	0
43.0	626.7	0	1	0	1	0	0	1	1	1	0	0
49.0	714.7	0	1	0	1	0	0	1	1	1	0	0
55.0	1969.7	0	1	0	1	0	0	1	1	1	0	0
62.0	1965.7	0	1	0	1	0	0	1	1	1	0	0
68.0	1930.7	0	1	0	1	0	0	1	1	1	0	0
74.0	1889.7	0	1	0	1	0	0	1	1	1	0	0
80.0	1921.7	0	1	0	1	0	0	1	1	1	0	0
86.0	1920.7	0	1	0	1	0	0	1	1	1	0	0
92.0	1928.7	0	1	0	1	0	0	1	1	1	0	0
98.0	1933.7	0	1	0	1	0	0	1	1	1	0	0
103.0	1922.7	0	1	0	1	0	0	1	1	1	0	0
109.0	1890.7	0	1	0	1	0	0	1	1	1	0	0
115.0	1922.7	0	1	0	1	0	0	1	1	1	0	0
121.0	741.7	0	1	0	1	0	0	1	1	1	0	0
140.0	715.7	0	1	0	1	0	0	1	1	1	0	0
146.0	520.7	0	1	0	1	0	0	1	1	1	0	0
195.0	702.7	0	1	0	1	0	0	1	1	1	0	0
201.0	710.7	0	1	0	1	0	0	1	1	1	0	0
207.0	1970.7	0	1	0	1	0	0	1	1	1	0	0
213.0	1913.7	0	1	0	1	0	0	1	1	1	0	0
219.0	1907.7	0	1	0	1	0	0	1	1	1	0	0

Figure 4.1: Example of output file after preprocessing step. The files contain: time stamp, total house's real power, and the binary status of each individual channel.

1.1 PRINCIPAL COMPONENT ANALYSIS

The Principal Component Analysis was made using software Matlab R16a. The code for House 6 in REDD is shown in Figure 4.2.



```

Editor - D:\Doutorado UNICAMP\TESE\REDD\house_6\process_PCA.m
process_PCA.m x plot_PCA.m x +
1 - clear all;
2 - close all;
3 - PATTERNS_CLEAN_TIME=load('REDD-HOUSE6-CLEAN.txt');
4 - colunas=length(PATTERNS_CLEAN_TIME(1,:));
5 - PATTERNS_CLEAN=PATTERNS_CLEAN_TIME(:,2:colunas);
6
7 - [coeff, score, latent,tsquared,explained,mu]=pca(PATTERNS_CLEAN');
8
9 - latent
10 - explained
11 - plot_PCA;
12
13 - fid = fopen('PCA_3D-HOUSE6.txt','w');
14 - for m=1:length(score(:,1))
15 -     fprintf(fid,'%10.3f %10.3f %10.3f\r\n', score(m,1:3));
16 - end
17 - fclose(fid);

```

Figure 4.2: code used in Matlab R2016a to process PCA.

PCA outputs in software Matlab are:

- Coeff: directions p_i (for data projection).
- Score: original data projected over the directions p_i .
- Latent: contains the variance of each Principal Component.
- Explained: contains the total variance cumulated, in percentage, as each Principal Component is considered. It starts with the percentage variance of the first PC. The second position is the percentage variance of the first and second PCs together, until they reach 100% (all PCs are included).

The outputs “tsquared” and “mu” were not used, and for this reason are not described above.

It is important to note that the reference data for the PCA algorithm is not the system status set matrix, but the transposed. This means that the PCA looks for redundancy not among the system status, but among the appliances behavior. This way, if the behavior of two appliances are very similar, the appliances will be points close to each other in the results.

The results shown refer to the “clean” version of the system status set. This means that the samples that had all the appliances as “off” were removed.

In the following pages, it is shown, for House 4 in UK-DALE dataset, the following figures:

1. Variance behavior: the bars represent the variance of each PC, and the red line the accumulated variance, in %. The behavior of this red curve determines if the PCA is a good tool for dimensionality reduction or not.
2. PCA 1D: this picture shows the data projection (appliances behavior) over the first Principal Component. The relation between appliances is already visible.
3. PCA2D: shows the data projection over the first PC in x-axis, and over the second PC in y-axis. Now some appliance agglomerations from PCA1D can appear separated (or not).
4. PCA3D: shows a 3D plot with the data projection over the first, second and third PC in the x, y and z-axis. These pictures show the most information captured from the first 3 PCs and is the farthest we can go without any other visualization routine/analysis.

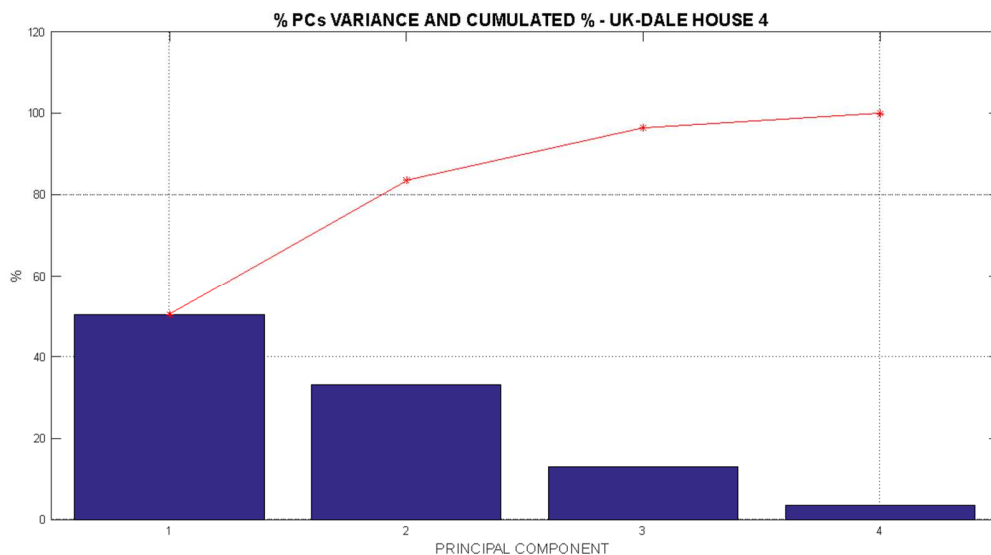


Figure 4.3: Variance for each PC and cumulated for UK-DALE, House 4.

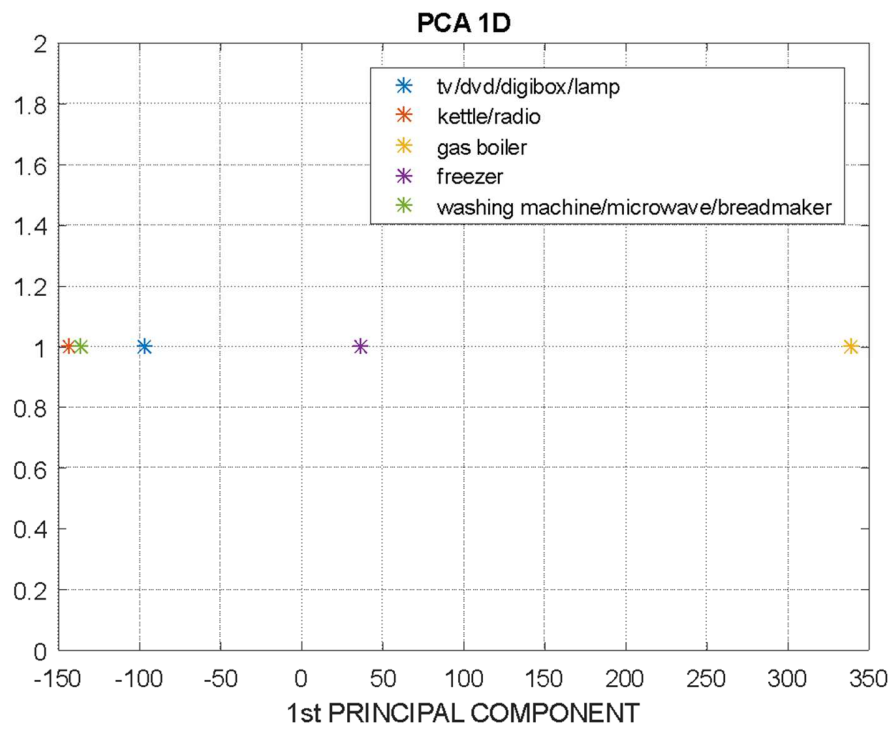


Figure 4.4: Data Projection over First Principal Component - UK-DALE, House 4

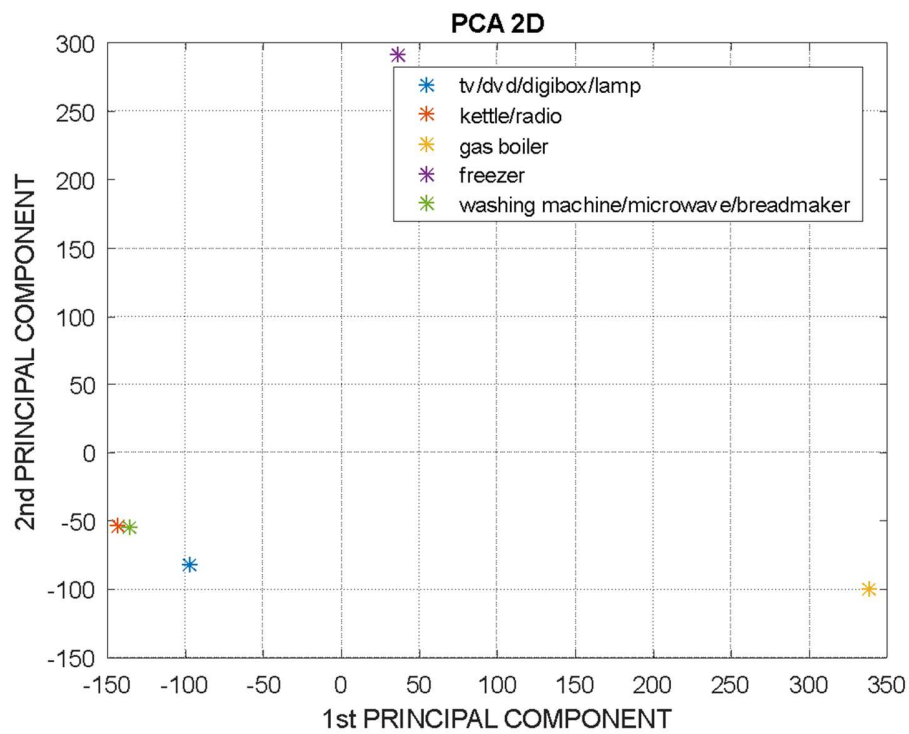


Figure 4.5: Data Projection over First and Second Principal Components - UK-DALE, House 4

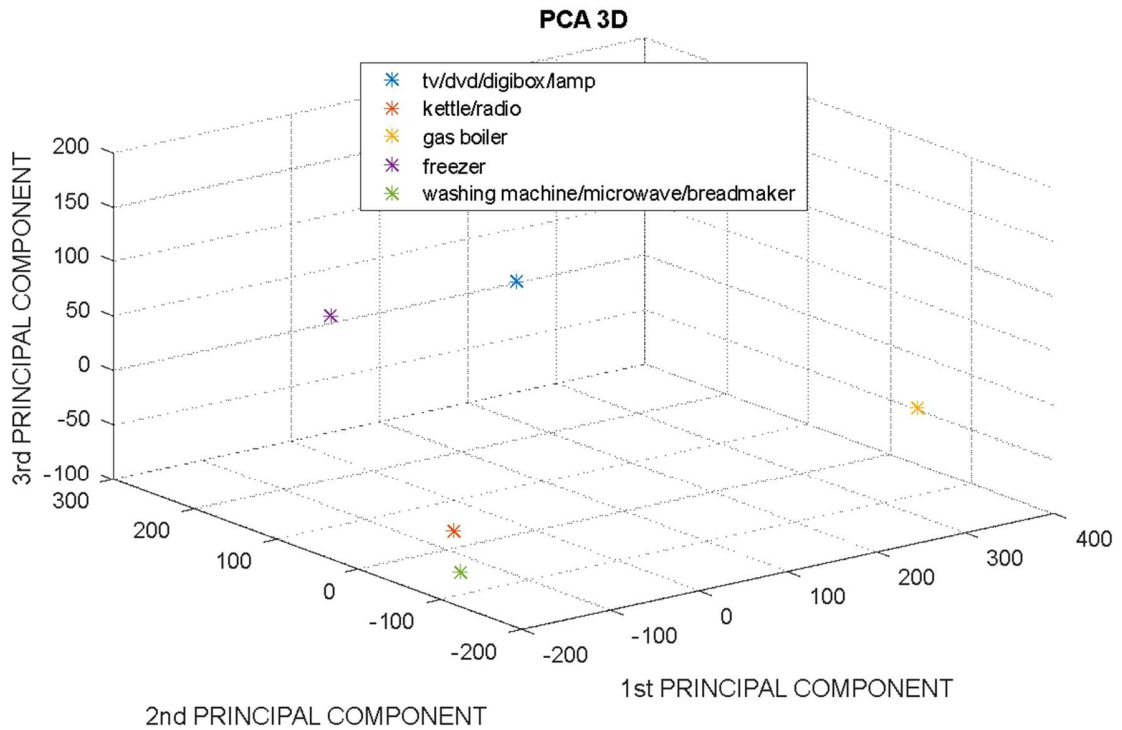


Figure 4.6: Data Projection over First, Second and Third Principal Components - UK-DALE, House 4

The same figures for UK-DALE and REDD datasets are shown in Appendix 2.

1.2 MINIMUM SPANNING TREE AND K-MEANS

The PCA results in the previous Section showed that the relation between some appliances do exist. However, even with the 3D plot, it is impossible to make a deep analysis of the groups. At this point the Minimum Spanning Tree toll makes the results visualization easier. For that it was used software MATLAB R2016a, and the code is shown in Figure 4.7.


```

clear all;
close all;

POINTS_PCA_3D=load('PCA_3D-HOUSE1.txt');
distances=dist(POINTS_PCA_3D);
node_names={'1-boiler','2-solar thermal pump','3-laptop','4-washing machine','5-
dishwasher','6-TV','7-kitchen lights','8-htpc','9-kettle','10-toaster','11-
fridge','12-microwave','13-lcd office','14-hifi office','15-breadmaker','16-amp
livingroom','17-adsl router','18-livingrooms lamp','19-soldering iron','20-gigE &
USBhub','21-kitchen dt lamp','22-bedroom ds lamp','23-lighting circuit','24-
livingrooms lamp2','25-iPad charger','26-subwoofer livingroom','27-livingroom lamp
tv','28-DAB radio livingroom','29-kitchen lamp2','30-kitchen phone&stereo','31-
utilityrm lamp','32-samsung charger','33-bedroom d lamp','34-coffee machine','35-
kitchen radio','36-bedroom chargers','37-gas oven','38-data logger pc','39-childs
table lamp','40-childs ds lamp','41-baby monitor tx','42-battery charger','43-
office lamp1','44-office lamp2','45-office lamp3','46-office pc','47-office
fan','48-LED printer'};
G=graph(distances,node_names,'upper','OmitSelfLoops');
T=minspantree(G);
figure;
p2=plot(T,'EdgeLabel',T.Edges.Weight,'EdgeColor','r','NodeColor','k','LineWidth',
2);
hold on;

```

Figure 4.7: Matlab R2016a code for generating the MST for UK-DALE, House 1

Figures 4.8 show the MST result for REDD, House 1. The results for all UK-DALE and REDD Houses are shown in appendix 3. In all cases, the appliances with distance smaller than 1 was considered as a single point. The figures were generated with software AutoCAD 2021.

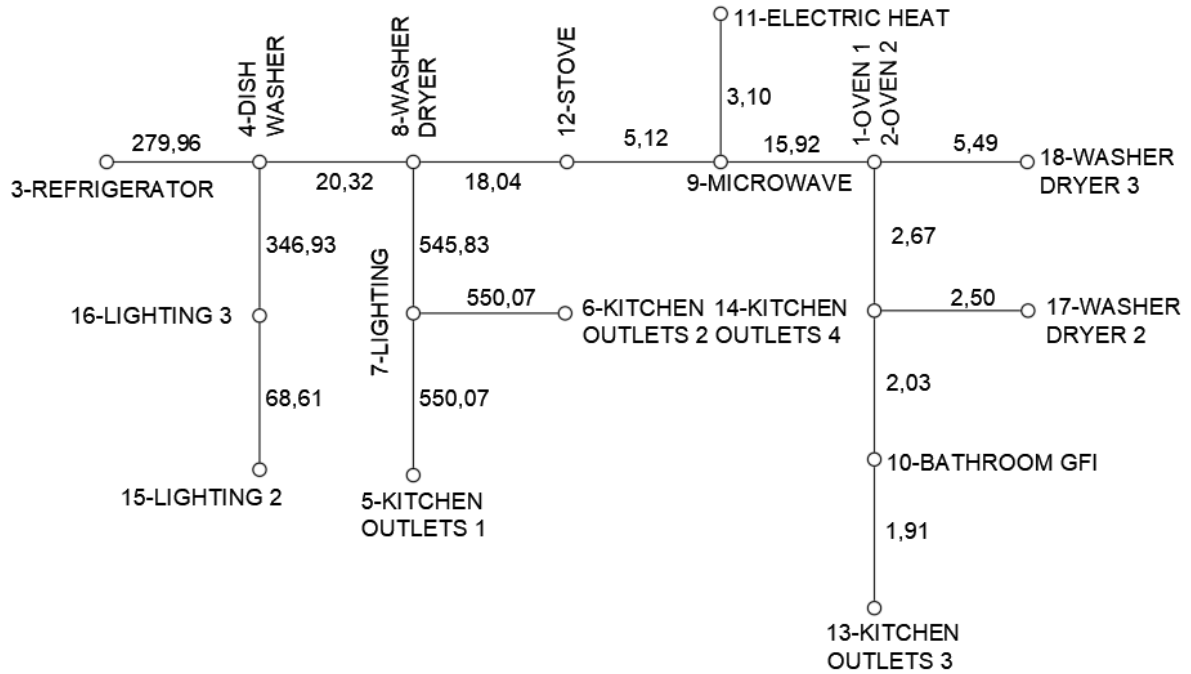
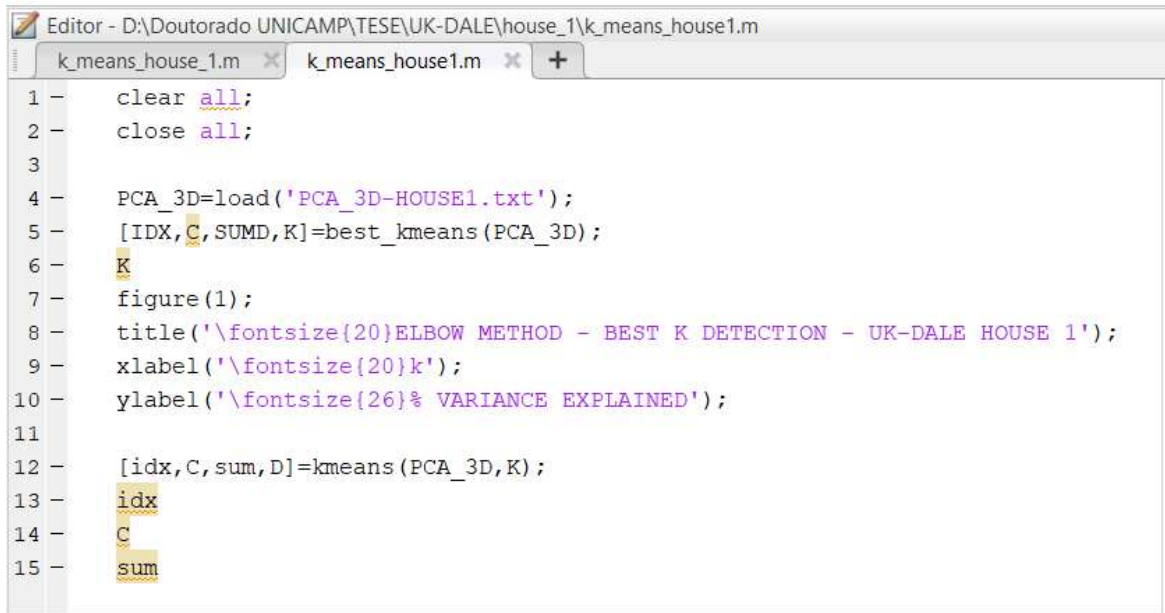


Figure 4.8: MST for REDD, House 1, including the distances between nodes. The appliances with distances smaller than 1 were considered as a single node.

1.3 K-MEANS RESULTS FOR PCA 3D

The k-means method was used to determine the quantity and composition of the appliance's clusters. The programming was made in software Matlab R2016a, and the code is shown in Figure 4.9.



```

1 - clear all;
2 - close all;
3
4 - PCA_3D=load('PCA_3D-HOUSE1.txt');
5 - [IDX,C,SUMD,K]=best_kmeans(PCA_3D);
6 - K
7 - figure(1);
8 - title('\fontsize{20}ELBOW METHOD - BEST K DETECTION - UK-DALE HOUSE 1');
9 - xlabel('\fontsize{20}k');
10 - ylabel('\fontsize{26}% VARIANCE EXPLAINED');
11
12 - [idx,C,sum,D]=kmeans(PCA_3D,K);
13 - idx
14 - C
15 - sum

```

Figure 4.9: Code used in Matlab for k-means.

The function “best_kmeans” is an iterative routine that returns in “K” the best number of clusters also according to the elbow method in the variance explained curve.

The function “kmeans(X,k)” returns the “k” clusters in the “idx” vector. The outputs include:

- Idx: performs k-means clustering to partition the observations of the data matrix X into k clusters and returns a vector (idx) containing cluster indices of each observation. Rows of X correspond to points and columns correspond to variables. By default, kmeans uses the squared Euclidean distance measure and the k-means++ algorithm for cluster center initialization.
- C: returns the k cluster centroid locations in the matrix C .
- Sum: returns the within-cluster sums of point-to-centroid distances in the vector sum.
- D: returns distances from each point to every centroid in matrix D .

Appendix C shows k-means results in tables, and the variance curves used to determine the best k . The pictures ahead show the same results using the MST representation.

evening demand peak. Another very interesting node is the one that gathers kettle and soldering iron. These two appliances are not related at all, and still appear as “frequently used at the same time”. If we focus on the appliances with more significant rated power, one modulation opportunity that is easy to achieve is regarding the washing machine and the dishwasher. These two appear at the same cluster, and both are appliances that can be programmed to work during the night, for example.

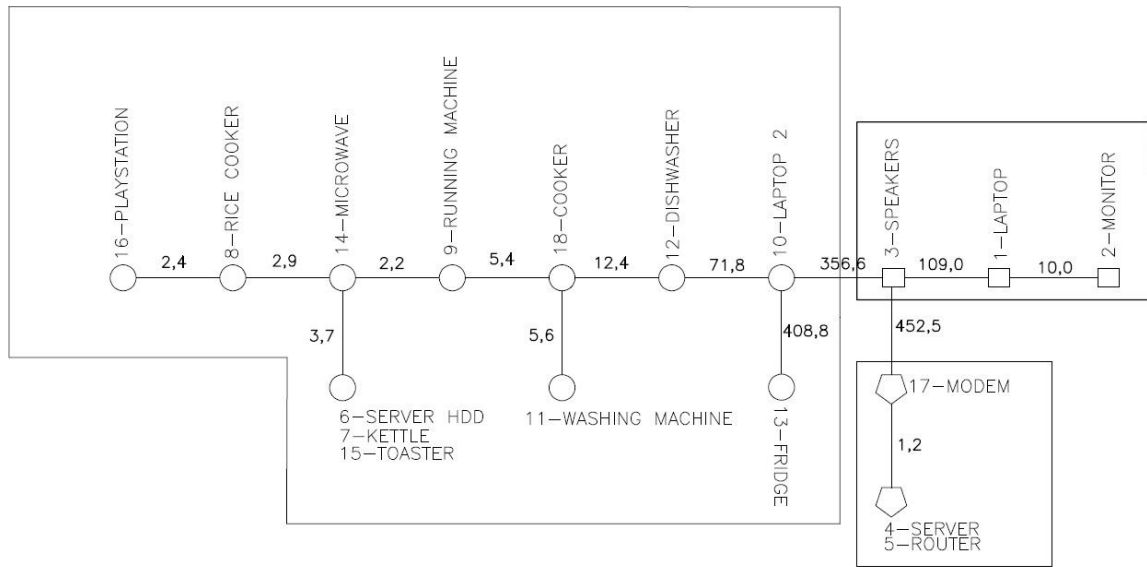


Figure 4.11: MST with k -means results represented over the MST for UK-DALE, House 2, with $k=3$.

The clusters in UK-DALE House 2 show the classic habit of using all the kitchen appliances at the same time (Rice cooker, microwave, kettle, toaster, cooker), but also reveals a good opportunity of modulation recommendation regarding the washing machine and dishwasher. Another interesting data that can be inferred from the results is that, at least during the monitored period, there are a minimum of 3 people inside the house: one to use the Playstation, one to use the running machine, and other one to operate the kitchen appliances. The two smaller clusters join the appliances that have their purpose related, so it is logical that they are in the same cluster: server, modem and router, and laptop, monitor and speakers.

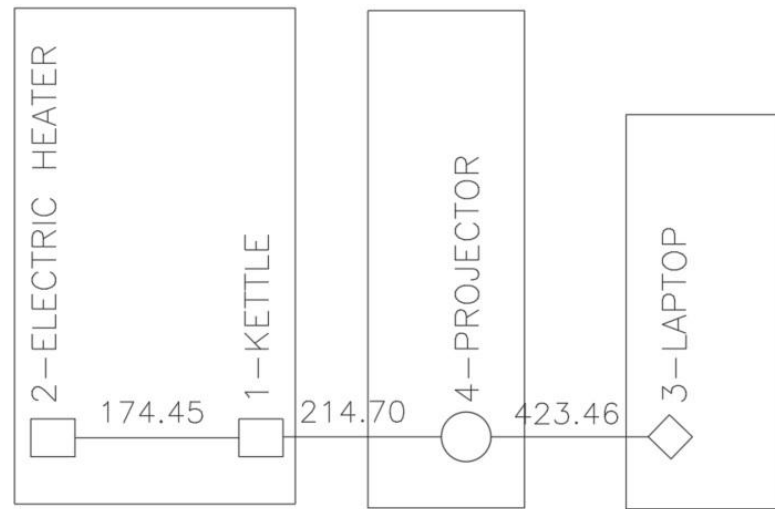


Figure 4.12: MST with k -means results represented over the MST for UK-DALE, House 3, with $k=3$.

The result for UK-DALE House 3 is not interesting because the information available is about only 4 individual channels.

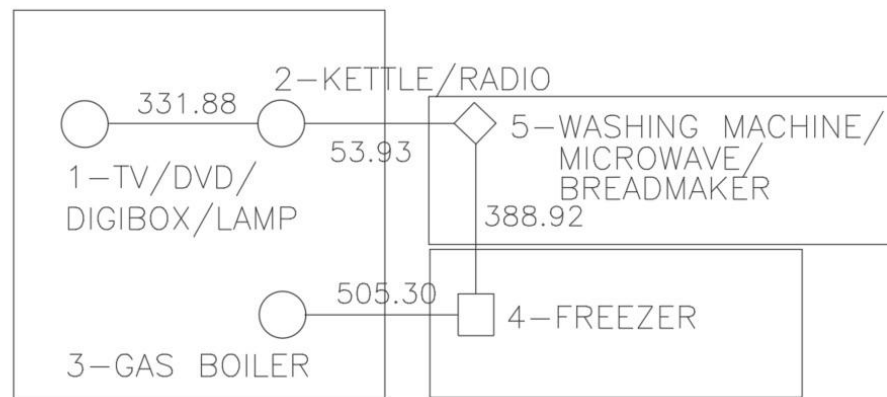


Figure 4.13: MST with k -means results represented over the MST for UK-DALE, House 4, with $k=3$.

UK-DALE House 4 also has too few appliances monitored to bring an interesting discussion. But even with so few appliances, this user could improve its energy efficiency by using the washing machine during the night, or at least not together with the microwave and breadmaker.

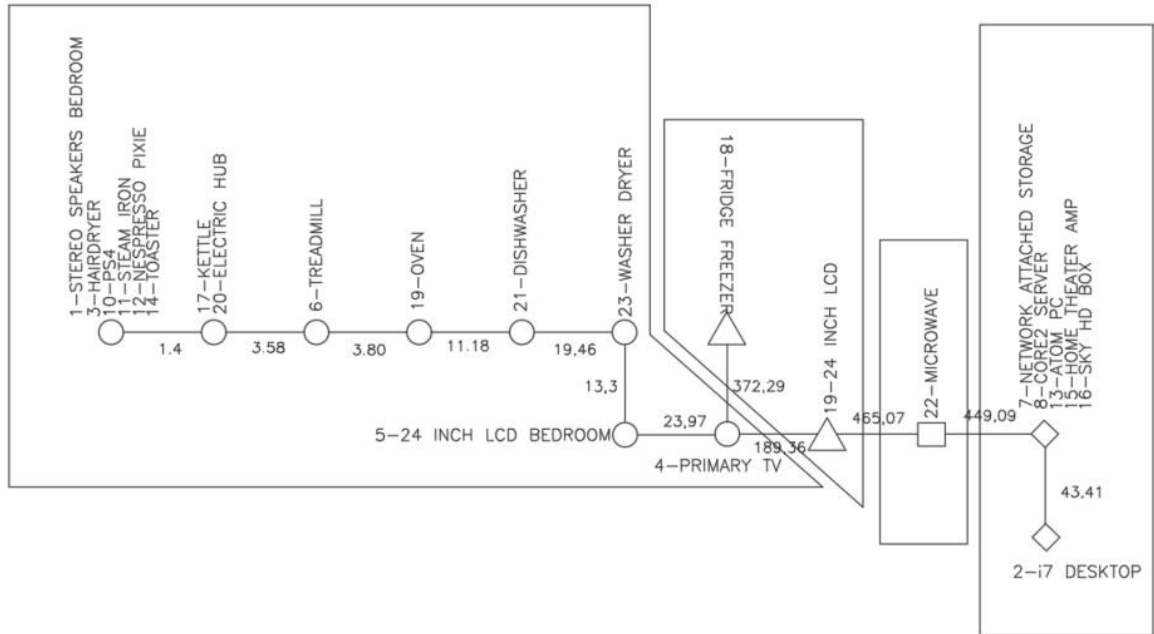


Figure 4.14: MST with k -means results represented over the MST for UK-DALE, House 5, with $k=4$.

For House 5 in UK-DALE, the interesting results are in the nodes with more than 1 appliance: stereo speakers' bedroom + hairdryer + PS4 + steam iron + Nespresso pixie + toaster, and network attached storage + core2 server + atom PC + home theater AMP + Sky HD Box. Here also it can be inferred that there are at least 3 people in the house, because it is impossible to use the hairdryer, PS4 and Steam iron at the same time. The dishwasher and washer dryer represent a load modulation opportunity in almost every house.

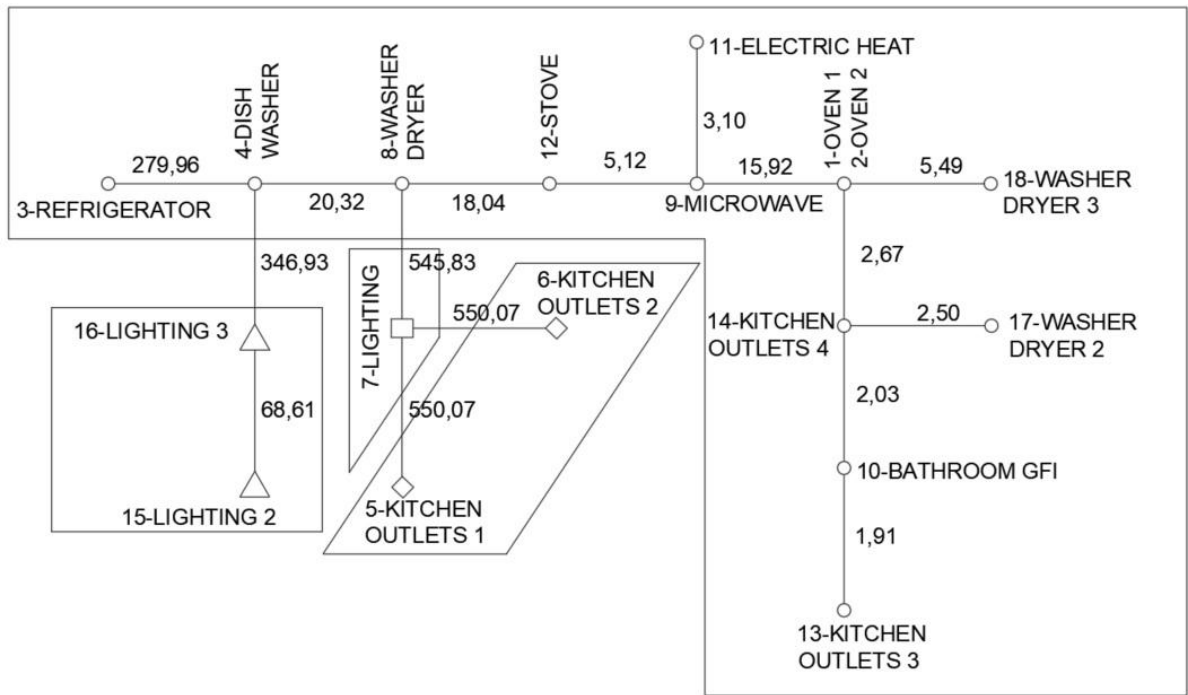


Figure 4.15: MST with k -means results represented over the MST for REDD, House 1, with $k=4$

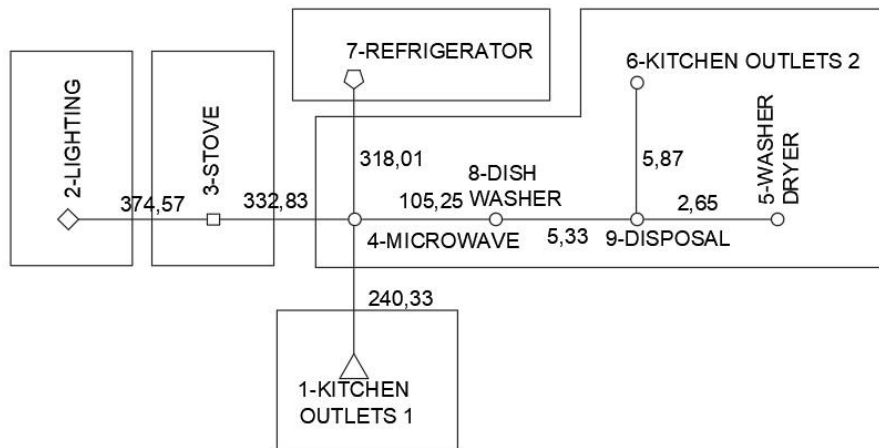


Figure 4.16: MST with k -means results represented over the MST for REDD, House 2, with $k=5$

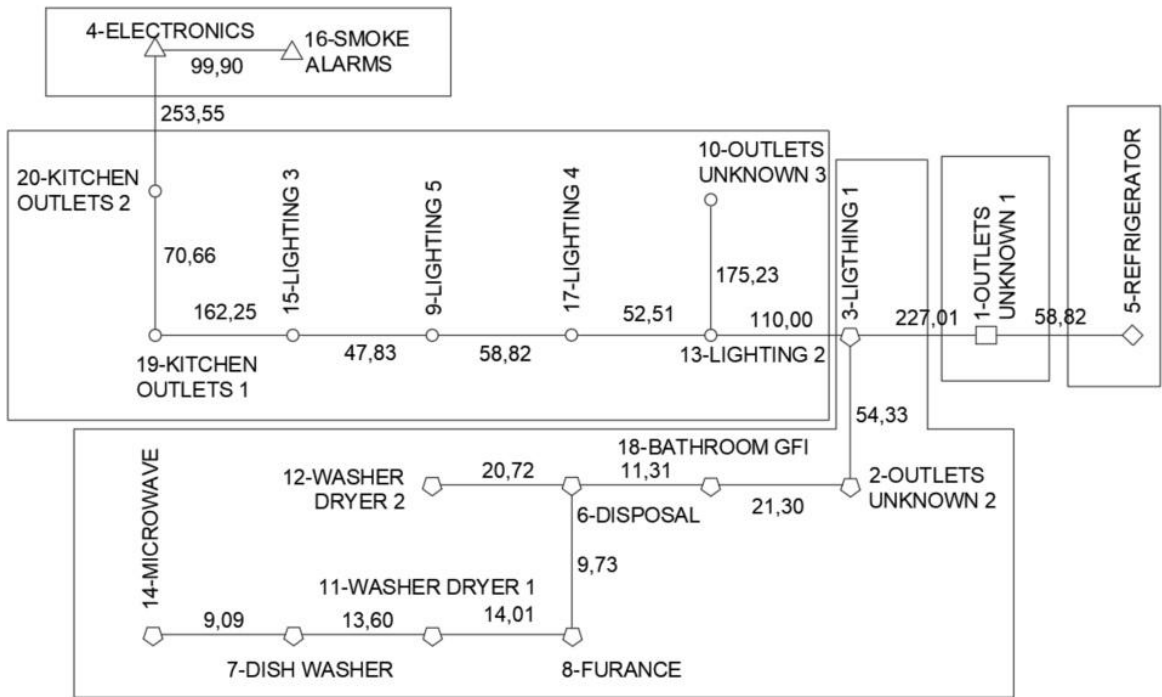


Figure 4.17: MST with k-means results represented over the MST for REDD, House 3, with k=5

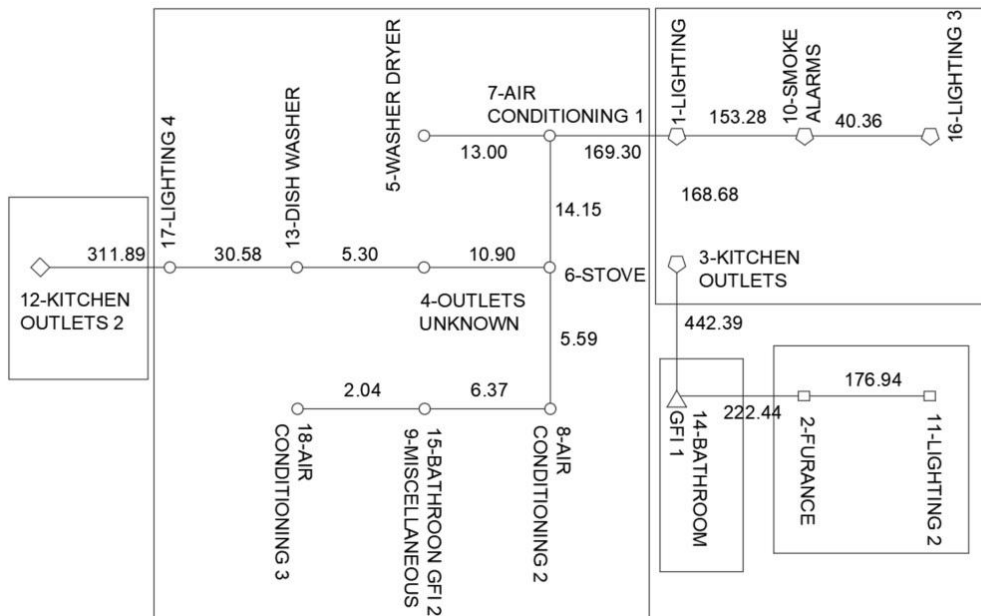


Figure 4.18: MST with k-means results represented over the MST for REDD, House 4, with k=5

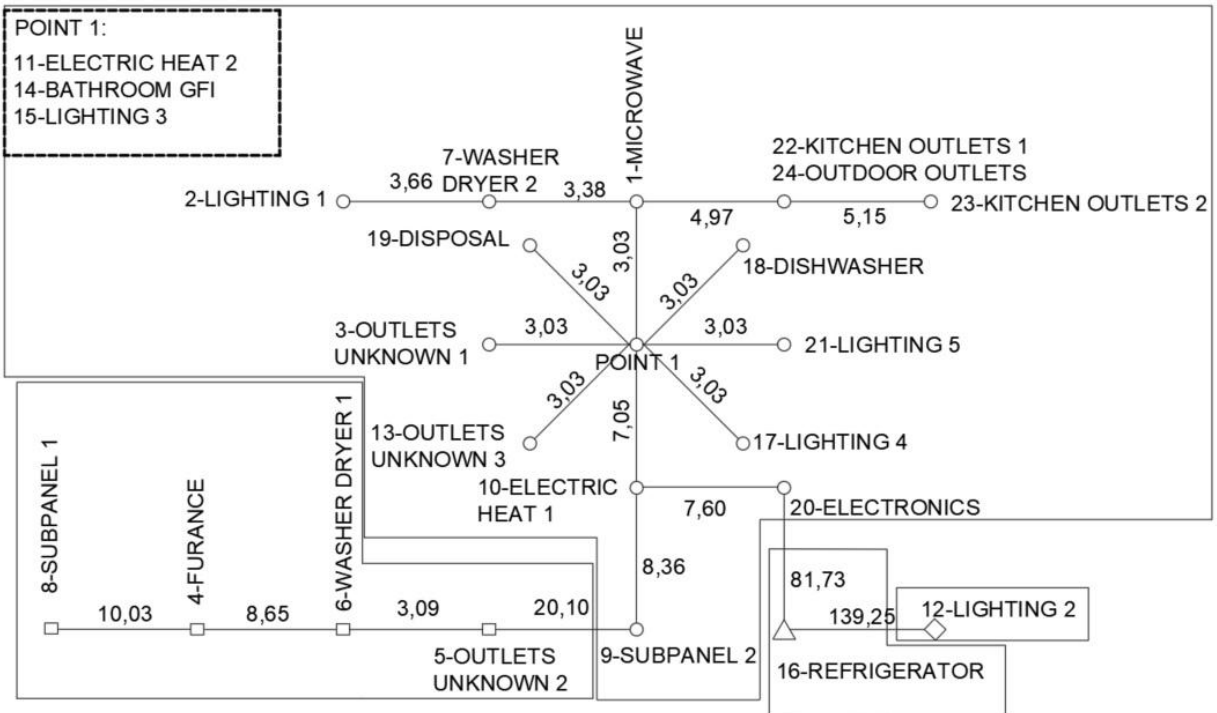


Figure 4.19: MST with k-means results represented over the MST for REDD, House 5, with k=4

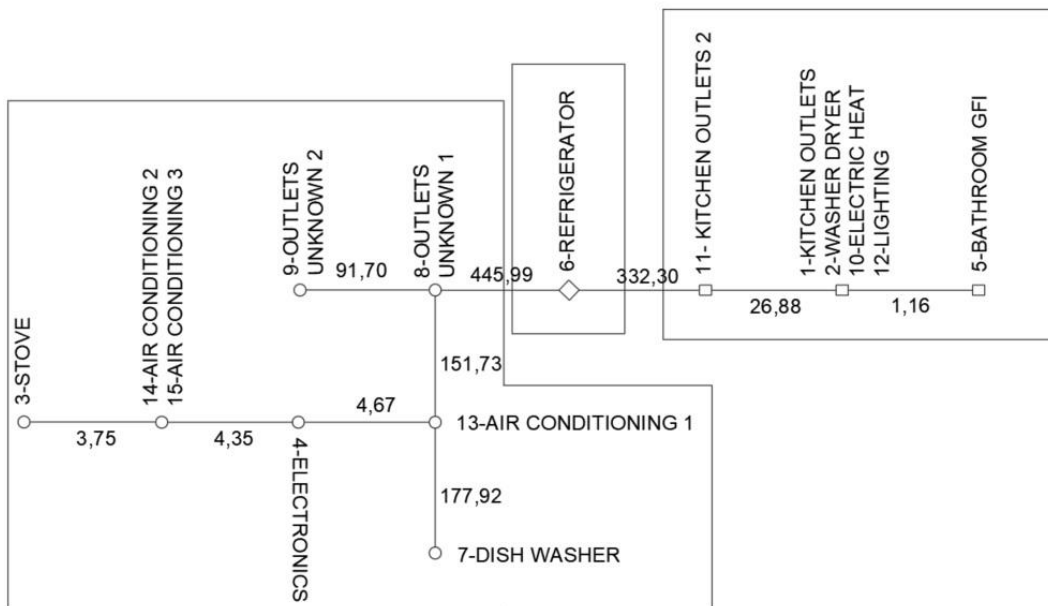


Figure 4.20: MST with k-means results represented over the MST for REDD, House 6, with k=3

5. SELF ORGANIZING MAPS RESULTS

The methodology described in Chapter 3 was also applied to the 2 datasets selected (UK-DALE and REDD). The simulations were made with SimpSOM Library in Python language, using Google Collaboratory to run the code.

The package proposes a hexagonal grid of neurons, closed. This means that the edges of the figures presented below are connected to each other, the right edge is connected to the left edge, and the upper and lower edges as well. To make the results comparable, for all the 5 houses in UK-DALE and the 6 houses in REDD were used to train a 40x40 grid for 50,000 epochs. The results shown below represent the distances between the neurons and its neighbors after the training in a color scale. The neuron agglomerations are represented by dark regions.

The results presented in this work represent one way of revealing the clusters existence in residential systems. The SOM methodology provides a wide range for exploration, by varying the grid configuration, such as:

- Shapes: can be linear, square, cylindrical, spherical, and other;
- the grid size is analogous to pixels in a figure: the most, the better, but the largest the grid, the most computer consuming is the training;
- and the number of training epochs: it is expected that, by increasing the training epochs quantity, the results visualization would became more clear (the colours in the pictures would be lighter or darker).

The exploration of these parameters variation and the impact in the results is described of one possibility of future work.

The code used to train the grid for UK-DALE House 5 is shown in Figure 5.1.

The Self Organizing Maps after training are shown in Figures 5.2 to 5.12. The distance between the neurons is shown in a color scale, from dark blue to yellow. The closer the neurons, the darker the regions, this is, the dark blue regions represent neuron agglomerations. On the other hand, yellow lines represent walls separating one agglomeration from another, or light green regions represent an area of low neuron density (transition between one cluster to another).

```

!pip install SimpSOM

import pandas as pd
import SimpSOM as sps
from sklearn.cluster import KMeans
import numpy as np

from google.colab import drive
drive.mount('/content/drive')

raw_data=np.genfromtxt('/content/drive/My Drive/DOUTORADO/HOUSE5_35DAYS_CLEAN.txt')
#labels=['red','green','blue','yellow','magenta','cyan','indigo']
#raw_data

net = sps.somNet(40, 40, raw_data, PBC=True)

net.train(0.01, 50000)
net.save('/content/drive/My Drive/DOUTORADO/SOM/UKDALEHOUSE5_weights')
#net.nodes_graph(colum=0)
net.diff_graph()

#net.nodes_graph(colnum=0) #plota o grafico
net.diff_graph()

net.project(raw_data, labels=labels)
net.cluster(raw_data, type='qthresh')

Y=net.find_bmu(np.transpose([[1,0,0]]))
Y.weights #mostra os pesos desse nó

Z=net.project(np.asarray([[0,1,1]]),show=True) #projeta no grid um ponto (ou conjunto de pont

#X=np.load('/content/colorExample_weights.npy') #arquivo com os pesos

```

Figure 5.1: Code used to train the 40x40 hexagonal grid to UK-DALE, House 5.

5.1 SOM RESULTS FOR UK-DALE

Figures 5.2 to 5.12 show the SOM after training with the distance between the neurons in a color scale. Some of the agglomerations were pointed with red circles to make it easier to the reader to follow the discussions.

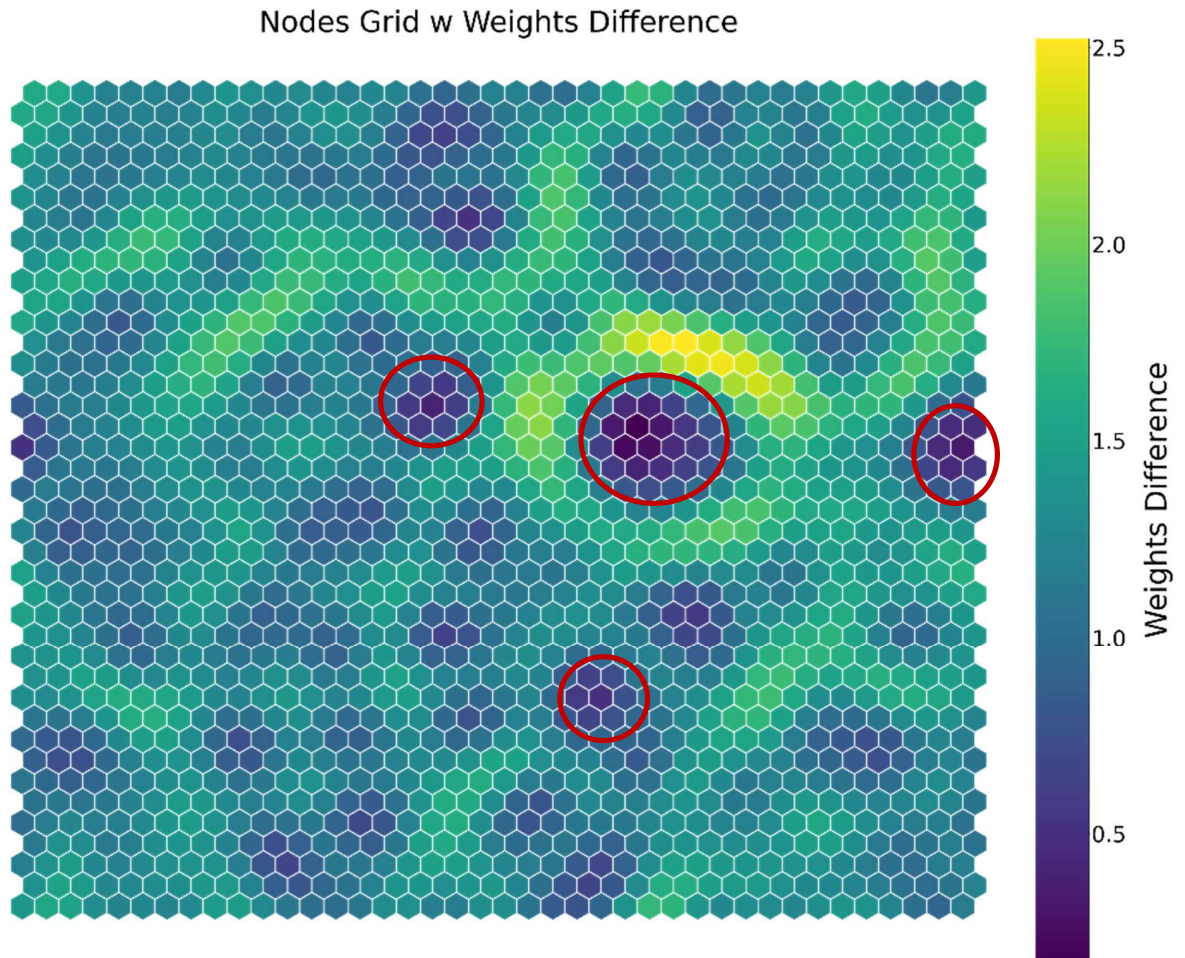


Figure 5.2: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 1.

The SOM maps results shown in this kind of figure are more qualitative than quantitative. For example, in Figure 5.2, the agglomerations are visible: two of them very clear (remembering that the map is closed, this is, the right border is connected to the left border, and the same happens to the upper and lower borders), and a couple more visible, but hard to decide if it is three clusters' or one. The result shows some similarity with the shown in Figure 4.10, with the PCA + k-means result. One cluster gathers most of the house appliances, and some other peripheral clusters. The distances between the clusters shown in Figure 4.10 also show that they are not very far from one another.

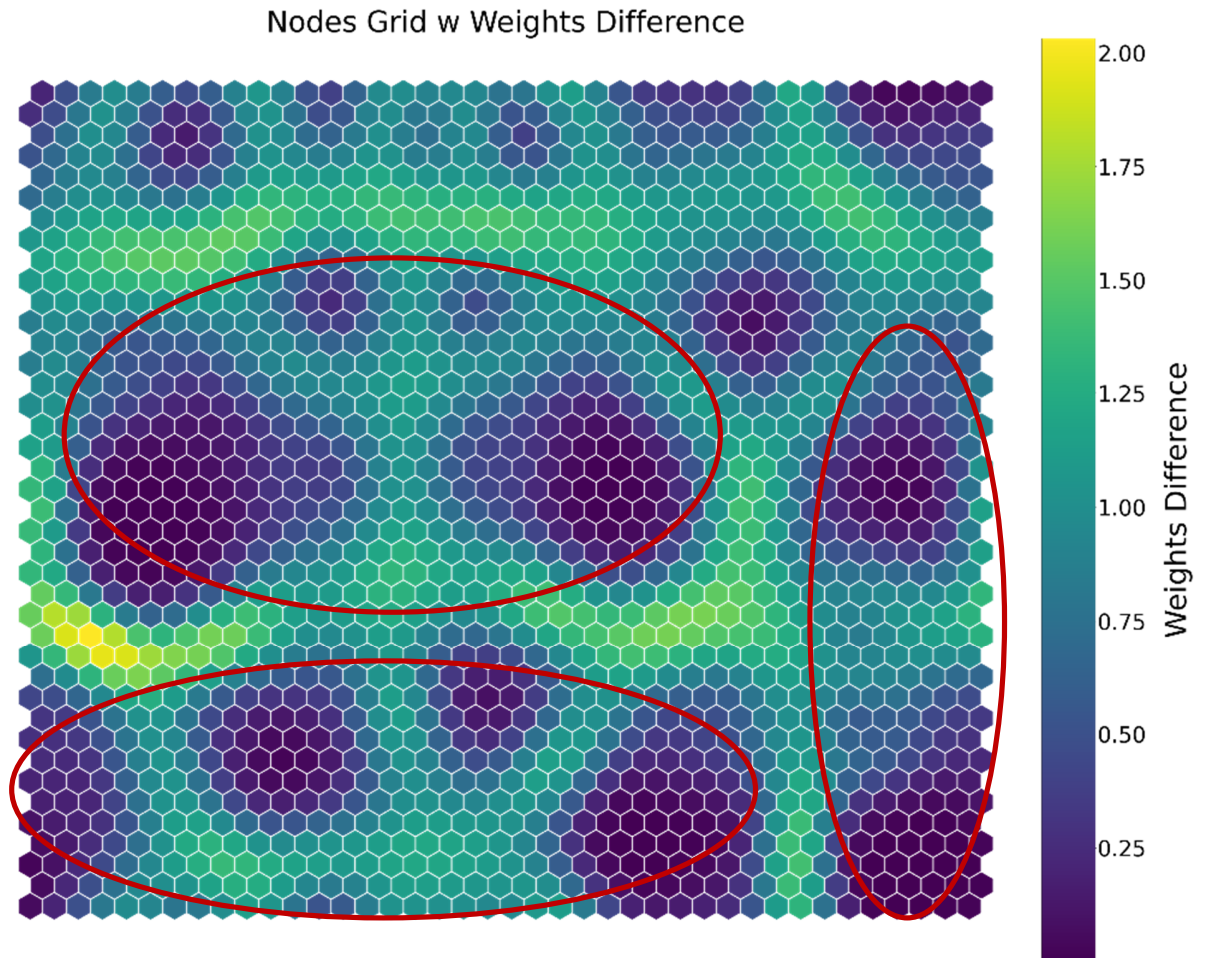


Figure 5.3: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 2.

When it comes to Figure 5.3, the clusters (neuron agglomerations) are more evident, but it is still hard to state if a light green region is part of a one cluster or a limit between them. The PCA and k-means result showed 3 clusters (see Figure 4.11), and if we focus on the yellow / light green lines in Figure 5.3 as divisions, it suggests 3 or 4 clusters.

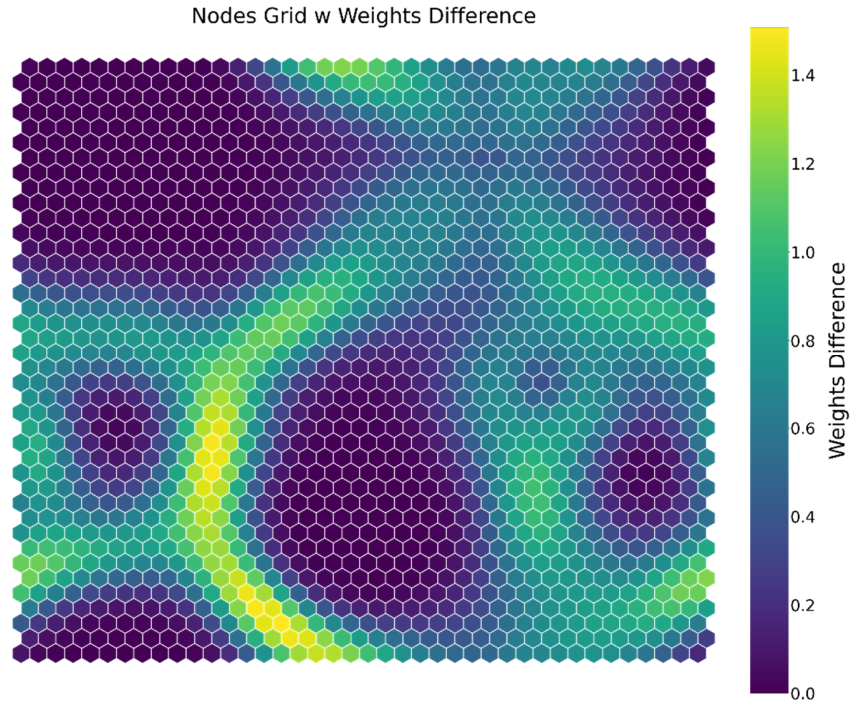


Figure 5.4: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 3.

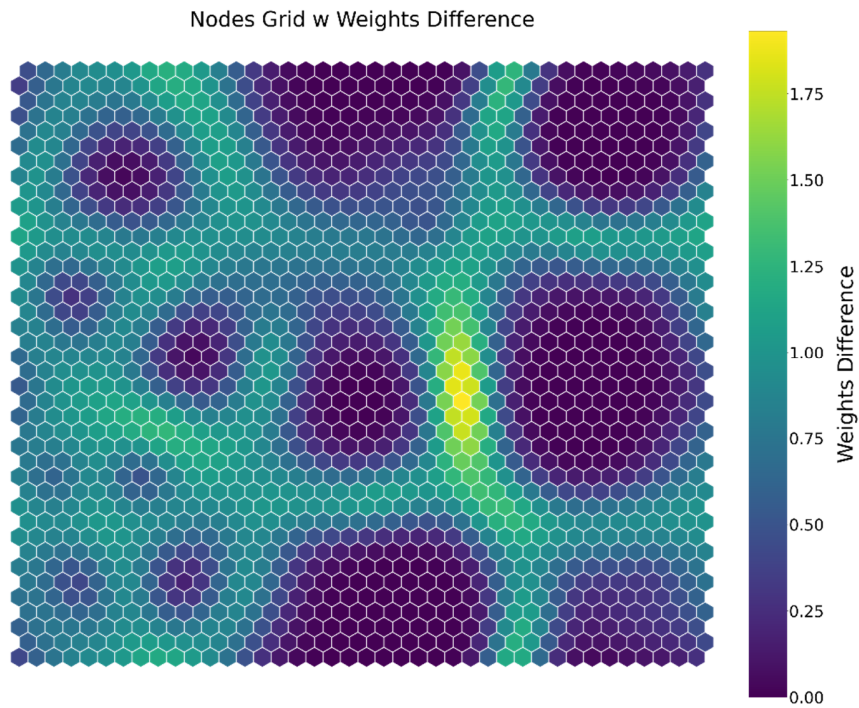


Figure 5.5: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 4.

Figures 5.4 and 5.5, as refer to Houses 3 and 5 from UK-DALE, with 4 and 5 individual channels respectively, make the results analysis possibilities too limited, and for this reason are not explored.

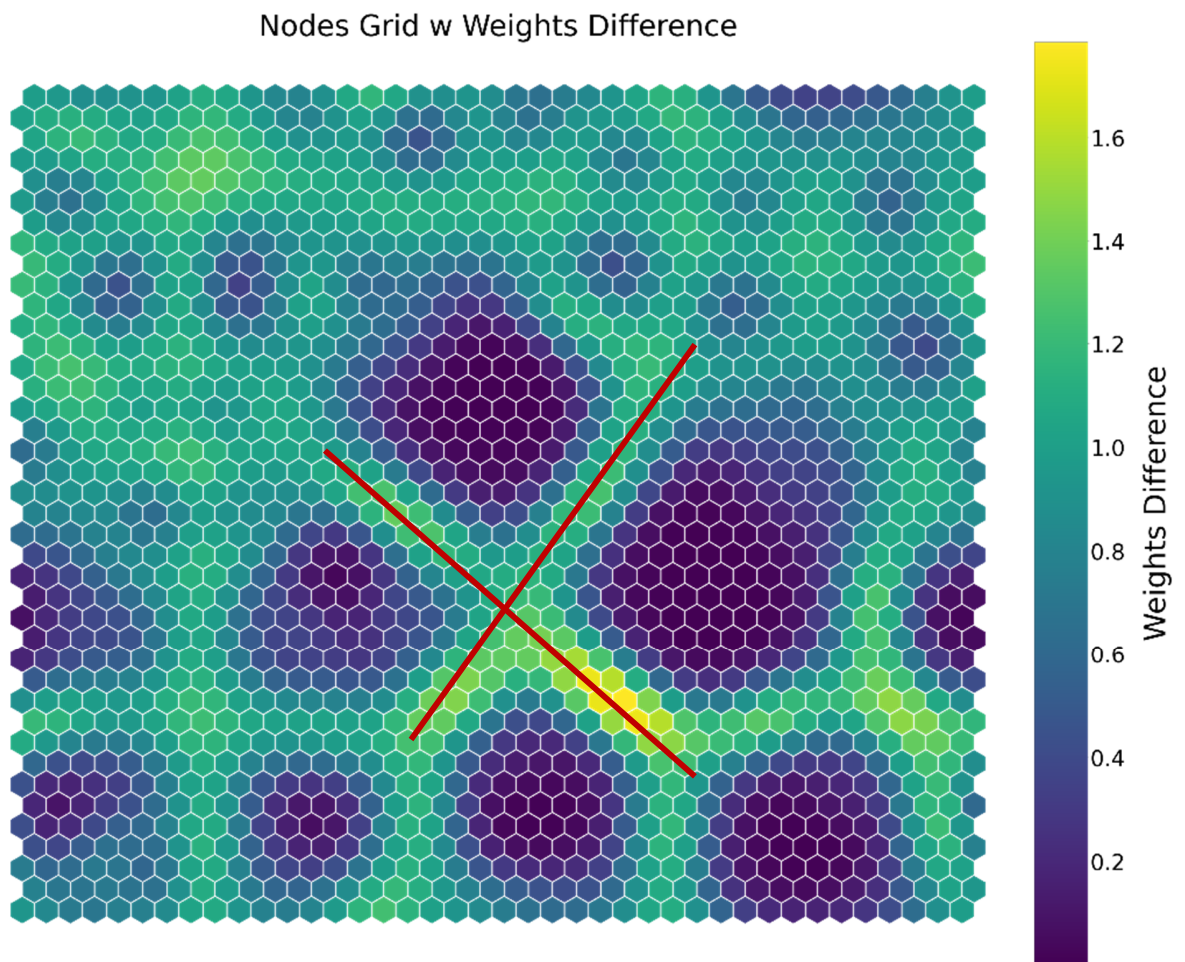


Figure 5.6: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for UK-DALE, House 5.

Figure 5.6 shows some agglomerations very clearly. The yellow / light green lines in the center of the figure suggest an X shape, separating the neurons in 4 main clusters.

5.2 REDD RESULTS

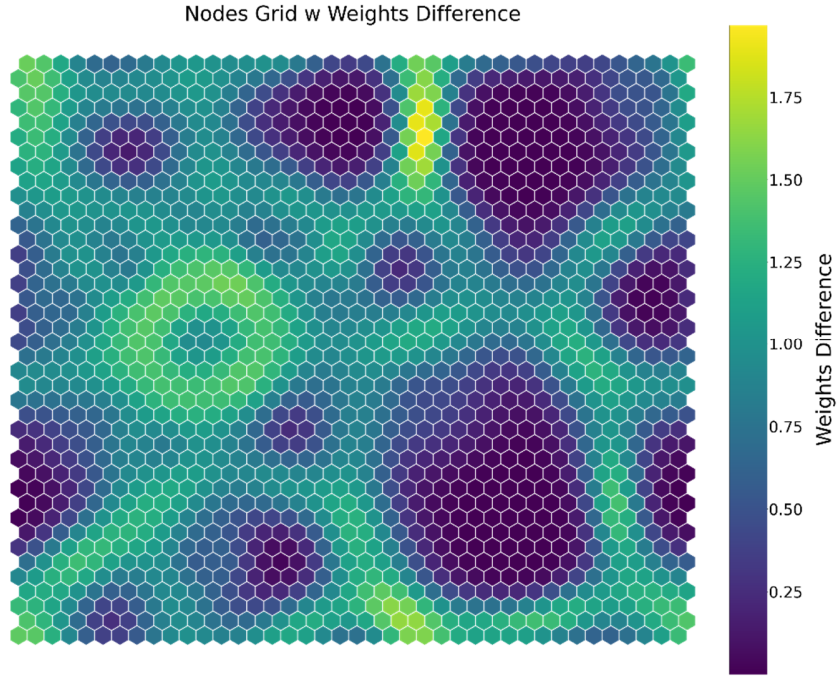


Figure 5.7: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 1.

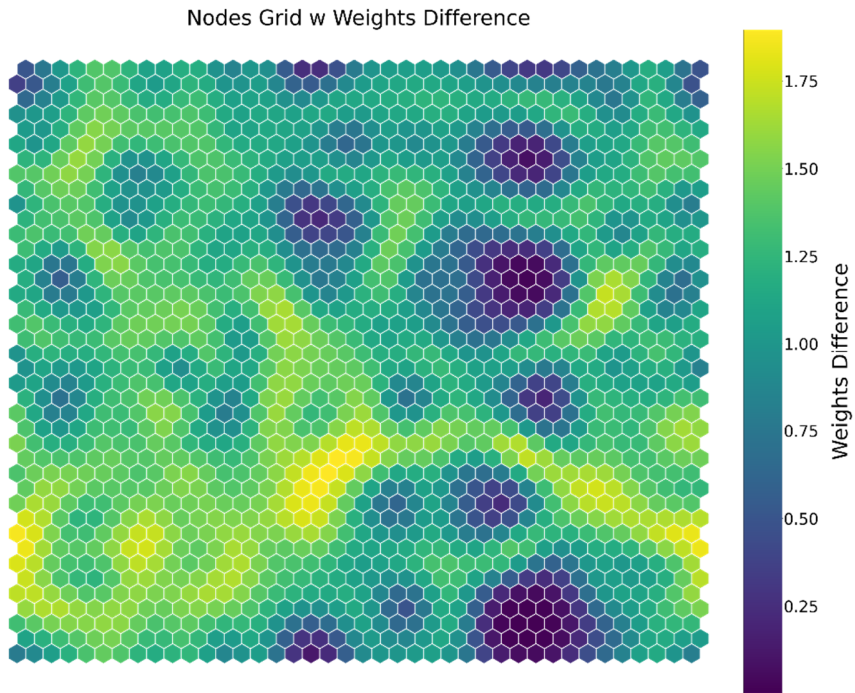


Figure 5.8: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 2.

The same line of analysis was followed when looking at the Figures 5.7 to 5.12, from the REDD dataset. Comparing Figure 5.7 to Figure 4.15, in the last it is shown 4 clusters with a large distance between them (the separating branch lengths are 346, 545 and 550). In a similar way, the light areas between the dark regions in Figure 5.7 seems to be more distinguished than in other Figures.

But the SOM results analysis must not be only a comparison between one linear method and one nonlinear. The loss of reference in the appliances resulting from the dimensionality reduction allows Figures 5.2 to 5.12 to show some other aspects from the system dynamic, like the classification of the system status as usual or unusual.

Every time a system status is presented to the map during the training, not only the winner neuron moves toward the sample, but the neighbors also. This means that if a common status is presented to the trained map, the winning neuron must be one from one dark region. The opposite is also true: if a non-usual status is presented to the trained grid, the winning neuron will be from a light region. This concept is the starting point to use the trained SOM as system status classifier.

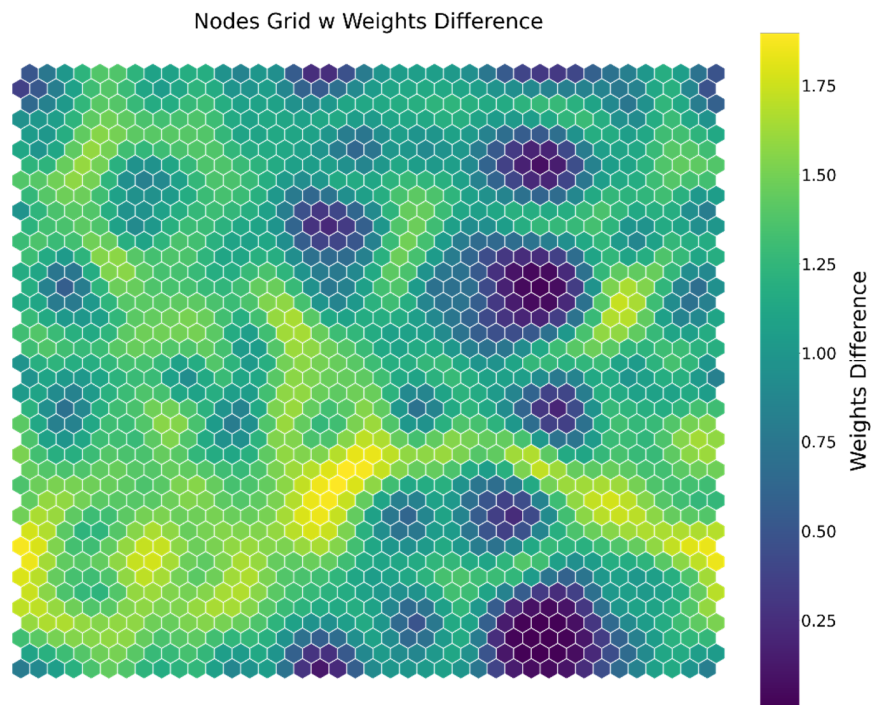


Figure 5.9: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 3.

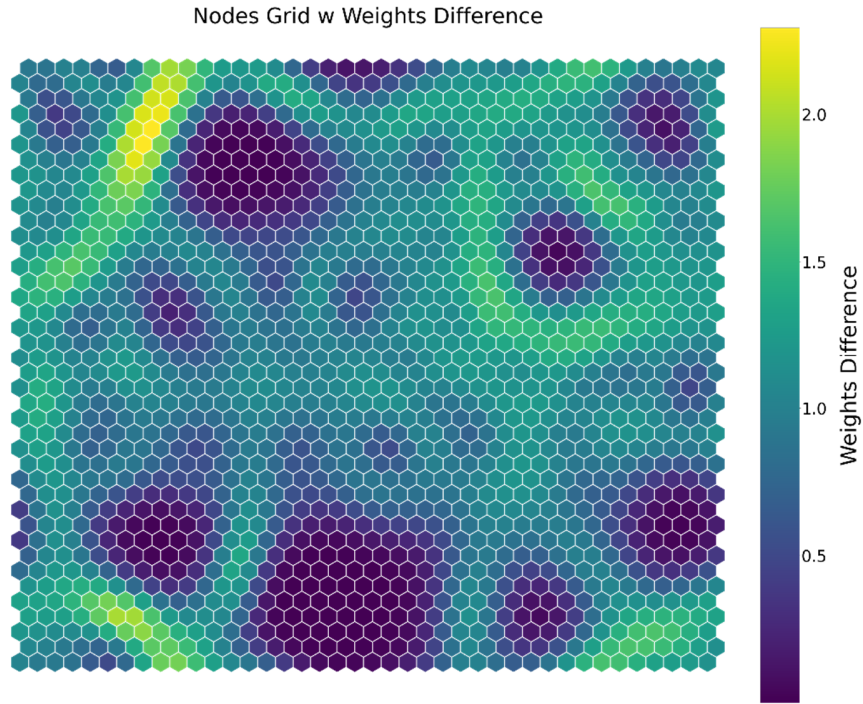


Figure 5.10: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 4.

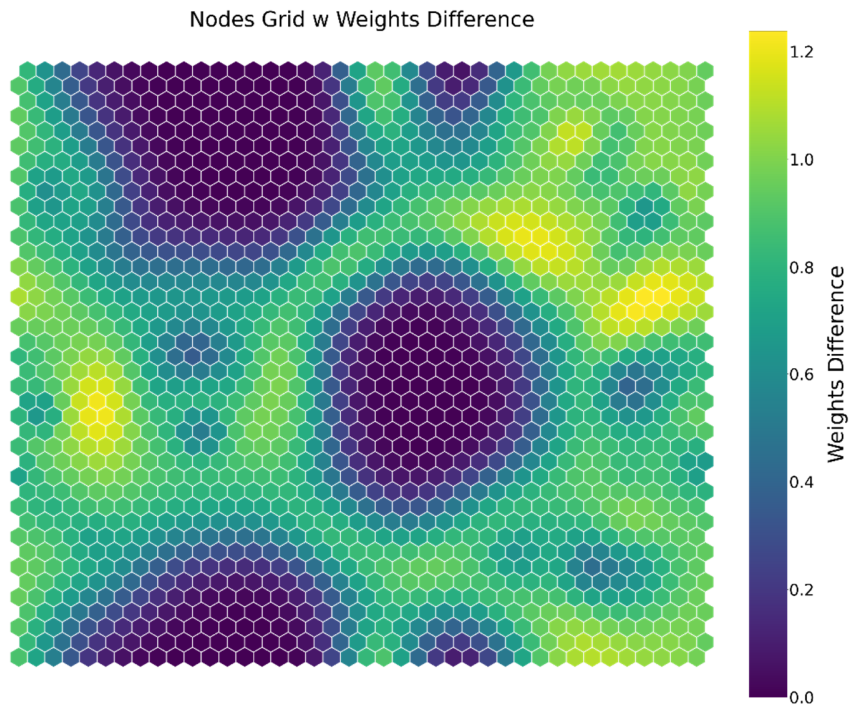


Figure 5.11: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 5.

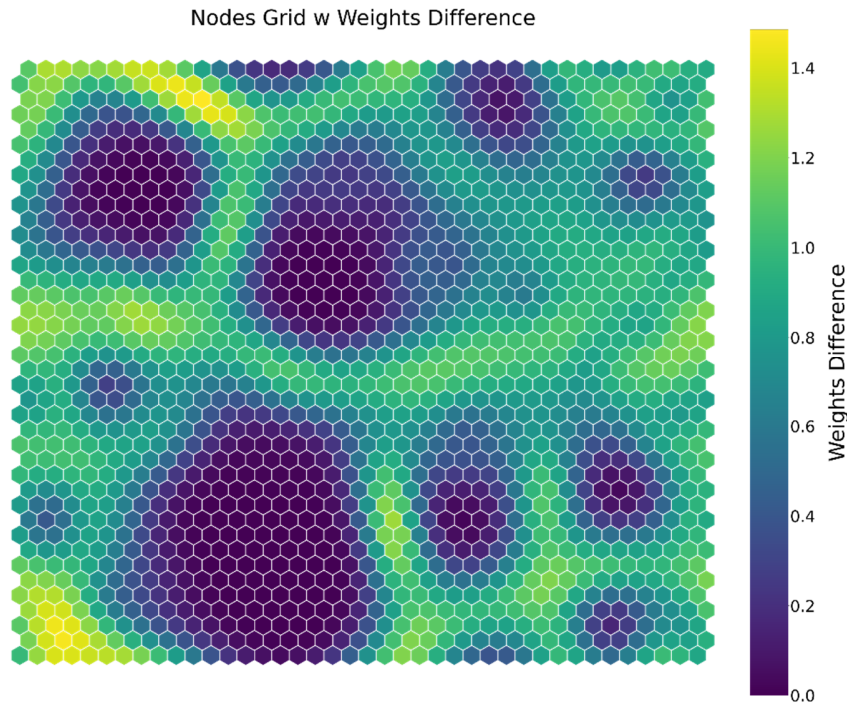


Figure 5.12: Flat view of trained 40x40 hexagonal SOM, after 50,000 epochs training, for REDD, House 6.

6. DISCUSSION

This work used two different methods of dimensionality reduction to find a natural association between appliances in residential installations. The algorithms were applied in data sets with measurements of actual residences with individual channels from 4 to 52, with monitoring periods from less than three days to more than one month. The data sets selected are the UK-DALE and the REDD datasets.

6.1 PCA AND K-MEANS FOR UK-DALE AND REDD DATASETS

In the case of the clusters shown in Figure 9, for the UK-DALE dataset, House 1, the larger group with all the appliances with a significant rated power (soldering iron, dishwasher, and washing machine) indicates a good potential for load modulation. The appliances represented as a single point also provide very interesting results, for example, the soldering iron + kettle, and can be used as a guide for detailed user feedback.

From the clustering results, it is also possible to infer the minimum number of people living in the house. Looking at the type of appliances in the same group, for example, it would be very unlikely that a single person would use several kitchen appliances and the soldering iron at the same time, and thus, there should be at least two people in the house at the same time. Following the same logic, observation of the lights suggests that there are people in the office and the living room during the busy periods.

In the case of the clusters in the UK-DALE dataset, House 2 (see Figure 11), some of the clusters are obvious, like “modem+server+router”, but the washing machine together with the PlayStation and the other kitchen appliances reveal some habits of the household population that can be exploited. Furthermore, 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 a video game.

In the case of the clusters for the UK-DALE dataset, House 5 (see Figure 13), the group with the office appliances (desktop+sky HD box+core2 server and others) makes a lot of sense, but the group with the kitchen appliances together with the hair dryer, steam iron, and washer dryer suggests that the house has a good load modulation potential. In this case, feedback from the energy supplier to the final user suggesting paying more attention to the use of these appliances together could be advisable. Regarding the number of occupants in the house, it seems that there are at least two people present during the busiest moment, one using the kitchen and the other playing a video game (PS4).

In the case of the clusters for the REDD dataset, House 1 (see Figure 14), the larger group includes almost all the appliances monitored. This means that everything is always used together, despite one lighting circuit (maybe this was always off during the monitoring period) and two circuits of kitchen outlets. House 1 in this data set is an interesting one, because it seems to have three washer dryers. The washer dryer and the dish washer are used together, also with the main appliances such as the stove, microwave, and oven.

In House 2, also for the REDD dataset, (see Figure 15), one interesting point about this result is that the refrigerator represents a single cluster, probably because of its very particular “on/off” cycles. Furthermore, this family probably either does not have the habit of cooking (the stove is insulated from the kitchen appliances, but the microwave is very close), or the cooker was broken for the entire monitoring period.

The linear approach (PCA and k-means) was very efficient in revealing the existence of appliance clusters. Because of the method's linearity, the reference in the appliances was not lost during the dimensionality reduction, and the final result is a number of groups for each house, containing appliances that are often used together. The final information can be of great use for the energy suppliers in order to suggest small changes in the consumers' behavior that can improve the energy efficiency of the residence.

Nevertheless, the main contribution of this method is not only the definition of appliance clusters and the range of analysis options that it brings, but also that the method does not require any information other than a large set of system statuses. If associated with an efficient disaggregation algorithm, the method can extract useful information about the occupants' behavior by using only the smart meter information.

6.2 SELF ORGANIZING MAPS FOR THE UK-DALE AND REDD DATASETS

The method of SOM was used as the nonlinear tool to perform the dimensionality reduction and to reveal patterns at the same time. Different from PCA and k-means, the SOM results depend on the map size, geometry, and other parameters involving the training itself. The same configuration was used for the nine houses that were analysed in order to obtain comparable results.

Agglomerations of neurons are visible in the grid from the distances between neurons and their neighbours. In our case, this distance is represented by a colour scale: dark blue means closer neurons while light yellow means further ones. In this way, the dark regions in Figures 20a to 21f indicate clusters. For example, in Figure 21e (REDD House 5), three main clusters are visually clear, but in Figure 21c, the map shows agglomerations that could be either interpreted as part of one big cluster, or of several small ones. Therefore, the actual number of clusters might not be visually defined in many cases.

Despite this fact, the results of SOM can be considered by following at least two different lines of interpretation. First, if the intention is to compare the linear and nonlinear methods for the same kind of results (finding appliance clusters that are statistically related), another step would be necessary to identify which appliances are included in each of the clusters revealed. This, however, would not add much to the conclusions, as the results from PCA are already reliable.

The second line, which is also the main contribution of SOM to this work, follows the fact that the neurons in the dark regions represent system statuses that are very usual, and those in the light regions represent statuses that are not usual. Thus, the trained SOM can be used as a classifier of “usual” or “not usual” system status. If the training is carried out with a sufficiently large set of system statuses that do not contain any malfunctions, the classification can be extended to the “healthy” or “fault or malfunction” statuses.

One very important difference between the SOM and PCA methods is that with PCA, as the method is nonparametric, the interpretation of the results is straightforward. After PCA reduces the data dimensionality, k-means and the elbow method show the best cluster configuration. On the other hand, this method requires some user inputs, mainly to decide how many PCs will be considered, and the best number of clusters. These can be interpreted as parameters for the whole formulation.

With SOM, first, the grid parameters must be chosen carefully. If the grid is defined to be too small, or there are not enough training epochs, the clusters may not become visible. However, once the parameters are chosen, the result is obtained in only one step: the grid training. Interpretation of the results is not as straightforward as in PCA and k-means, but this gives the reader more liberty to discuss the results in a nonbinary way. For example, in Figures 20a to 21f, the agglomerations are visible, but the number of clusters can be hard to determine exactly.

7. CONCLUSION AND FUTURE WORK

7.1 CONCLUSION

In this paper, we have proposed a technique for the detection of appliance utilization patterns. These patterns are identified from only the system status behavior, which is represented by large system status datasets, by using dimensionality reduction and clustering algorithms. PCA, k-means, and the elbow method are used to define the clusters, and the minimum spanning tree is used to visualize the results and show the appearance of utilization patterns. The SOM technique is used to create a system status classifier. Thus, the proposed methodology uses low-computational cost algorithms that do not require any information about households.

To demonstrate the effectiveness of the proposed techniques, we applied them to two public datasets from two different countries with different usage patterns, the United Kingdom (UK-DALE) and the US (REDD). The techniques were very effective in revealing usage patterns of appliances with no need for any personal information of the households.

Using the proposed clustering techniques, system operators can implement effective demand-side management. Further, the system status classifier can be used to detect appliance malfunctions through system status analyses alone. In the future, we will improve the methodology by incorporating a good performance disaggregation algorithm so that the system status set could be obtained from only the smart meter information. Further, we will improve the classifier formulation, giving each neuron a label, and using the classifier as a real-time monitor. We will try to develop and analyze different classifiers that function seasonally (winter, spring, summer, or autumn), or produce different maps for weekdays and weekends.

7.2 FUTURE WORK

As future work, it can be suggested:

- Apply the methodology described to other public datasets, not only the ones already published before year 2021, but also to other ones that are being created at this moment. The new datasets, as result of new load monitoring programs, shall have better information quantity and quality compared to the ones already available. This way, it is expected that soon it will be possible to apply the methodology for data collected for longer periods, and that can be subdivided according week days and week-ends, season of the year, and other information that can lead to a deeper understanding of load usage in residences.
- Apply other clustering methodologies to the data already used and other datasets and compare the results. The clusters definition is the most valuable information to create good feedback letters to the final energy users, and thus is the key to transform this work into a practical application.
- Explore the parameters variation of Self Organizing Maps in the maps results: grid geometry, grid size and training epochs quantity.
- Continue developing the idea of using the trained SOM to work as a system status classifier. For that, the SOM training must be longer (more epochs) and with a larger set of information. The results will be very interesting and possible of large discussions. As the methodology is parametric, it is also possible to train maps with different sizes (number of neurons) and geometries to compare the results.
- Develop user feedback letters containing specific information about the residence usage profile and study the impact of this information in the households change of habits.
- Associate the patterns recognition methodology with a high-performance disaggregation algorithm. This step is also very important to transform this work into a practical application.

REFERENCES

- [1] Roy Dong, Lillian Ratliff, Henrik Ohlsson, and S. Shankar Sastry , A dynamical systems approach to energy disaggregation, *IEEE 52nd Annual Conference on Decision and Control (CDC)*, 2013, pp. 6335-6340.
- [2] Hart, G.W. 1992. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*. 80, 12 (1992), 1870–1891.
- [3] Zeifman, M. and Roth, K. 2011. Nonintrusive appliance load monitoring: Review and outlook. *IEEE Transactions on Consumer Electronics*. (2011), 76–84.
- [4] Nipun Batra, Oliver Parson, Mario Berges, Amaejeet Singh and Alex Rogers, A comparison of non-intrusive load monitoring methods for commercial and residential buildings
- [5] Hsueh-Hsien Chang, Ching-Lung Lin and Jin-Kwei Lee, Load identification in Non-intrusive Load Monitoring using steady-state and turn-on transient energy algorithms, *Proceedings of the 2010 14th International Conference on Computer Supported Cooperative Work in Design*
- [6] Leslie K. Norford, Steven B. Leeb, Non Intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms, *Energy and Buildings*, 24 (1996), 51-64
- [7] Y. Agarwal, T. Weng, and R. K. Gupta. The energy dashboard: improving the visibility of energy consumption at a campus-wide scale. In *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 55-60. ACM, 2009.
- [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.
- [9] *Proceedings of 2nd KDD Workshop on Data Mining Applications in Sustainability*, pages 12-16, Beijing, China, 2012.
- [10] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert. Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*, 52:213-234, 2013.

- [11] N. Batra, P. Arjunan, A. Singh, and P. Singh. Experiences with occupancy based building management systems. In *Intelligent Sensors, Sensor Networks and Information Processing*, 2013. IEEE Eighth International Conference on, pages 153-158. IEEE, 2013.
- [12] N. Batra, H. Dutta, and A. Singh. INDiC: Improved Non-Intrusive load monitoring using load Division and Calibration. In *International Conference of Machine Learning and Applications*, Miami, FL, USA, 2013.
- [13] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava. NILMTK: An Open Source Toolkit for Non-intrusive Load Monitoring. In *Fifth International Conference on Future Energy Systems (ACM e-Energy)*, Cambridge, UK, 2014.
- [14] C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, and S. Santini. The eco data set and the performance of non-intrusive load monitoring algorithms. In *Proceedings of the First ACM International Conference on Embedded Systems For Energy-Efficient Buildings*. ACM, 2014.
- [15] R. Fontugne, J. Ortiz, N. Tremblay, P. Borgnat, P. Flandrin, K. Fukuda, D. Culler, and H. Esaki. Strip, bind, and search: a method for identifying abnormal energy consumption in buildings. In *Proceedings of the 12th international conference on Information processing in sensor networks*, pages 129-140. ACM, 2013.
- [16] M. Gulati, S. Sundar Ram, and A. Singh. An in depth study into using emi signatures for appliance identification. In *Proceedings of the First ACM International Conference on Embedded Systems For Energy-Efficient Buildings*. ACM, 2014.
- [17] J. Kelly and W. Knottenbelt. UK-DALE: A dataset recording UK Domestic Appliance-Level Electricity demand and whole-house demand. *ArXiv e-prints*, 2014.
- [18] H. Kim, M. Marwah, M. F. Arlitt, G. Lyon, and J. Han. Unsupervised Disaggregation of Low Frequency Power Measurements. In *Proceedings of 11th SIAM International Conference on Data Mining*, pages 747-758, Mesa, AZ, USA, 2011.
- [19] J. Z. Kolter, S. Batra, and A. Y. Ng. Energy Disaggregation via Discriminative Sparse Coding. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pages 1153-1161, Vancouver, BC, Canada, 2010.

- [20] K. D. Lee, S. B. Leeb, L. K. Norford, P. R. Armstrong, J. Holloway, and S. R. Shaw. Estimation of variable-speed-drive power consumption from harmonic content. *Energy Conversion, IEEE Transactions on*, 20(3):566-574, 2005.
- [21] J. Liang, S. Ng, G. Kendall, and J. Cheng. Load Signature Study - Part I: Basic Concept, Structure, and Methodology. *IEEE Transactions on Power Delivery*, 25(2):551-560, 2010.
- [22] S. Makonin, F. Popowich, L. Bartram, B. Gill, and I. V. Bajic. AMPds: A Public Dataset for Load Disaggregation and Eco-Feedback Research. In *IEEE Electrical Power and Energy Conference*, Halifax, NS, Canada, 2013.
- [23] A. Monacchi, D. Egarter, W. Elmenreich, S. D'Alessandro, and A. M. Tonello. Greend: An energy consumption dataset of households in italy and austria. arXiv preprint arXiv:1405.3100, 2014.
- [24] L. Norford, H. Xing, and D. Luo. Detection of HVAC Equipment Turn On-Turn Off Events with Non-Intrusive Electrical Load Monitoring. Technical report, Massachusetts Institute of Technology, MA, USA, 2001.
- [25] L. K. Norford and S. B. Leeb. Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms. *Energy and Buildings*, 24(1):51-64, 1996.
- [26] J. Schroeder. The energy consumption of elevators. *Elevator Technology*, (Editor Dr. G. Barney), Ellis Horwood, 1986.
- [27] H. Shao, M. Marwah, and N. Ramakrishnan. A temporal motif mining approach to unsupervised energy disaggregation: Applications to residential and commercial buildings. *Power (W)*, 250(1700):2250, 1700.
- [28] S. Thakur, M. Saha, A. Singh, and Y. Agarwal. WattShare: Detailed energy apportionment in shared living spaces within commercial buildings. In *Proceedings of the First ACM International Conference on Embedded Systems For Energy-Efficient Buildings*. ACM, 2014.
- [29] L. V. Thanayankizil, S. K. Ghai, D. Chakraborty, and D. P. Seetharam. Softgreen: Towards energy management of green office buildings with soft sensors.
- [30] Henrik Ohlsson, Lillian Ratliff, Roy Dong and S. Shankar Sastry: Blind Identification of ARX Models with Piecewise Constant Inputs.

- [31] K. Abed-Meraim, J.-F. Cardoso, A.Y. Gorokhov, P. Loubaton, and E. Moulines. On subspace methods for blind identification of singleinput multiple-output FIR systems. *IEEE Transactions on Signal Processing*, 45(1):42–55, Jan.
- [32] K. Abed-Meraim, W. Qiu, and Y. Hua. Blind system identification. *Proceedings of the IEEE*, 85(8):1310–1322, 1997.
- [33] R. Dong, L. Ratliff, H. Ohlsson, and S. S. Shankar. A dynamical systems approach to energy disaggregation. In *Proceedings of the 52th IEEE Conference on Decision and Control*, Florence, Italy, December 2013. Submitted to.
- [34] D. Gesbert, P. Duhamel, and S. Mayrargue. On-line blind multichannel equalization based on mutually referenced filters. *IEEE Transactions on Signal Processing*, 45(9):2307–2317, 1997.
- [35] Y. Hua. Fast maximum likelihood for blind identification of multiple FIR channels. *IEEE Transactions on Signal Processing*, 44(3):661–672, 1996.
- [36] J. I. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.
- [37] S. Talwar, M. Viberg, and A. Paulraj. Blind estimation of multiple cochannel digital signals using an antenna array. *IEEE Signal Processing Letters*, 1(2):29–31, 1994.
- [38] L. Jung Lennart. *System Identification – Theory for the User*, second edition.
- [39] California Energy Comission. CEC End-User Survey, CEC-400-2006-005, March 2006. <http://buildingsdatabook.eren.doe.gov/>
- [40] DOE. *Buildings Energy Data Book*, Department of Energy, March 2009. <http://buildingsbatabook.eren.doe.gov/>.
- [41] U.S. Energy Information Administration, “End-Use Consumption of Electricity 2001” <http://www.eia.gov/emeu/recs/recs2001/enduse2001/enduse2001.html>
- [42] A. Yang, S. Li, C. Ren, H. Liu, Y. Han and L. Liu, "Situational Awareness System in the Smart Campus," in *IEEE Access*, vol. 6, pp. 63976-63986, 2018, doi: 10.1109/ACCESS.2018.2877428
- [43] Z. Yorio, R. Oram, S. El-Tawab, A. Salman, M. H. Heydari and B. B. Park, "Data analysis and information security of an Internet of Things (IoT) intelligent transit system," 2018 *Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, 2018, pp. 24-29, doi: 10.1109/SIEDS.2018.8374744.

- [44] W. Muhamad, N. B. Kurniawan, Suhardi and S. Yazid, "Smart campus features, technologies, and applications: A systematic literature review," 2017 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, Indonesia, 2017, pp. 384-391, doi: 10.1109/ICITSI.2017.8267975.
- [45] Ian Ayres, Sophie Raseman, Alice Shih, Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage, *The Journal of Law, Economics, and Organization*, Volume 29, Issue 5, October 2013, Pages 992–1022, <https://doi.org/10.1093/jleo/ews020>
- [46] Noah J. Goldstein, Robert B. Cialdini, Vladas Griskevicius, A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels, *Journal of Consumer Research*, Volume 35, Issue 3, October 2008, Pages 472–482, <https://doi.org/10.1086/586910> Satu Pätäri, Kirsi Sinkkonen,
- [47] Energy Service Companies and Energy Performance Contracting: is there a need to renew the business model? Insights from a Delphi study, *Journal of Cleaner Production*, Volume 66, 2014, Pages 264-271, ISSN 0959-6526, <https://doi.org/10.1016/j.jclepro.2013.10.017>. (<https://www.sciencedirect.com/science/article/pii/S095965261300680X>).
- [48] Kolter, J.Z.; Johnson, M.J. REDD: A public data set for energy disaggregation research. Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, 2011, Vol. 25, pp. 59–62.
- [49] Anderson, K.; Ocleanu, A.; Benitez, D.; Carlson, D.; Rowe, A.; Berges, M. BLUED: A fully labeled public dataset for event-based non-intrusive load monitoring research. Proceedings of the 2nd KDD workshop on data mining applications in sustainability (SustKDD), 2012, pp. 1–5.
- [50] García, S.; Parejo, A.; Personal, E.; Guerrero, J.I.; Biscarri, F.; León, C. A retrospective analysis of the impact of the COVID-19 restrictions on energy consumption at a disaggregated level. *Applied Energy* 2021, p. 116547.
- [51] Vega, A.; Amaya, D.; Santamaría, F.; Rivas, E. Active demand-side management strategies focused on the residential sector. *The Electricity Journal* 2020, 33, 106734.

- [52] Hussain, H.M.; Nardelli, P.H. A Heuristic-based Home Energy Management System for Demand Response. 2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS). IEEE, 2020, Vol. 1, pp. 285–290.
- [53] Hui, H.; Ding, Y.; Shi, Q.; Li, F.; Song, Y.; Yan, J. 5G network-based Internet of Things for demand response in smart grid: A survey on application potential. *Applied Energy* 2020, 257, 113972.
- [54] Wei, Y.; Xia, L.; Pan, S.; Wu, J.; Zhang, X.; Han, M.; Zhang, W.; Xie, J.; Li, Q. Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks. *Applied energy* 2019, 240, 276–294.
- [55] Zhao, H.; Yan, X.; Ren, H. Quantifying flexibility of residential electric vehicle charging loads using non-intrusive load extracting algorithm in demand response. *Sustainable Cities and Society* 2019, 50, 101664.
- [56] Ushakova, A.; Mikhaylov, S.J. Big data to the rescue? Challenges in analysing granular household electricity consumption in the United Kingdom. *Energy Research & Social Science* 2020, 64, 101428.
- [57] Narayanan, A.; De Sena, A.S.; Gutierrez-Rojas, D.; Melgarejo, D.C.; Hussain, H.M.; Ullah, M.; Bayhan, S.; Nardelli, P.H. Key Advances in Pervasive Edge Computing for Industrial Internet of Things in 5G and Beyond. *IEEE Access* 2020, 8, 206734–206754.
- [58] Chen, Y.C.; Chu, C.M.; Tsao, S.L.; Tsai, T.C.-Detecting users' behaviors based on nonintrusive load monitoring technologies. 2013 10th IEEE International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2013, pp. 804–809.
- [59] Beckel, C.; Sadamori, L.; Staake, T.; Santini, S. Revealing household characteristics from smart meter data. *Energy* 2014, 78, 397–410.
- [60] Aleksei Mashlakov, Evangelos Pournaras, Pedro H.J. Nardelli, Samuli Honkapuro, Decentralized cooperative scheduling of prosumer flexibility under forecast uncertainties, *Applied Energy*, Volume 290, 2021, 116706, ISSN 0306-2619, <https://doi.org/10.1016/j.apenergy.2021.116706>.
- [61] Parag, Y., Sovacool, B. Electricity market design for the prosumer era. *Nat Energy* 1, 16032 (2016). <https://doi.org/10.1038/nenergy.2016.32>.

- [62] COWI, 2016. Impact Assessment Study on Downstream Flexibility, Price Flexibility, Demand Response and Smart Metering. EUROPEAN COMMISSION DG ENERGY, Brussels.
- [63] EA Technology, Southern Electric Power Distribution, 2016. My Electric Avenue (I2EV) Project Close-Down Report. EA Technology & Southern Electric Power Distribution.
- [64] EcoGrid EU, 2016. EcoGrid EU – A Prototype for European Smart Grids Deliverable D6.7 Overall Evaluation and Conclusion. EcoGrid EU.
- [65] EPRI, 2011. The Effect on Electricity Consumption of the Commonwealth Edison Customer Applications Program: Phase 2 Final Analysis. Electric Power Research Institute, Palo Alto, CA.
- [66] Bryony Parrish, Phil Heptonstall, Rob Gross, Benjamin K. Sovacool, A systematic review of motivations, enablers and barriers for consumer engagement with residential demand response, *Energy Policy*, Volume 138, 2020, 111221, ISSN 0301-4215, <https://doi.org/10.1016/j.enpol.2019.111221>.
- [67] Gils, H.C., 2014. Assessment of the theoretical demand response potential in Europe. *Energy* 67, 1–18. <https://doi.org/10.1016/j.energy.2014.02.019>.
- [68] EPRI, 2012. Understanding Electric Utility Customers - Summary Report what We Know and what We Need to Know. Electric Power Research Institute, Palo Alto, CA. Available at. <http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId%4000000000001025856>.
- [69] H. M. Hussain and P. H. J. Nardelli, A Heuristic-based Home Energy Management System for Demand Response, 2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS), Tampere, Finland, 2020, pp. 285-290, doi: 10.1109/ICPS48405.2020.9274742.
- [70] Sahbasadat Rajamand, Effect of demand response program of loads in cost optimization of microgrid considering uncertain parameters in PV/WT, market price and load demand, *Energy*, Volume 194, 2020, 116917, ISSN 0360-5442, <https://doi.org/10.1016/j.energy.2020.116917>. (<https://www.sciencedirect.com/science/article/pii/S0360544220300244>).

- [71] Hongxun Hui, Yi Ding, Qingxin Shi, Fangxing Li, Yonghua Song, Jinyue Yan, 5G network-based Internet of Things for demand response in smart grid: A survey on application potential, *Applied Energy*, Volume 257, 2020, 113972, ISSN 0306-2619, <https://doi.org/10.1016/j.apenergy.2019.113972>. (<https://www.sciencedirect.com/science/article/pii/S0306261919316599>).
- [72] US Department of Energy . An assessment of energy technologies and research opportunities. USA, Washington D.C: US Department of Energy; 2015.
- [73] Sayed Saeed Hosseini, Kodjo Agbossou, Sousso Kelouwani, Alben Cardenas, Non-intrusive load monitoring through home energy management systems: A comprehensive review, *Renewable and Sustainable Energy Reviews*, Volume 79, 2017, Pages 1266-1274, ISSN 1364-0321, <https://doi.org/10.1016/j.rser.2017.05.096>. (<https://www.sciencedirect.com/science/article/pii/S1364032117307359>).
- [74] American Energy Society . Energy future: think efficiency. California, USA: American Energy Society; 2008.
- [75] Hafiz Khurram Iqbal, Farhan Hassan Malik, Aoun Muhammad, Muhammad Ali Qureshi, Muhammad Nawaz Abbasi, Abdul Rehman Chishti, A critical review of state-of-the-art non-intrusive load monitoring datasets, *Electric Power Systems Research*, Volume 192, 2021, 106921, ISSN 0378-7796, <https://doi.org/10.1016/j.epsr.2020.106921>. (<https://www.sciencedirect.com/science/article/pii/S0378779620307197>).
- [76] Lorna A. Greening, David L. Greene, Carmen Difiglio, Energy efficiency and consumption — the rebound effect — a survey, *Energy Policy*, Volume 28, Issues 6–7, 2000, Pages 389-401, ISSN 0301-4215, [https://doi.org/10.1016/S0301-4215\(00\)00021-5](https://doi.org/10.1016/S0301-4215(00)00021-5). (<https://www.sciencedirect.com/science/article/pii/S0301421500000215>).
- [77] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, J. Albrecht, et al., Smart*: an open data set and tools for enabling research in sustainable homes, *SustKDD* 111 (112) (August 2012) 108.
- [78] N. Batra, O. Parson, M. Berges, A. Singh, A. Rogers, A comparison of non-intrusive load monitoring methods for commercial and residential buildings, arXiv preprint arXiv:1408.6595(2014).

- [79] VASSILEVA, Iana; CAMPILLO, Javier. Increasing energy efficiency in low-income households through targeting awareness and behavioral change. *Renewable energy*, v. 67, p. 59-63, 2014.
- [80] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. Smart*: An Open Data Set and Tools for Enabling Research in Sustainable Homes. Proceedings of the 2012 Workshop on Data Mining Applications in Sustainability (SustKDD 2012), Beijing, China, August 2012.
- [81] Nipun Batra, Oliver Parson, Mario Berges, Amarjeet Singh and Alex Rogers. A comparison of non-intrusive load monitoring methods for commercial and residential buildings. 2014 arXiv:1408.6595.
- [82] Jonathon Shlens, A. Tutorial on Principal Component Analysis <http://www.brain-mapping.org/NITP/PNA>. *Readings 2005*, 12,10.
- [83] Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 1988.

8. APENDIX A – DATASETS

Table 8.1: Number of samples in each individual channel, and the correspondent label for House 1, UK-DALE dataset. In this house, channels 22, 39, 40 and 41 were eliminated for having too few samples compared to the other channels.

INDIVIDUAL CHANNEL LABEL – HOUSE 1	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 1	SAMPLES
CH1 - main channel	21,837,636	CH28 - Subwoofer living room	18,645,123
CH2 - boiler	21,281,331	CH29 - Living room lamp tv	18,651,142
CH3 - Solar thermal pump	21,281,208	CH30 – DAB radio living room	677,711
CH4 - laptop	4,539,118	CH31 - Kitchen lamp 2	11,355,561
CH5 - Washing machine	19,555,935	CH32 - Kitchen phones & stereo	18,497,094
CH6 - dishwasher	19,819,392	CH33 - Utilityrm lamp	17,439,431
CH7 - TV	19,763,329	CH34 - Samsung charger	17,003,826
CH8 - Kitchen lights	21,449,386	CH35 - Bedroom d lamp	17,934,605
CH9 - HTPC	19,543,268	CH36 - Coffee machine	17,304,547
CH10 - kettle	18,881,051	CH37 - Kitchen radio	18,506,920
CH11 - toaster	19,404,267	CH38 - Bedroom chargers	12,207,918
CH12 - fridge	19,381,298	CH39 - Hair dryer	181,497
CH13 - microwave	19,406,625	CH40 - Straighteners	45,239
CH14 - Lcd office	4,143,648	CH41 - Iron	7,160
CH15 - Hifi office	4,157,704	CH42 - Gas oven	18,185,875
CH16 - breadmaker	13,351,860	CH43 - Data logger pc	18,621,408
CH17 - Amp livingroom	19,547,699	CH44 - Childs table lamp	18,327,970
CH18 - Adsl router	19,116,542	CH45 - Childs ds lamp	18,004,938
CH19 - Livingroom s lamp	18,609,054	CH46 - Baby monitor tx	17,820,706
CH20 - Soldering iron	617,699	CH47 - Battery charger	13,333,663
CH21 - gigE_&_USBhub	3,602,216	CH48 - Office lamp 1	18,179,281
CH22 - Hoover	134,467	CH49 - Office lamp 2	17,649,667
CH23 - Kitchen dt lamp	10,670,426	CH50 - Office lamp 3	3,323,283
CH24 - Bedroom ds lamp	16,218,474	CH51 - Office PC	3,391,394
CH25 - Lightning circuit	20,785,844	CH52 - Office fan	2,632,277
CH26 - Livingroom s lamp 2	18,618,217	CH53 - LED printer	2,977,937
CH27 - iPad charger	12,253,298		

Table 8.2: Number of samples in each individual channel, and the correspondent label for House 2, UK-DALE dataset. In this house, it is not necessary to eliminate any channel.

INDIVIDUAL CHANNEL LABEL – HOUSE 2	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 2	SAMPLES
CH1 - main channel	2,780,373	CH11 - laptop	2,804,685
CH2 - monitor	2,805,646	CH12 - washing machine	1,686,220
CH3 - speakers	2,801,065	CH13 - dish washer	1,687,175
CH4 - server	2,806,036	CH14 - fridge	1,687,285
CH5 - router	2,795,349	CH15 - microwave	1,685,519
CH6 - server hdd	2,094,586	CH16 - toaster	1,685,322
CH7 - kettle	2,094,523	CH17 - playstation	1,686,903
CH8 - rice cooker	2,080,995	CH18 - modem	1,677,592
CH9 - running machine	2,089,140	CH19 - cooker	1,679,203
CH10 - laptop2	1,878,770		

Table 8.3: Number of samples in each individual channel, and the correspondent label for House 3, UK-DALE dataset. In this house, it is not necessary to eliminate any channel.

INDIVIDUAL CHANNEL LABEL – HOUSE 3	SAMPLES
CH1 - main channel	512,327
CH2 – kettle	515,845
CH3 – electric heater	517,434
CH4 – laptop	517,595
CH5 - projector	517,957

Table 8.4: Number of samples in each individual channel, and the correspondent label for House 4, UK-DALE dataset. In this house, it is not necessary to eliminate any channel.

INDIVIDUAL CHANNEL LABEL – HOUSE 4	SAMPLES
CH1 - main channel	2,186,446
CH2 – tv / dvd / digibox / lamp	2,156,167
CH3 – kettle / radio	2,171,770
CH4 – gas boiler	2,198,131
CH5 - freezer	2,194,864
CH6 – washing machine / microwave / breadmaker	2,180,830

Table 8.5: Number of samples in each individual channel, and the correspondent label for House 5, UK-DALE dataset. In this house, channels 11 and, 25 are not included for having too few samples compared to the other channels.

INDIVIDUAL CHANNEL LABEL – HOUSE 5	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 5	SAMPLES
CH1 - main channel	1,763,101	CH14 – atom pc	985,998
CH2 – stereo speakers bedroom	984,881	CH15 – toaster	985,856
CH3 – i7 desktop	1,845,649	CH16 – home theatre amp	984,488
CH4 – hairdryer	1,854,414	CH17 – sky hd box	975,359
CH5 – primary tv	1,840,507	CH18 – kettle	985,855
CH6 – 24 inch lcd bedroom	1,857,061	CH19 – fridge freezer	1,842,971
CH7 – treadmill	1,851,917	CH20 – oven	1,842,883
CH8 – network attached storage	985,653	CH21 – electric hob	1,842,782
CH9 – core 2 server	1,848,688	CH22 - dishwasher	1,859,593
CH10 – 24 inch lcd	985,784	CH23 – microwave	1,859,545
CH11 – PS4	10,215	CH24 – washer dryer	1,859,562
CH12 – steam iron	985,779	CH25 – vacuum cleaner	79,850
CH13 – nespresso pixie	985,600		

Table 8.6: Number of samples in each individual channel, and the correspondent label for House 1, REDD dataset.

INDIVIDUAL CHANNEL LABEL – HOUSE 1	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 1	SAMPLES
CH1 - main channel 1	1,561,660	CH11 – microwave	745,878
CH2 – main channel 2	1,561,660	CH12 – bathroom gfi	745,878
CH3 – oven 1	745,878	CH13 – electric heat	745,878
CH4 – oven 2	745,878	CH14 – stove	745,878
CH5 – refrigerator	745,878	CH15 – kitchen outlets	745,878
CH6 – dishwasher	745,878	CH16 – kitchen outlets	745,878
CH7 – kitchen outlets	745,878	CH17 – lighting	745,878
CH8 – kitchen outlets	745,878	CH18 – lighting	745,878
CH9 – lighting	745,878	CH19 - washer dryer	745,878
CH10 – washer dryer	745,878	CH20 – washer dryer	745,878

Table 8.7: Number of samples in each individual channel, and the correspondent label for House 2, REDD dataset.

INDIVIDUAL CHANNEL LABEL – HOUSE 2	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 2	SAMPLES
CH1 - main channel 1	1,198,534	CH7 – washer dryer	318,759
CH2 – main channel 2	1,198,534	CH8 – kitchen outlets 2	318,759
CH3 – kitchen outlets 1	318,759	CH9 – refrigerator	318,759
CH4 – lighting	318,759	CH10 – dishwasher	318,759
CH5 – stove	318,759	CH11 – disposal	318,759
CH6 – microwave	318,759		

Table 8.8: Number of samples in each individual channel, and the correspondent label for House 3, REDD dataset.

INDIVIDUAL CHANNEL LABEL – HOUSE 3	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 3	SAMPLES
CH1 - main channel 1	1,427,284	CH12 – outlets unknown 3	404,107
CH2 – main channel 2	1,427,284	CH13 – washer dryer 1	404,107
CH3 – outlets unknown 1	404,107	CH14 – washer dryer 2	404,107
CH4 – outlets unknown 2	404,107	CH15 – lighting 3	404,107
CH5 – lighting 1	404,107	CH16 – microwave	404,107
CH6 – electronics	404,107	CH17 – lighting 4	404,107
CH7 – refrigerator	404,107	CH18 – smoke alarms	404,107
CH8 – disposal	404,107	CH19 – lighting 5	404,107
CH9 – dishwasher	404,107	CH20 – bathroom gfi	404,107
CH10 – furnace	404,107	CH21 – kitchen outlets 1	404,107
CH11 - lighting 2	404,107	CH22 – kitchen outlets 2	404,107

Table 8.9: Number of samples in each individual channel, and the correspondent label for House 4, REDD dataset.

INDIVIDUAL CHANNEL LABEL – HOUSE 4	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 4	SAMPLES
CH1 - main channel 1	1,679,839	CH11 – miscellaneous	570,363
CH2 – main channel 2	1,679,839	CH12 – smoke alarms	570,363
CH3 – lighting 1	570,363	CH13 – lighting 2	570,363
CH4 – furance	570,363	CH14 – kitchen outlets 2	570,363
CH5 – kitchen outlets 1	570,363	CH15 – dishwasher	570,363
CH6 – outlets unknown	570,363	CH16 – bathrrom gfi 1	570,363
CH7 – washer dryer	570,363	CH17 – bathrrom gfi 2	570,363
CH8 – stove	570,363	CH18 – lighting 3	570,363
CH9 – air conditioning 1	570,363	CH19 – lighting 4	570,363
CH10 – air conditioning 2	570,363	CH20 – air conditioning 3	570,363

Table 8.10: Number of samples in each individual channel, and the correspondent label for House 5, REDD dataset.

INDIVIDUAL CHANNEL LABEL – HOUSE 5	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 5	SAMPLES
CH1 - main channel 1	302,122	CH14 – lighting 2	80,417
CH2 – main channel 2	302,122	CH15 – outlets unknown 3	80,417
CH3 – microwave	80,417	CH16 – bathrrom gfi	80,417
CH4 – lighting 1	80,417	CH17 – lighting 3	80,417
CH5 – outlets unknown 1	80,417	CH18 – refrigerator	80,417
CH6 – furance	80,417	CH19 – lighting 4	80,417
CH7 – outlets unknown 2	80,417	CH20 – dishwasher	80,417
CH8 – washer dryer 1	80,417	CH21 – disposal	80,417
CH9 – washer dryer 2	80,417	CH22 – electronics	80,417
CH10 – subpanel 1	80,417	CH23 – lighting 4	80,417
CH11 – subpanel 2	80,417	CH24 – kitchen outlets 1	80,417
CH12 – electric heat 1	80,417	CH25 - kitchen outlets 2	80,417
CH13 – electric heat 2	80,417	CH26 – outdoor outlets	80,417

Table 8.11: Number of samples in each individual channel, and the correspondent label for House 6, REDD dataset.

INDIVIDUAL CHANNEL LABEL – HOUSE 6	SAMPLES	INDIVIDUAL CHANNEL LABEL – HOUSE 6	SAMPLES
CH1 - main channel 1	887,457	CH10 – outlets unknown 1	376,968
CH2 – main channel 2	887,457	CH11 – outlets unknown 2	376,968
CH3 – kitchen outlets	376,968	CH12 – electric heat	376,968
CH4 – washer dryer	376,968	CH13 – kitchen outlets 2	376,968
CH5 – stove	376,968	CH14 – lighting	376,968
CH6 – electronics	376,968	CH15 – air conditioning 1	376,968
CH7 – bathroom gfi	376,968	CH16 – air conditioning 2	376,968
CH8 – refrigerator	376,968	CH17 – air conditioning 3	376,968
CH9 - dishwasher	376,968		

9. APENDIX B – PRINCIPAL COMPONENT ANALISYS RESULTS

9.1 UK-DALE, HOUSE 1

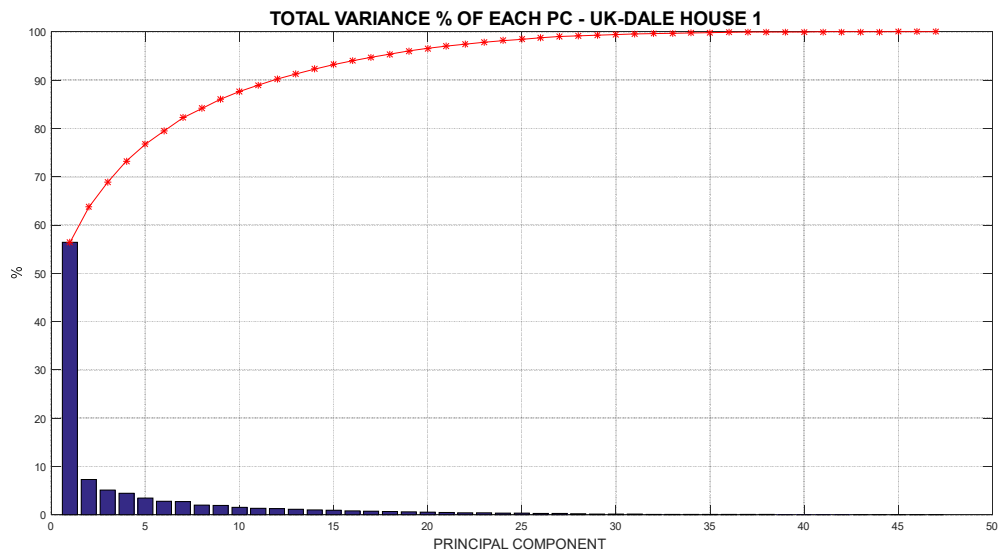


Figure 9.1: Latent (individual variance of each PC) and total variance accumulated for UK-DALE, House 1

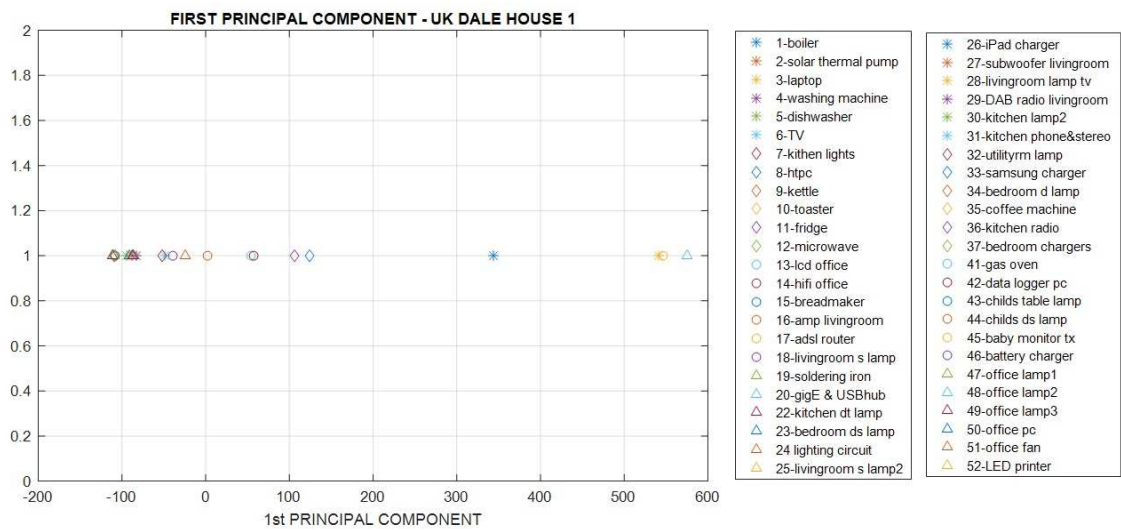


Figure 9.2: Data Projection over First Principal Component - UK-DALE, House 1

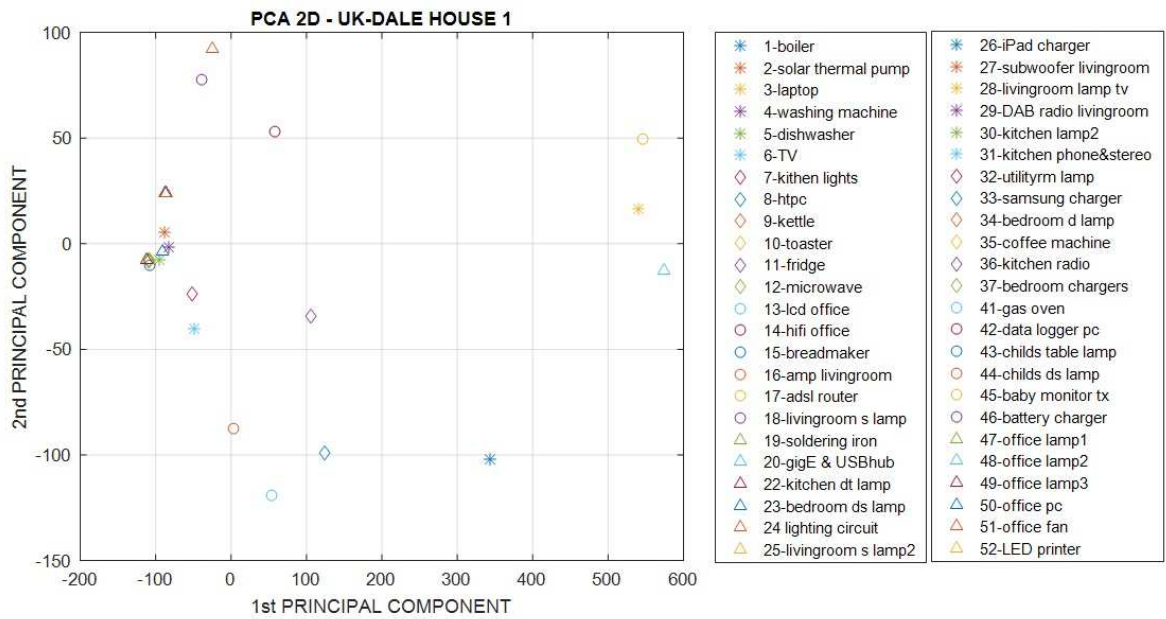


Figure 9.3: Data Projection over First and Second Principal Components - UK-DALE, House 1

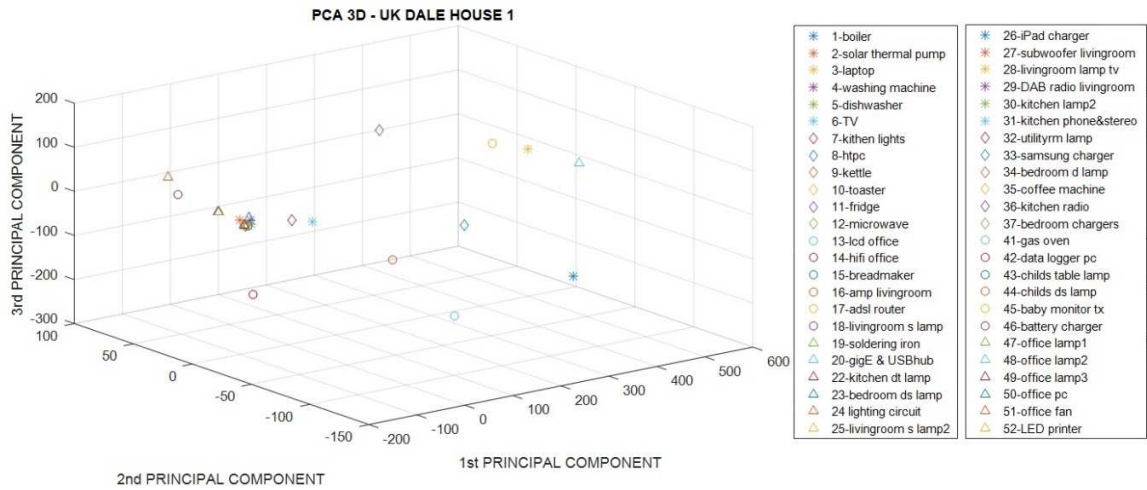


Figure 9.4: Data Projection over First, Second and Third Principal Components - UK-DALE, House 12

9.2 UK-DALE, HOUSE 2

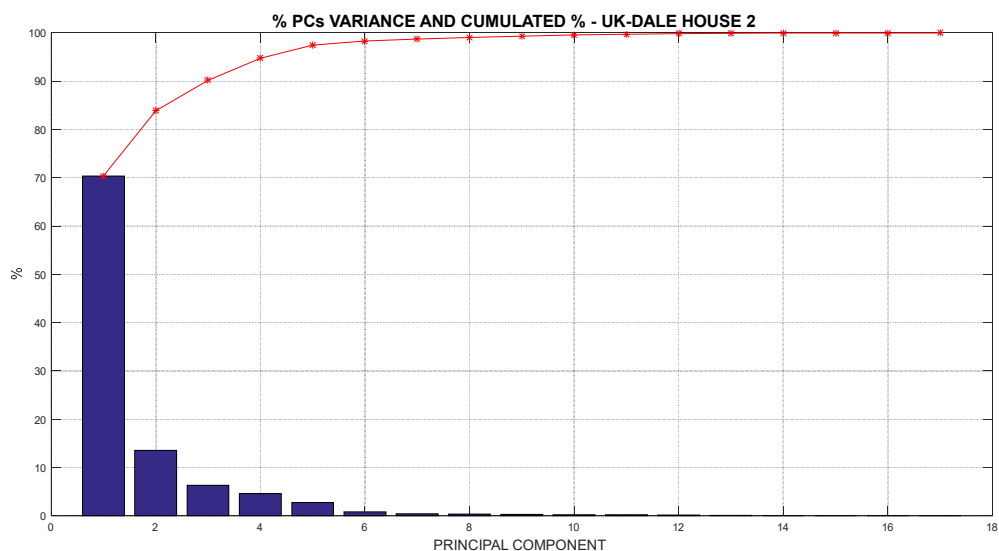


Figure 9.5: Variance for each PC and cummulated for UK-DALE, House 2

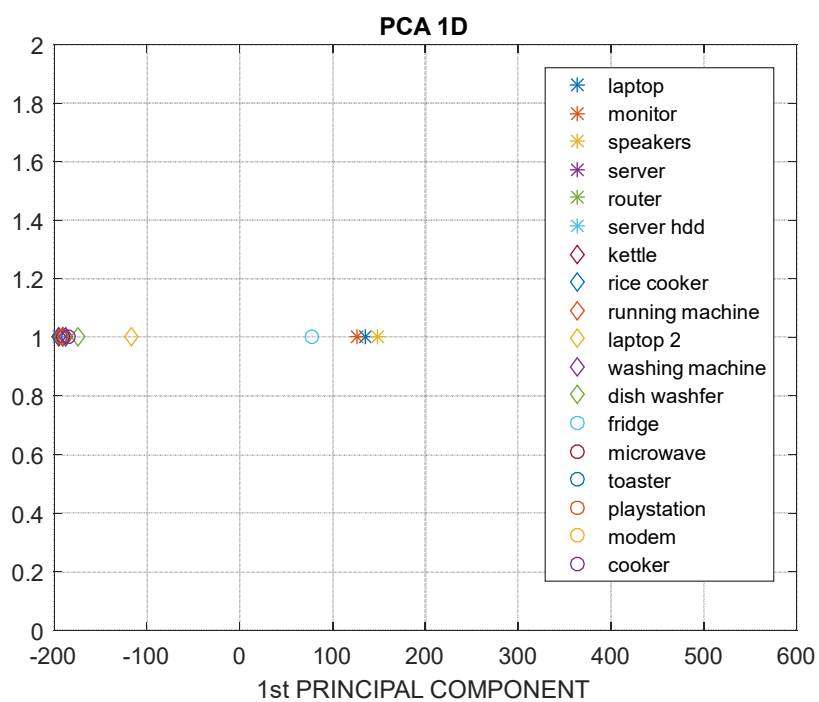


Figure 9.6: Data Projection over First Principal Component - UK-DALE, House 2

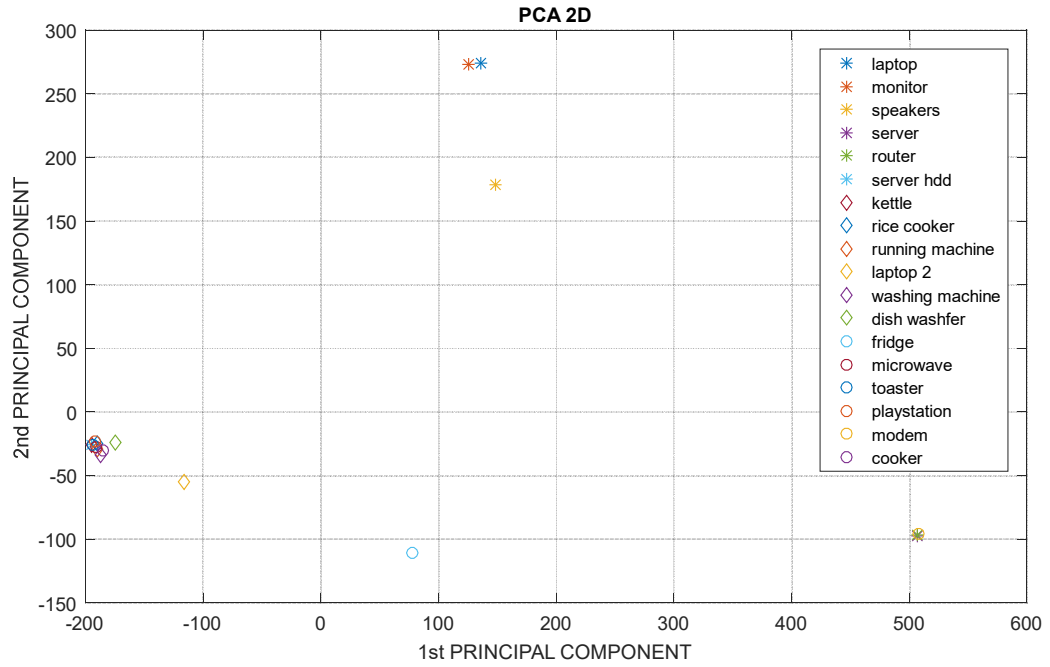


Figure 9.7: Data Projection over First and Second Principal Components - UK-DALE, House 2

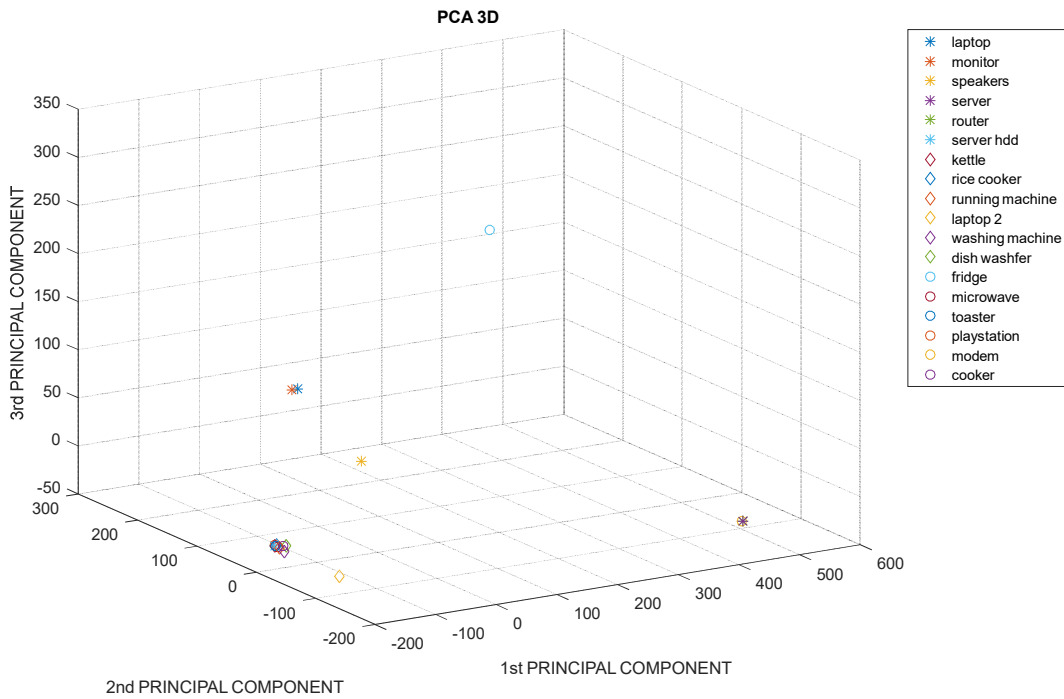


Figure 9.8: Data Projection over First, Second and Third Principal Components - UK-DALE, House 2

9.3 UK-DALE, HOUSE 3

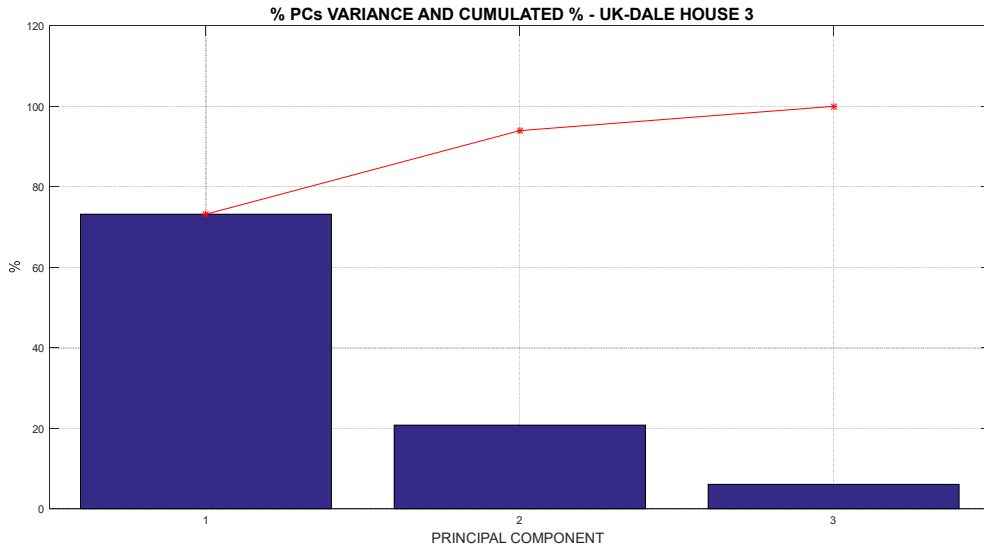


Figure 9.9: Variance for each PC and cumulated for UK-DALE, House 3

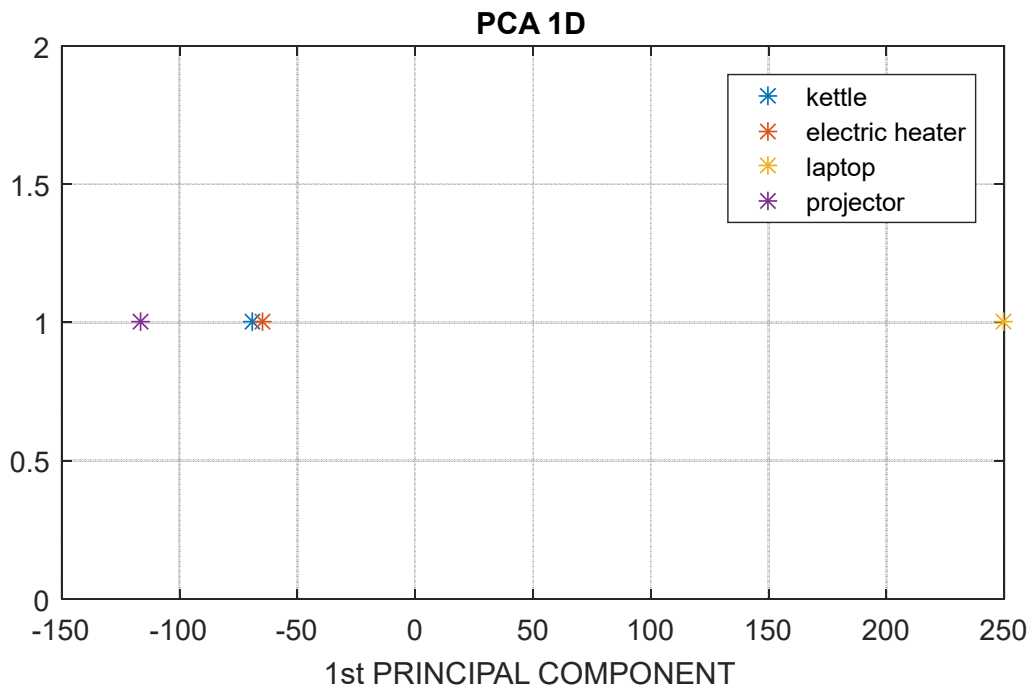


Figure 9.10: Data Projection over First Principal Component - UK-DALE, House 3

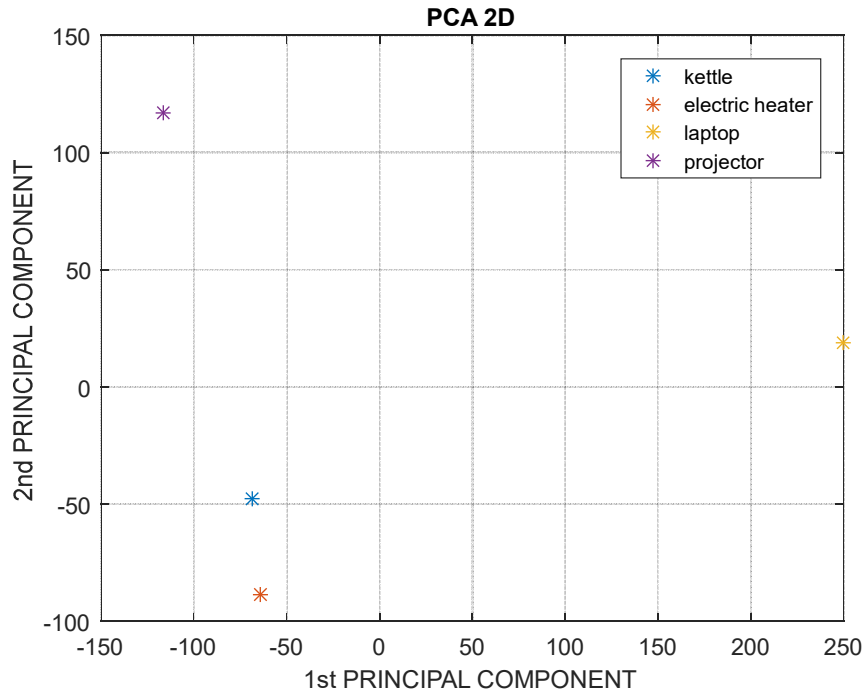


Figure 9.11: Data Projection over First and Second Principal Components - UK-DALE, House 3

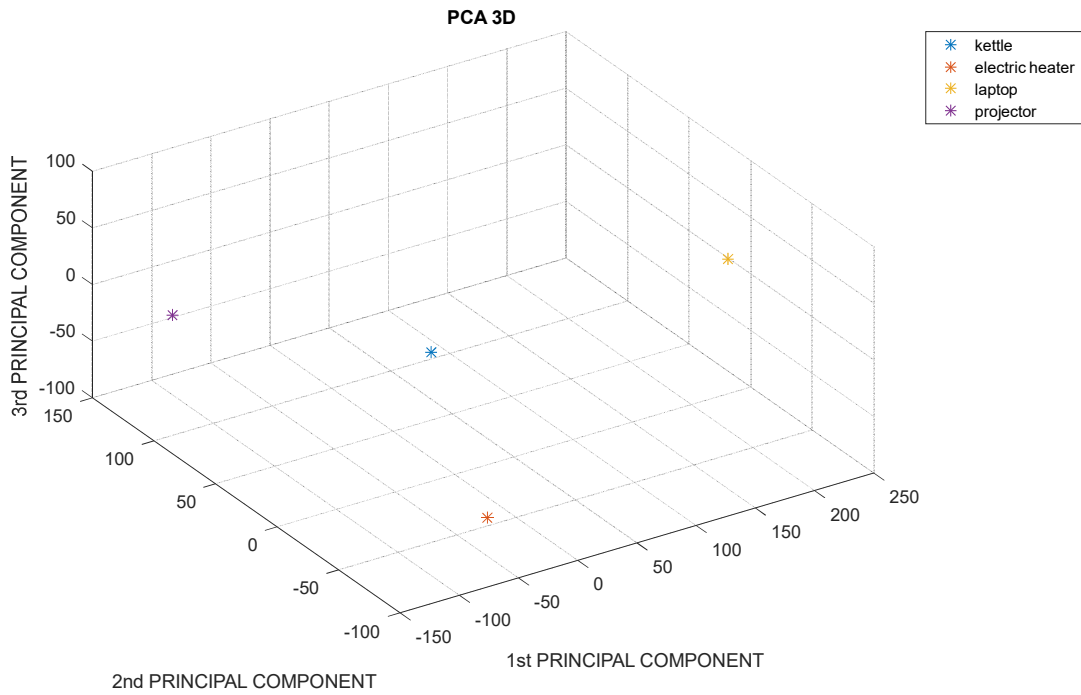


Figure 9.12: Data Projection over First, Second and Third Principal Components - UK-DALE, House 3

9.4 UK-DALE, HOUSE 4

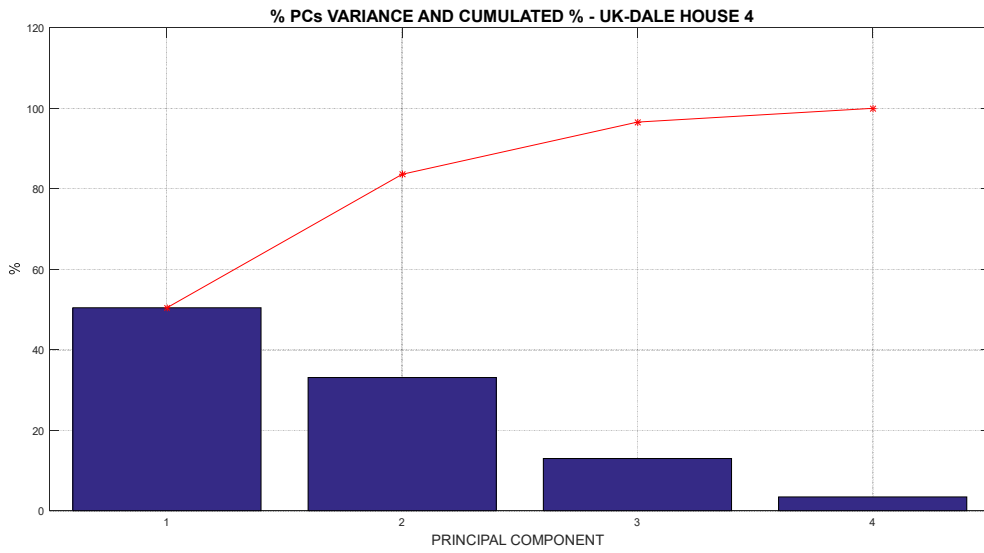


Figure 9.13: Variance for each PC and cumulated for UK-DALE, House 4

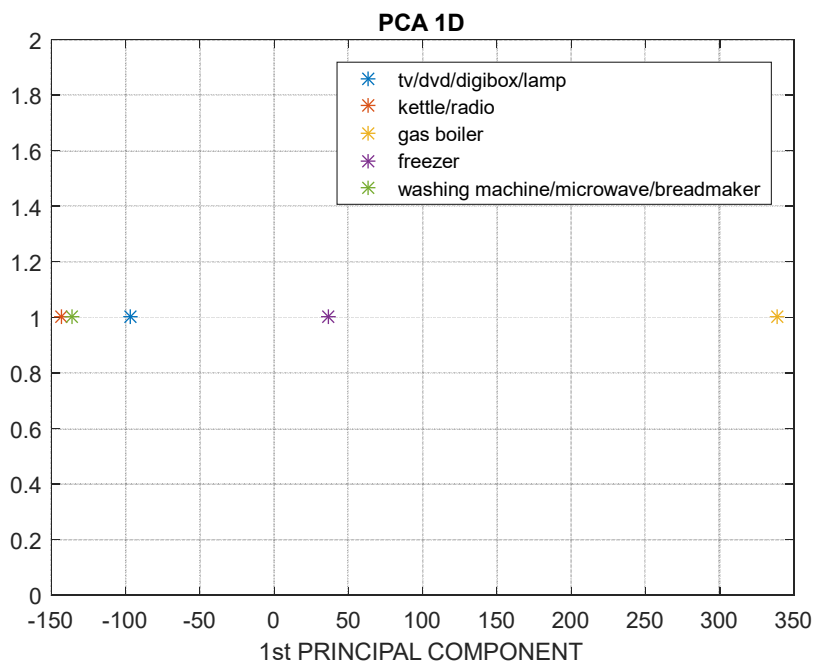


Figure 9.14: Data Projection over First Principal Component - UK-DALE, House 4

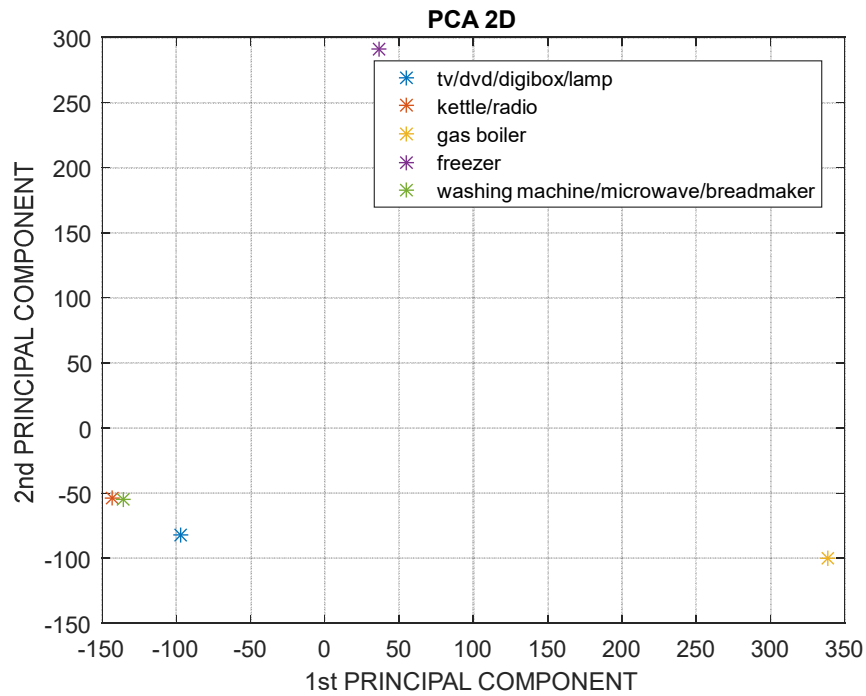


Figure 9.15: Data Projection over First and Second Principal Components - UK-DALE, House 4

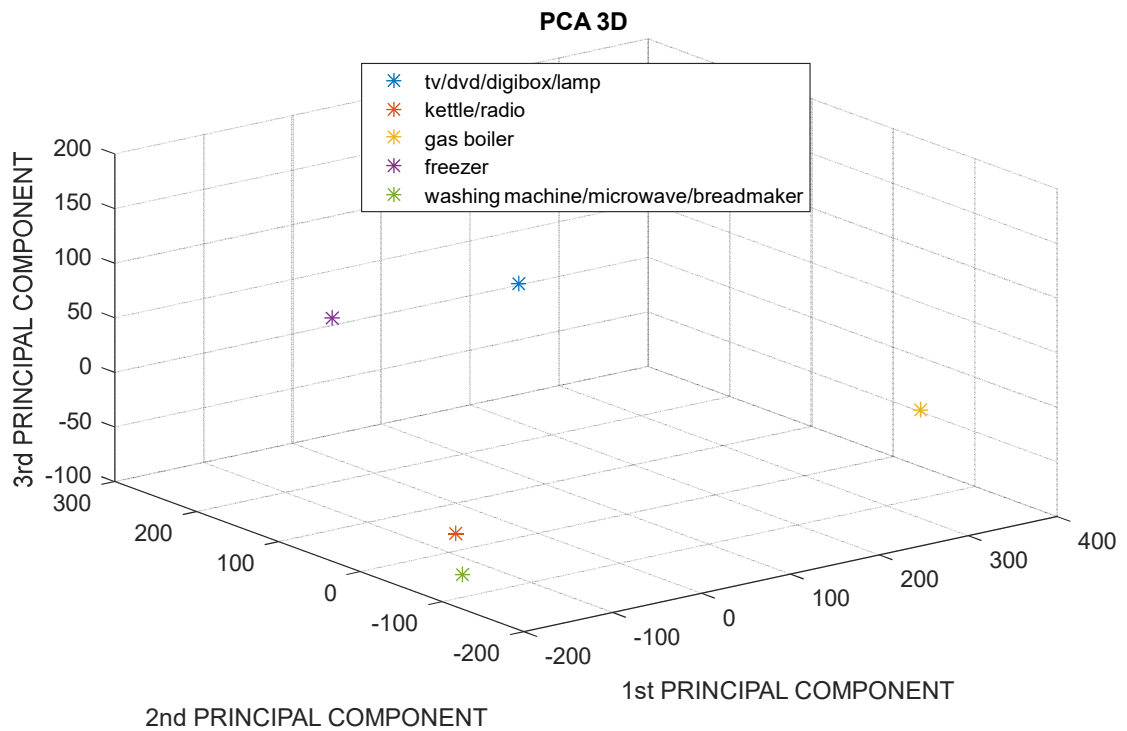


Figure 9.16: Data Projection over First, Second and Third Principal Components - UK-DALE, House

9.5 UK-DALE, HOUSE 5

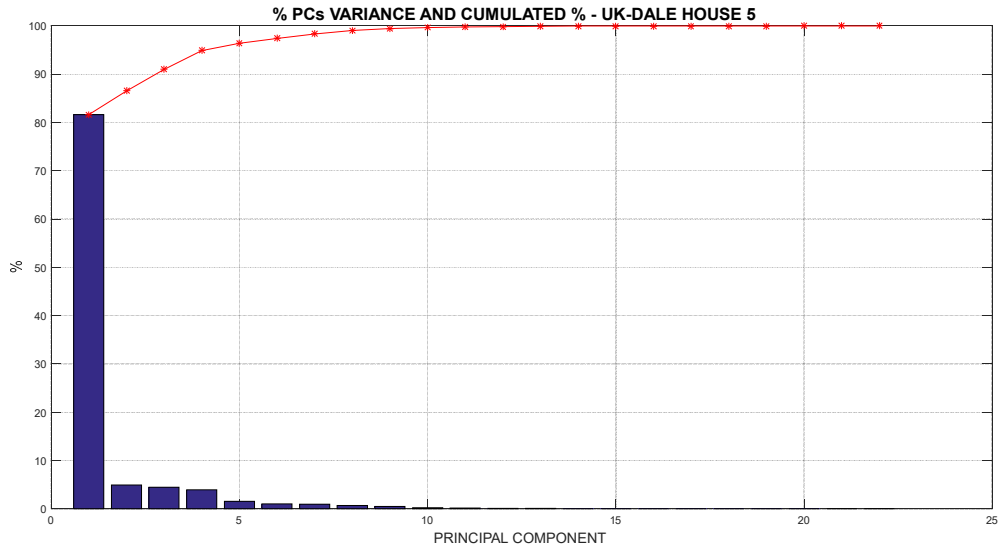


Figure 9.17: Variance for each PC and cummulated for UK-DALE, House 5

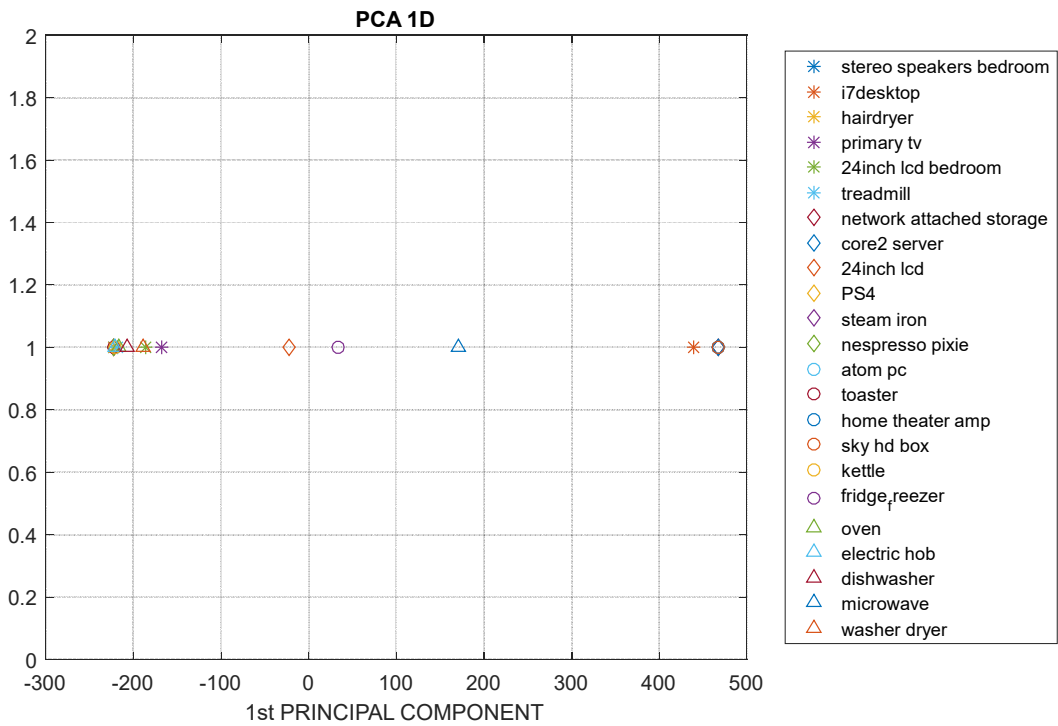


Figure 9.18: Data Projection over First Principal Component - UK-DALE, House 5

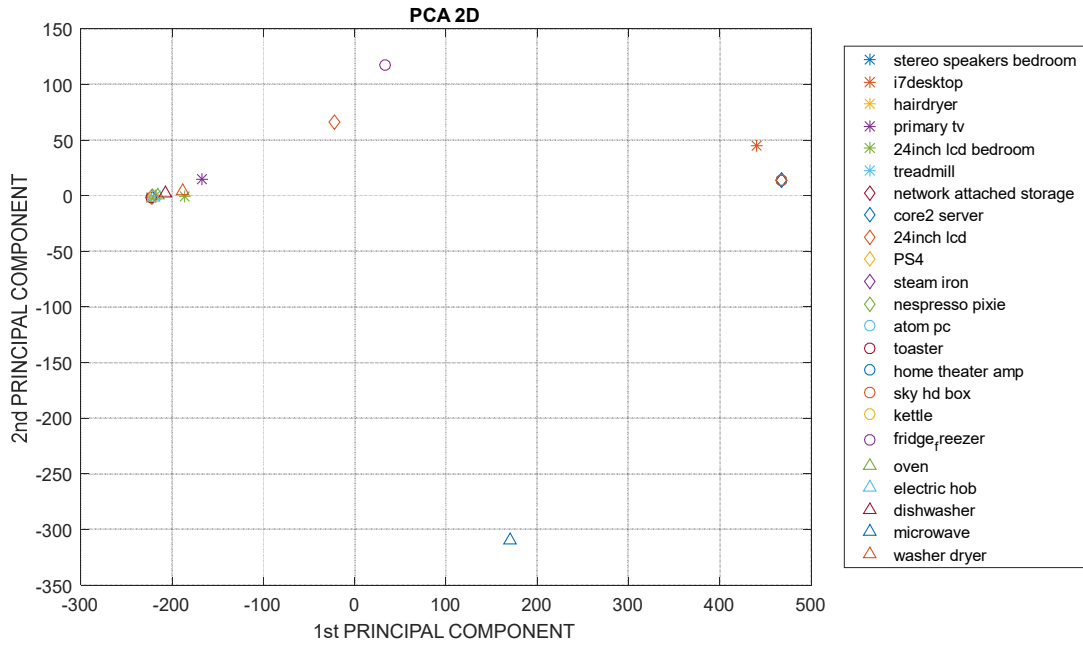


Figure 9.19: Data Projection over First and Second Principal Components - UK-DALE, House 5

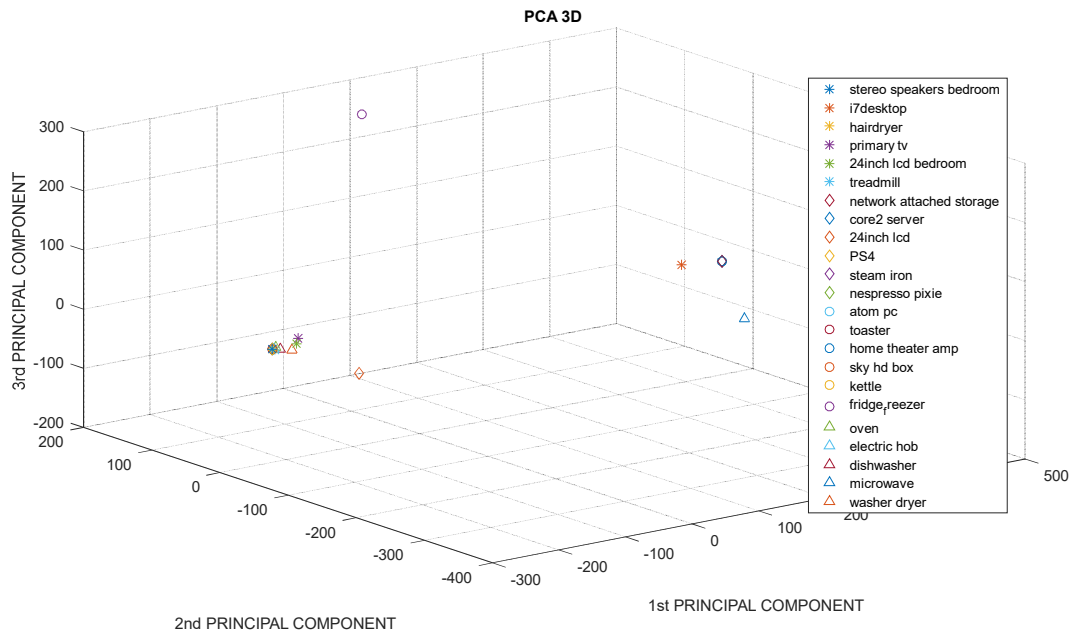


Figure 9.20: Data Projection over First, Second and Third Principal Components - UK-DALE, House 5

9.6 REDD, HOUSE 1

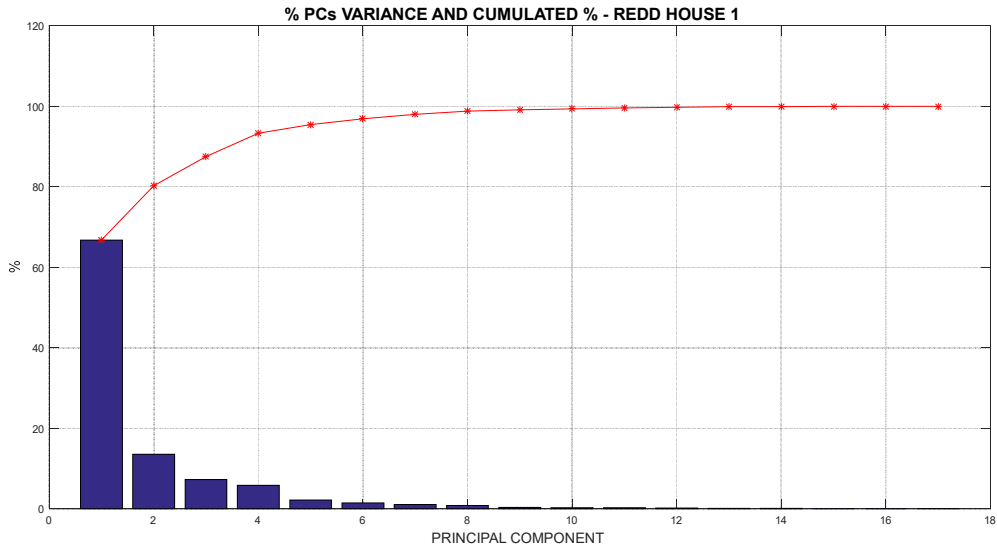


Figure 9.21: Variance for each PC and cumulated for REDD, House 1

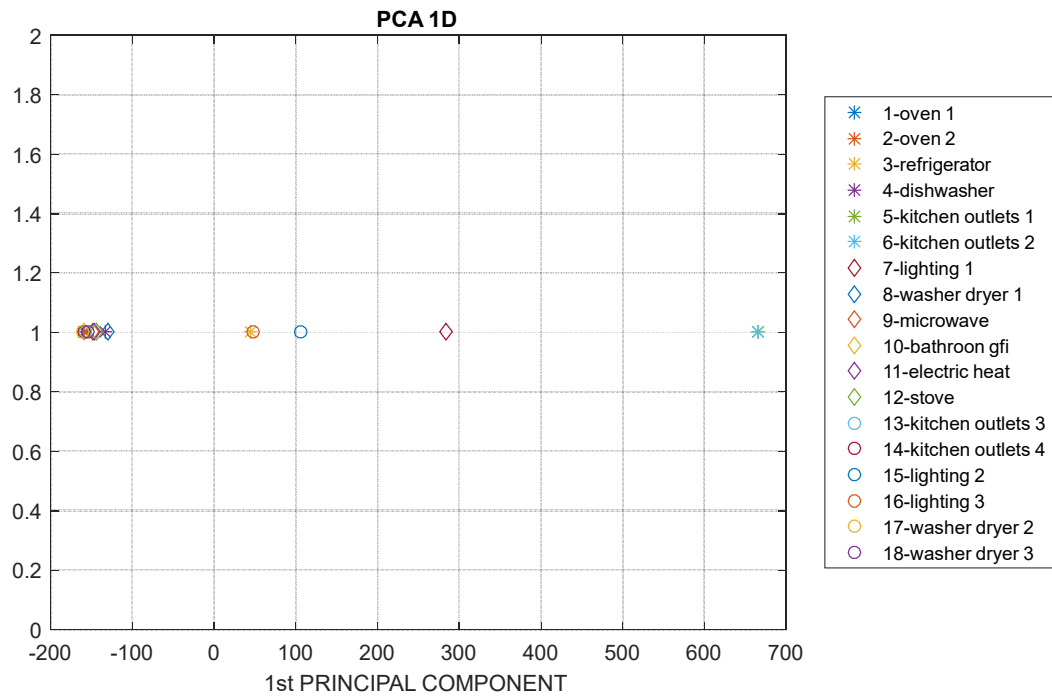


Figure 9.22: Data Projection over First Principal Component - REDD, House 1

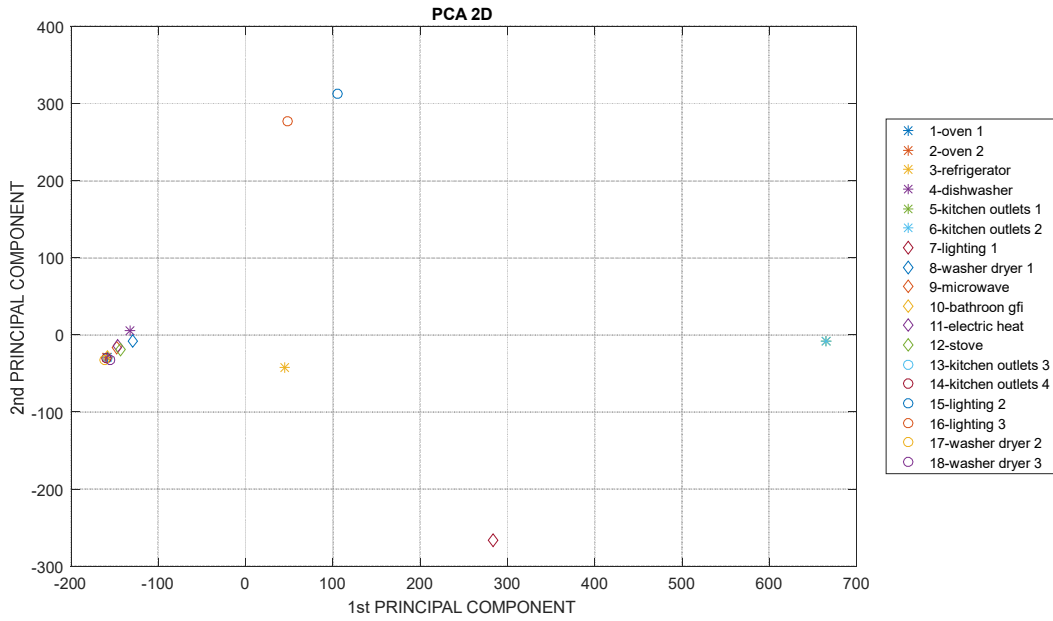


Figure 9.23: Data Projection over First and Second Principal Components - REDD, House 1

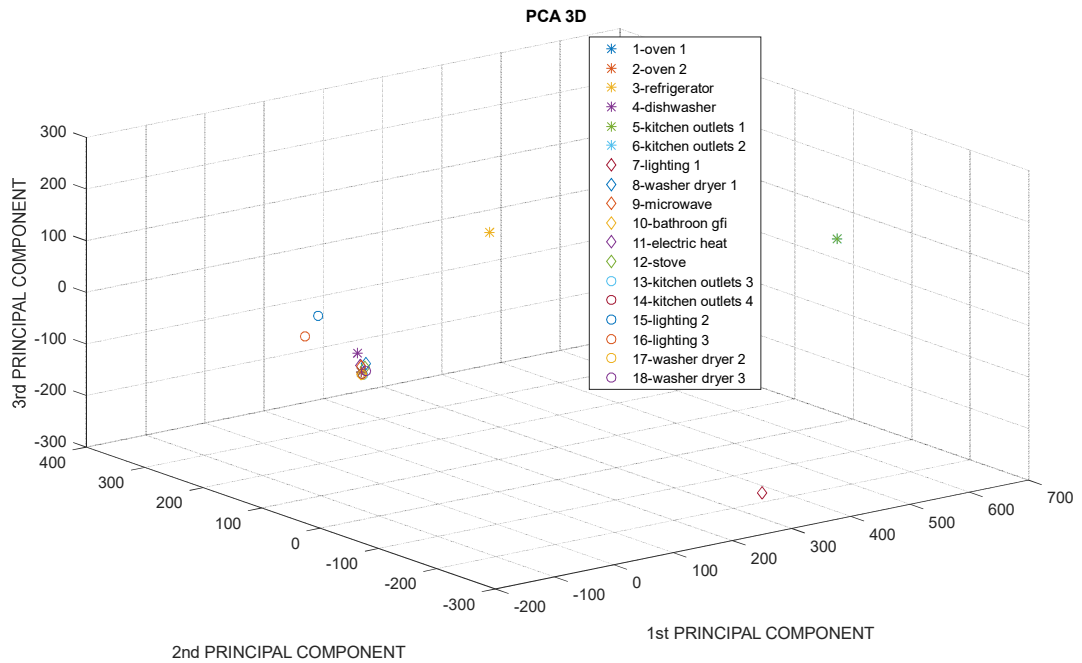


Figure 9.24: Data Projection over First, Second and Third Principal Components - REDD, House 1

9.7 REDD, HOUSE 2

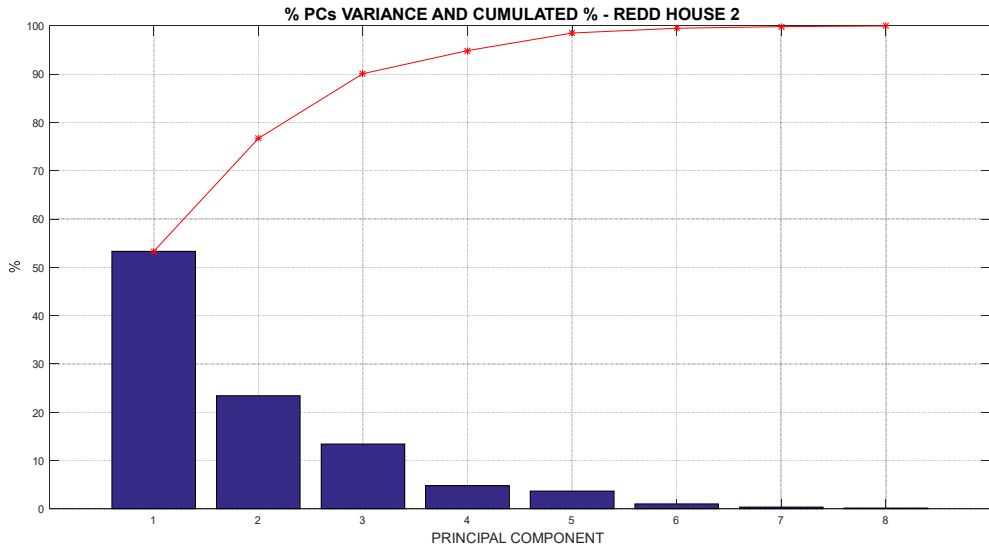


Figure 9.25: Variance for each PC and cumulated for REDD, House 2

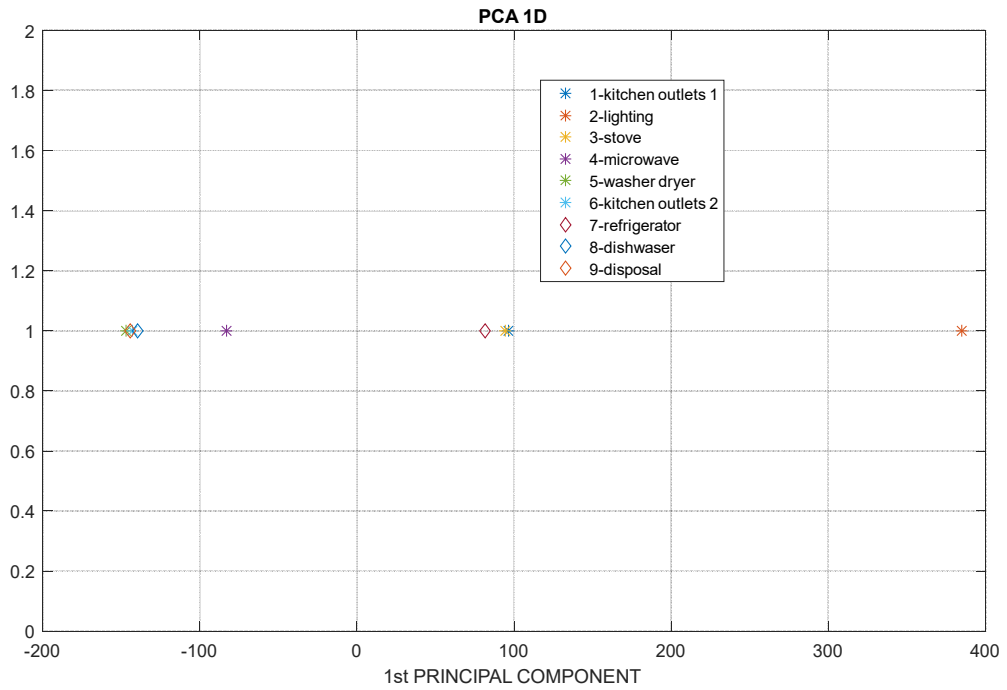


Figure 9.26: Data Projection over First Principal Component - REDD, House 2

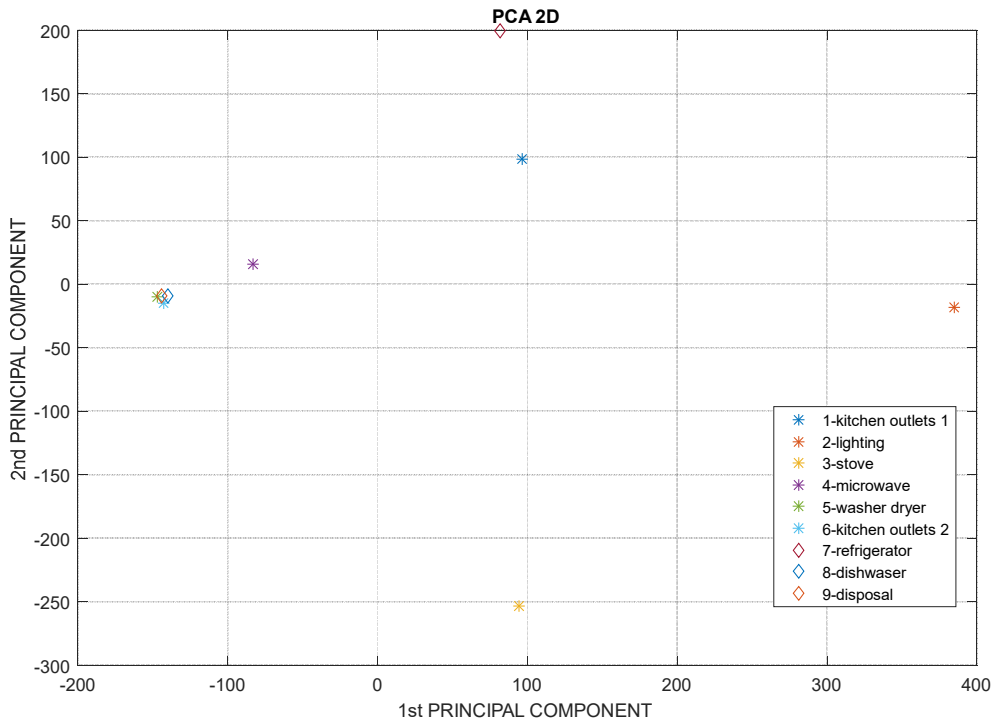


Figure 9.27: Data Projection over First and Second Principal Components - REDD, House 2

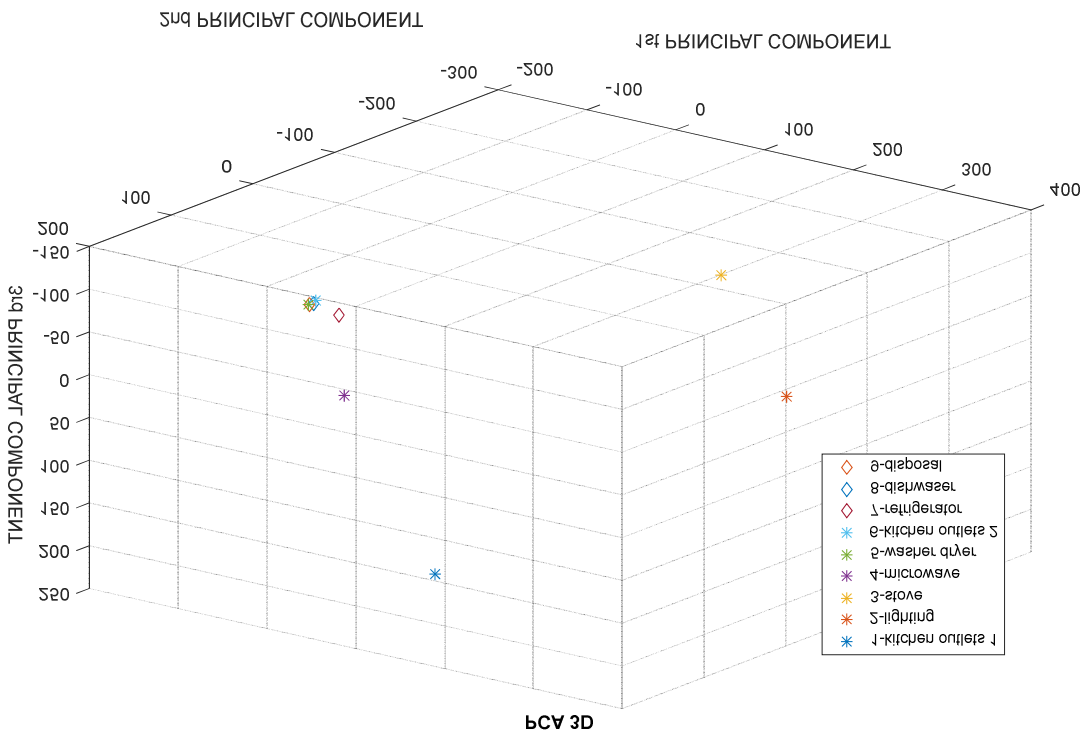


Figure 9.28: Data Projection over First, Second and Third Principal Components - REDD, House 2

9.8 REDD, HOUSE 3

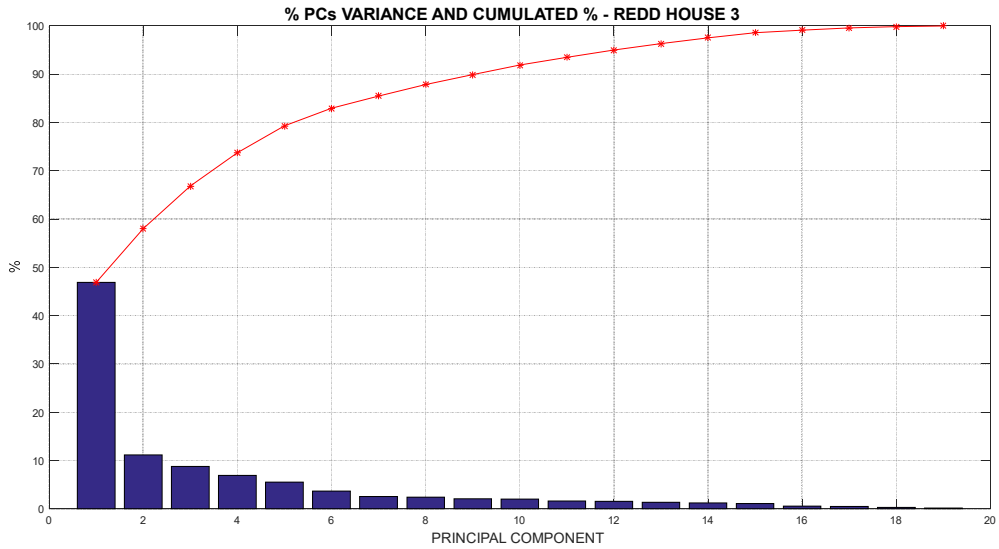


Figure 9.29: Variance for each PC and cumulated for REDD, House 3

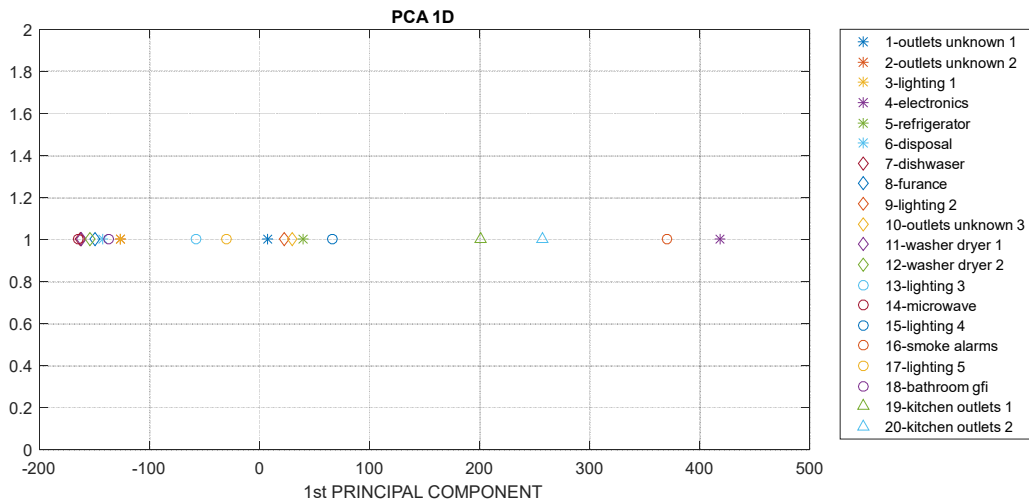


Figure 9.30: Data Projection over First Principal Component - REDD, House 3

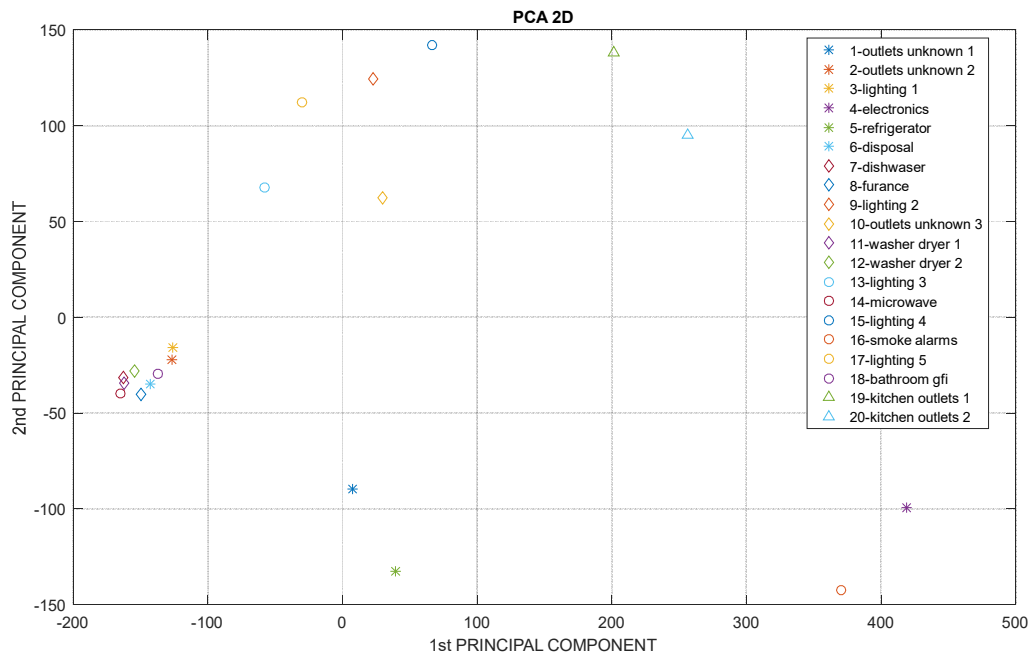


Figure 9.31: Data Projection over First and Second Principal Components - REDD, House 3

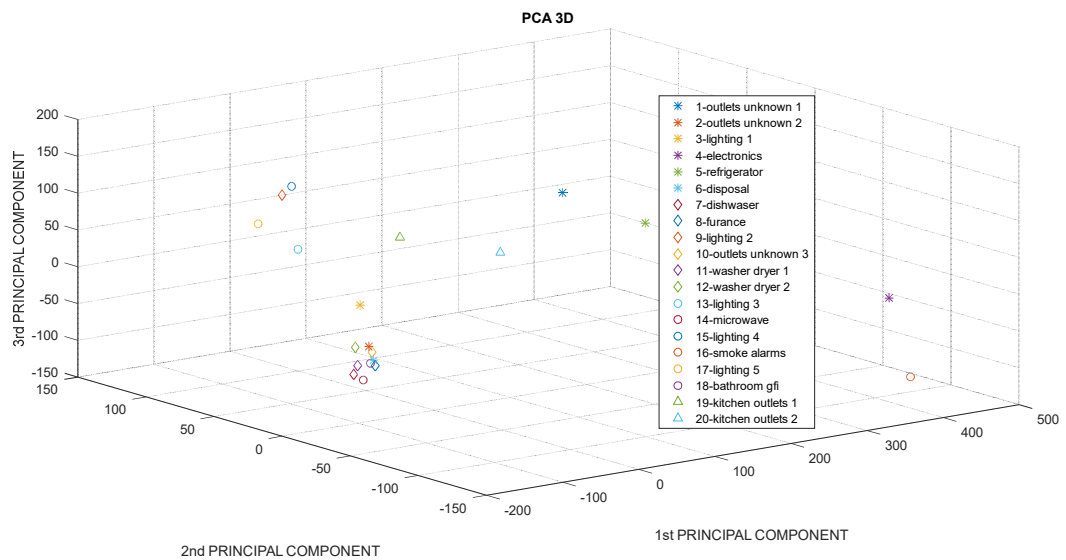


Figure 9.32: Data Projection over First, Second and Third Principal Components - REDD, House 3

9.9 REDD, HOUSE 4

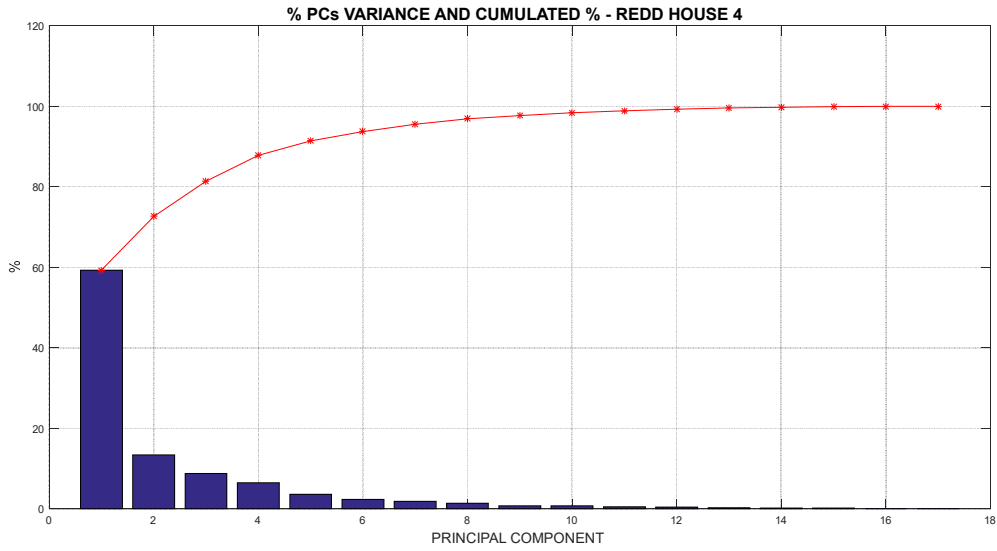


Figure 9.33: Variance for each PC and accumulated for REDD, House 4

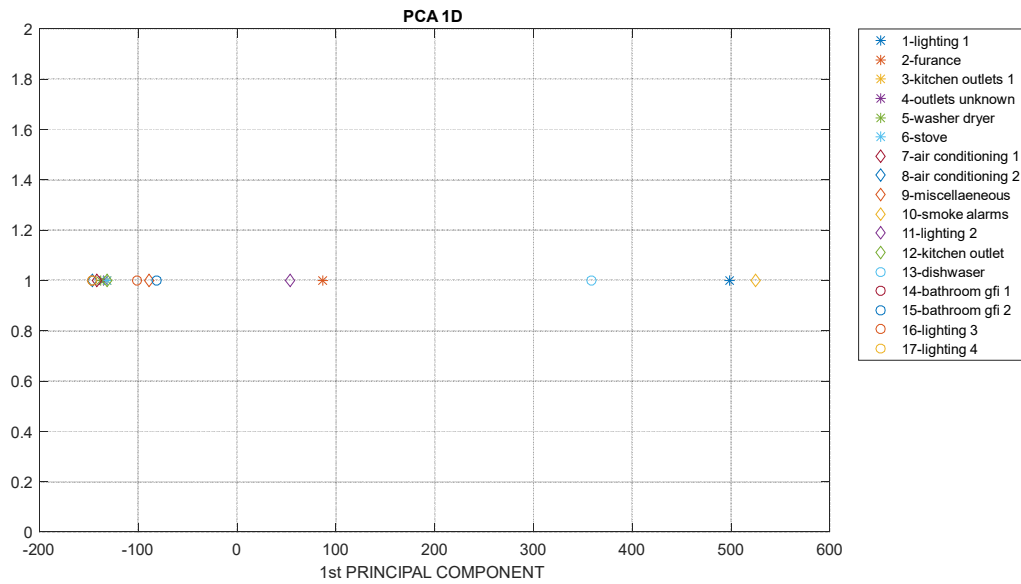


Figure 9.34: Data Projection over First Principal Component - REDD, House 4

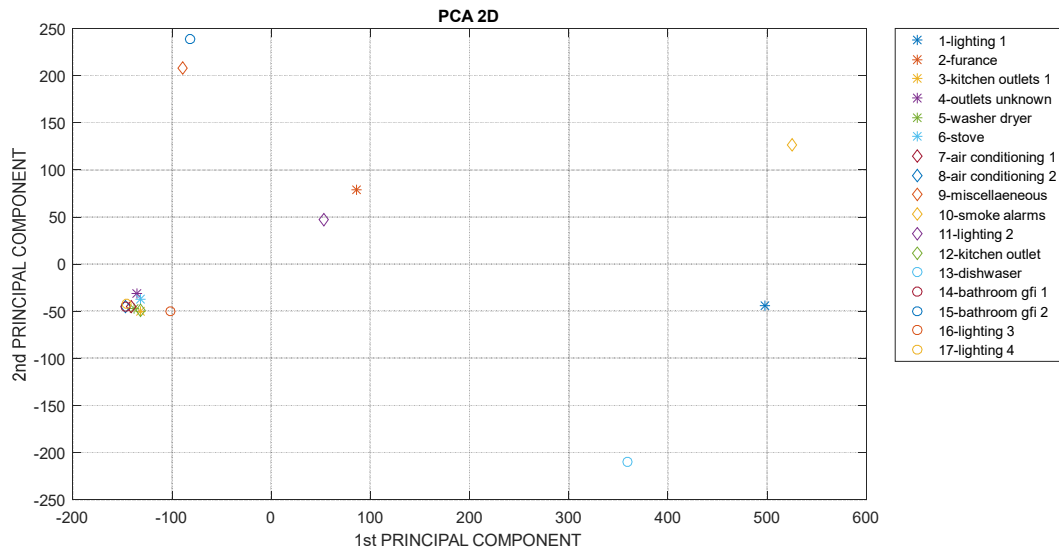


Figure 9.35: Data Projection over First and Second Principal Components - REDD, House 4

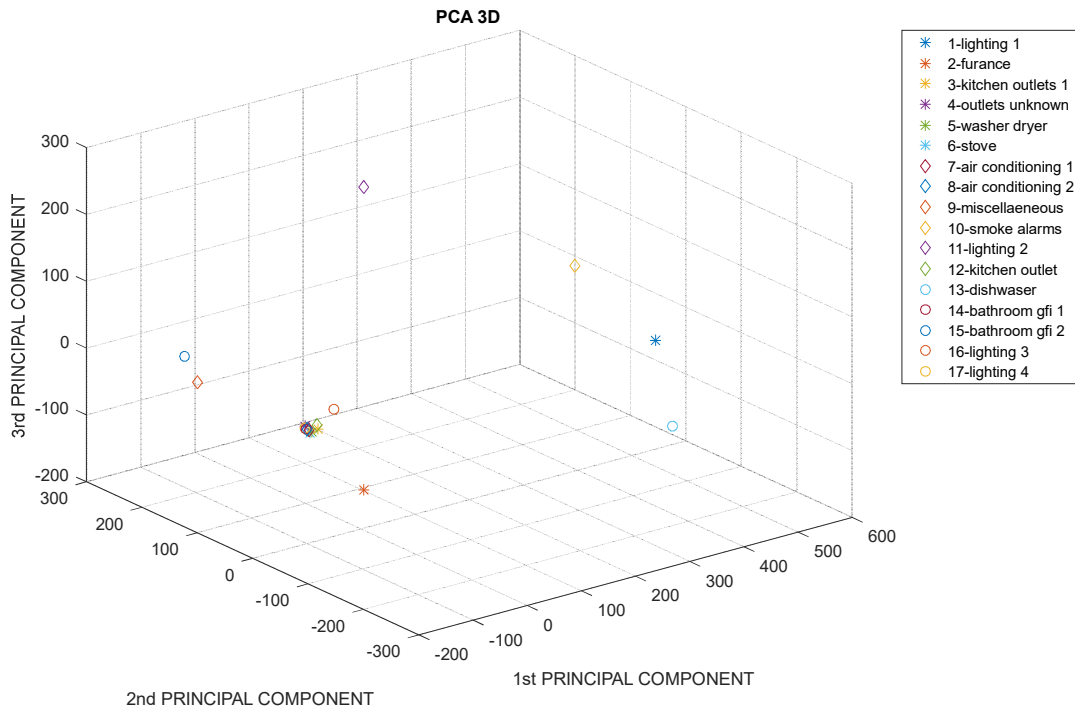


Figure 9.36: Data Projection over First, Second and Third Principal Components - REDD, House 4

9.10 REDD, HOUSE 5

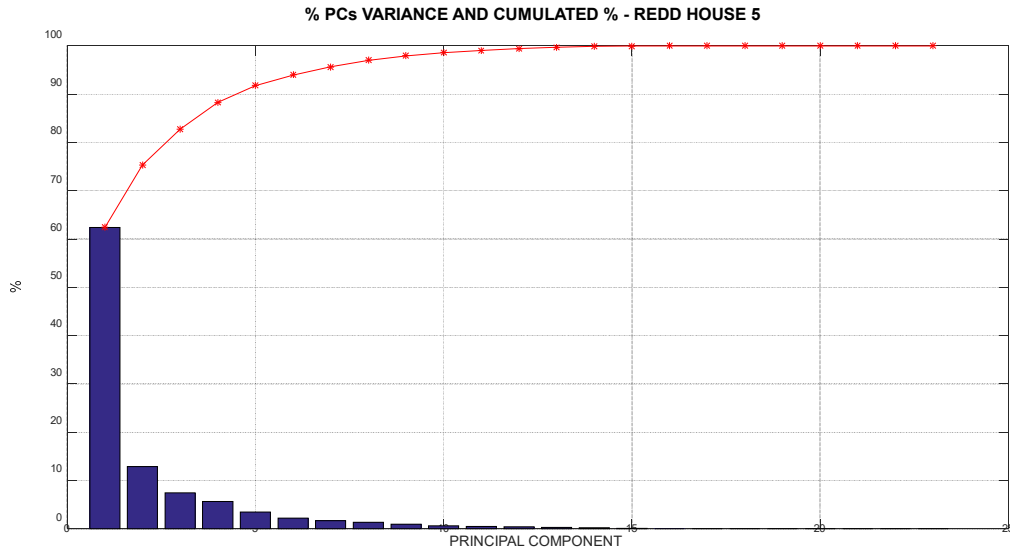


Figure 9.37: Variance for each PC and cummulated for REDD, House 5

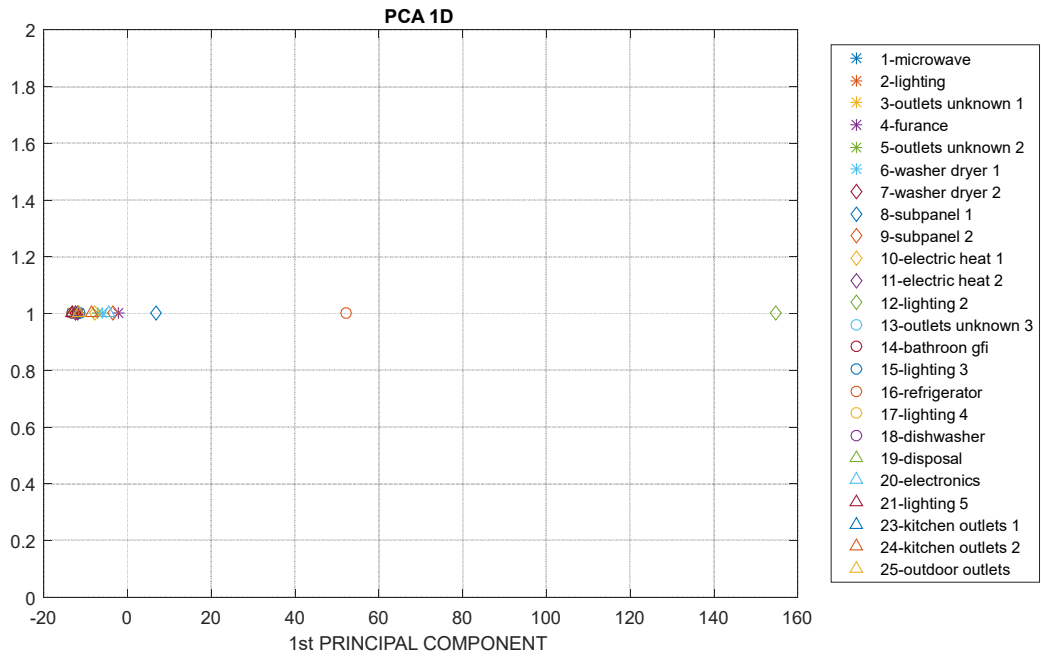


Figure 9.38: Data Projection over First Principal Component - REDD, House 5

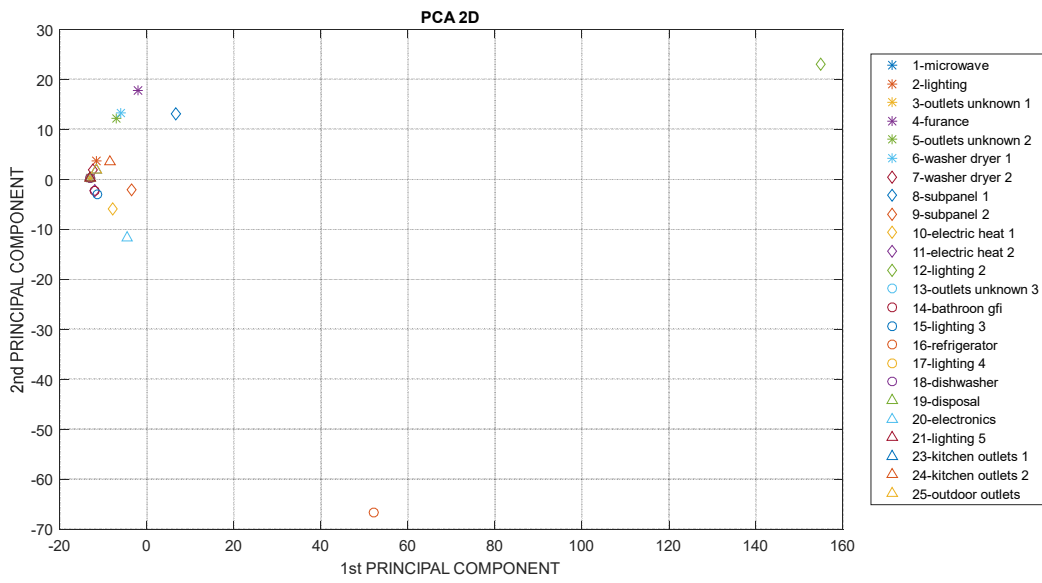


Figure 9.39: Data Projection over First and Second Principal Components - REDD, House 5

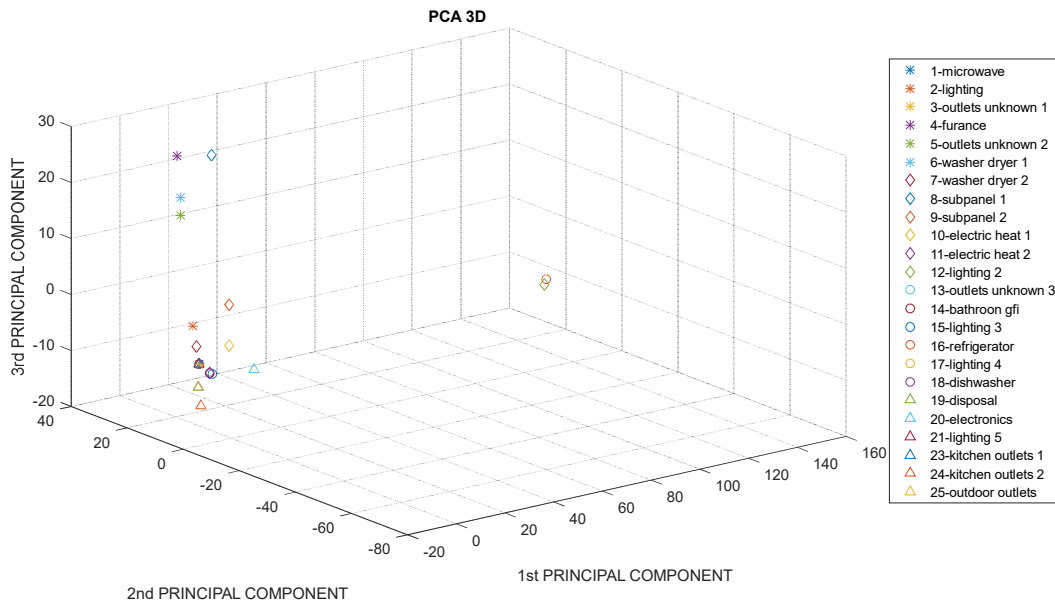


Figure 9.40: Data Projection over First, Second and Third Principal Components - REDD, House 5

9.11 REDD, HOUSE 6

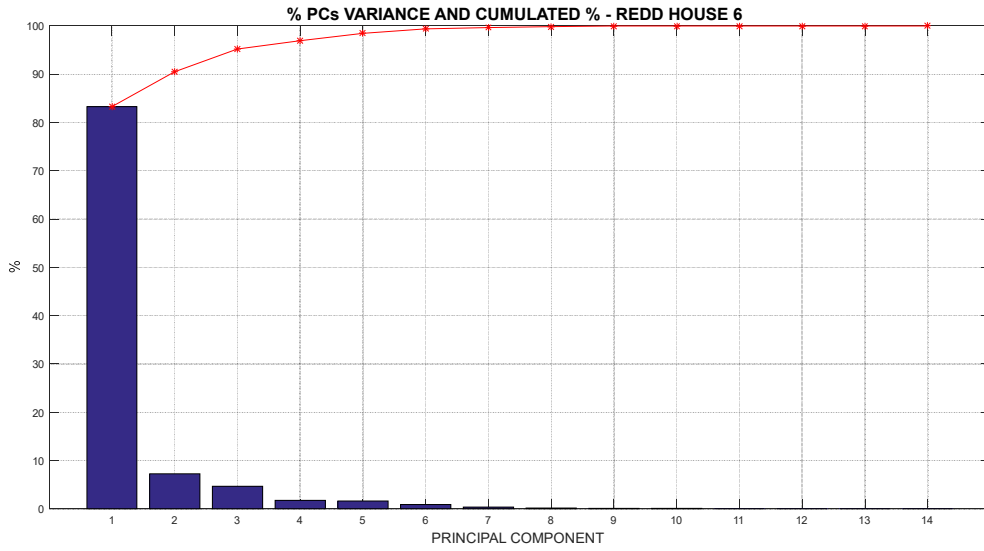


Figure 9.41: Variance for each PC and cumulated for REDD, House 6

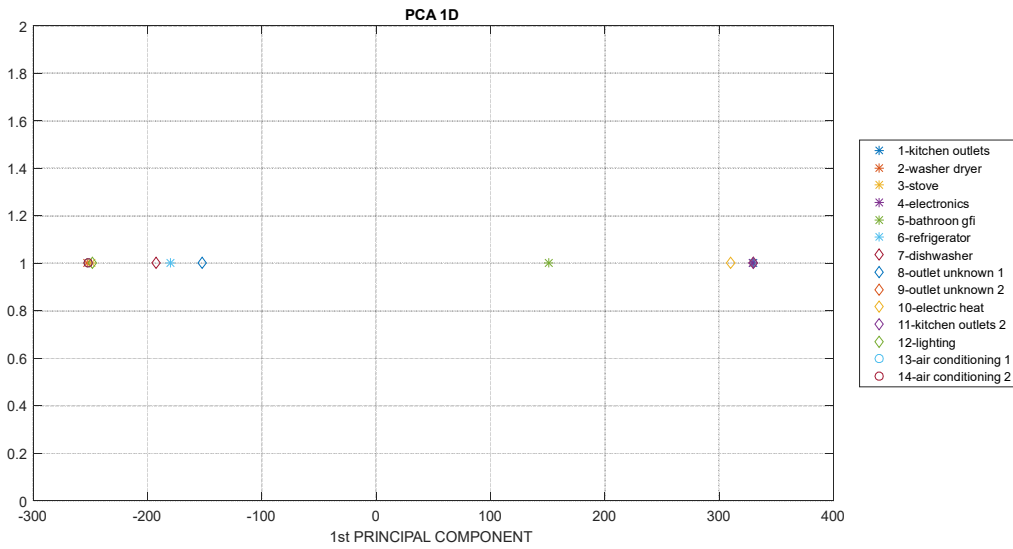


Figure 9.42: Data Projection over First Principal Component - REDD, House 6

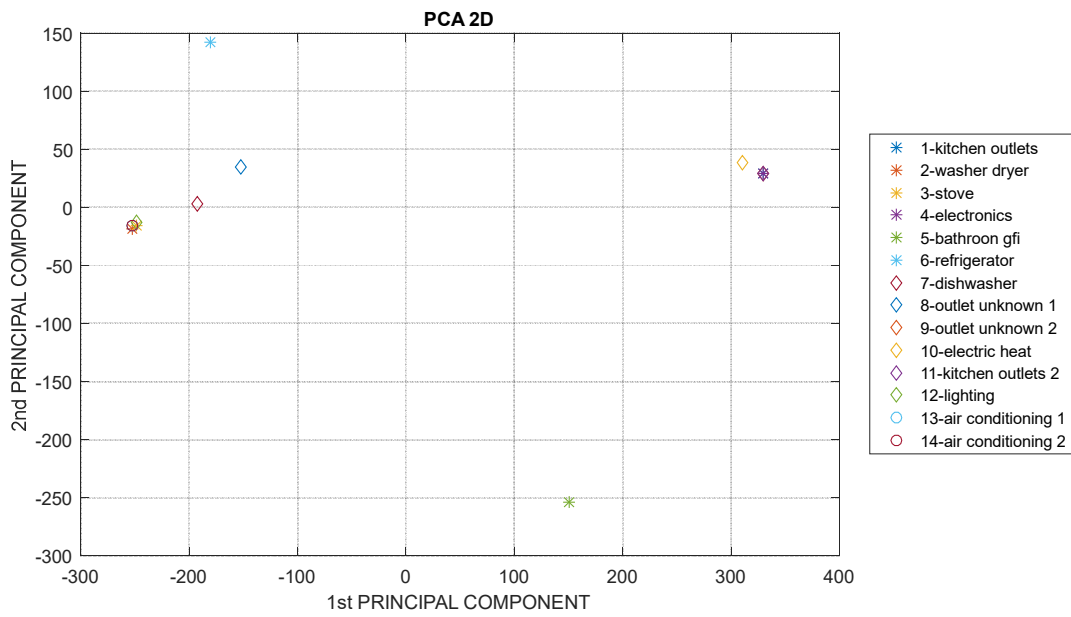


Figure 9.43: Data Projection over First and Second Principal Components - REDD, House 6

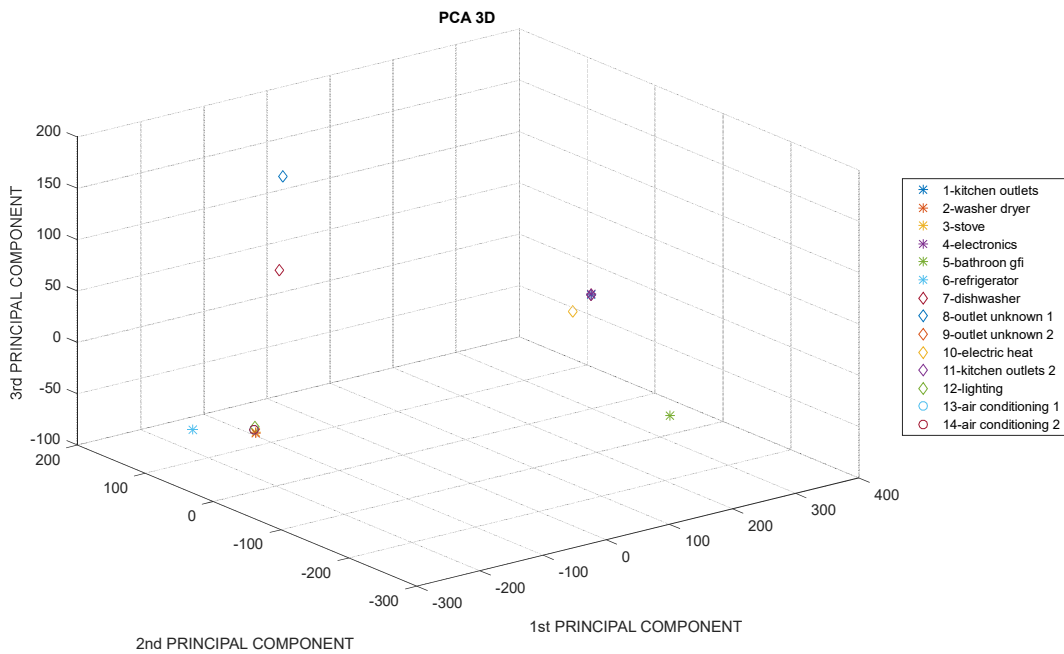


Figure 9.44: Data Projection over First, Second and Third Principal Components - REDD, House 6

10. APENDIX C – K-MEANS RESULTS

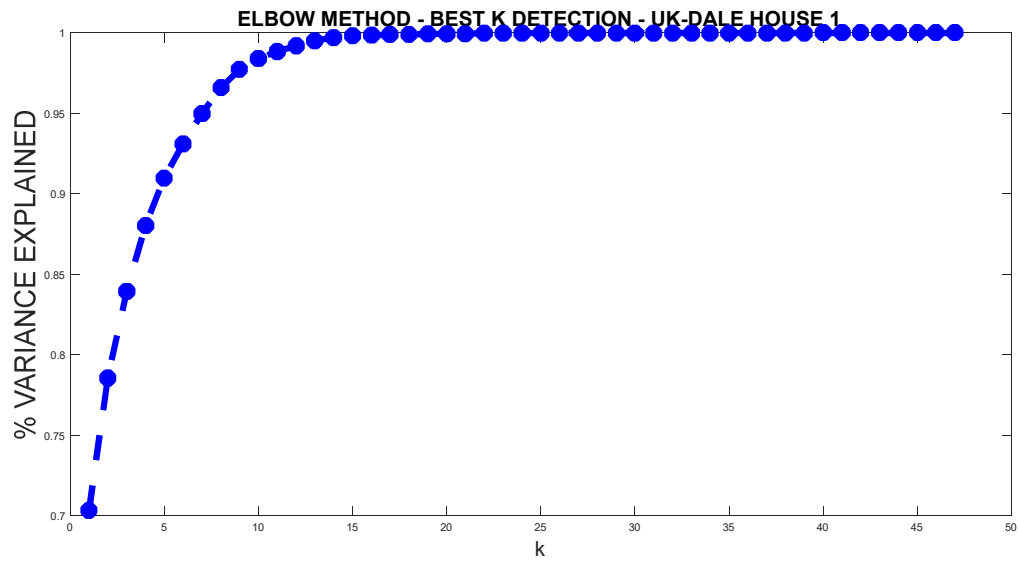


Figure 10.1: Variance Explained according to the number of clusters (k), for UK-DALE, House 1. The best k is 6.

Table 10.1: Clusters for UK-DALE House 1, $k=6$. The groups with “+” are considered as a single point

CLUSTER INDEX	CHANNELS
1	1-BOILER 32-UTILITYRM LAMP
2	2-SOLAR THERMAL PUMP 4-WASHING MACHINE 5-DISHWASHER 6-TV 7-KITCHEN LIGHTS 9-KETTLE + 19-SOLDERING IRON 10-TOASTER 12-MICROWAVE 15-BREADMAKER 16-AMP LIVING ROOM 18-LIVINGROOMS LAMP 22-KITCHEN DT LAMP + 27-SUBWOOFER LIVINGROOM + 29-DAB RADIO LIVINGROOM + 33-SAMSUNG CHARGERS + 34- BEDROOM D LAMP + 35- COFFEE MACHINE + 36 – KITCHEN RADIO + 41- GAS OVEN + 45- BABY MONITOR TX + 52-LED PRINTER 23-BEDROOM DS LAMP 24-LIGHTING CIRCUIT 25-LIVINGROOM LAMP 2 26-IPAD CHARGER 28-LIVINGROOM LAMP TV 31-KITCHEN PHONE AND STEREO 36-BEDROOM CHARGERS 37- BEDROOM CHARGERS 43-CHILDS TABLE LAMP 44- CHILDS DS LAMP 46-BATTERY CHARGER 47-OFFICE LAMP 2 48-OFFICE LAMP 2 50-OFFICE PC 51-OFFICE FAN 14-HIFI OFFICE

	13-LCD OFFICE
3	3-LAPTOP 17-ADSL ROUTER 20-GIGE and USBHUB
4	13- LCD OFFICE 14- HIFI OFFICE
5	8-HTPC 11-FRIDGE 30-KITCHEN LAMP 2 49-OFFICE LAMP 3
6	42-DATA LOGGER PC

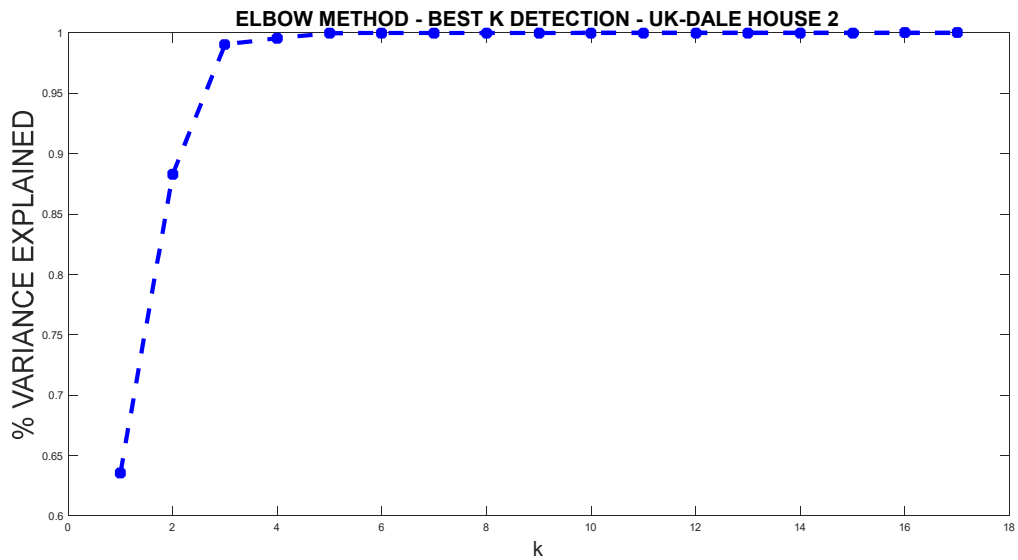


Figure 10.2: Variance Explained according to the number of clusters (k), for UK-DALE, House 2. The best k is 3.

Table 10.2: Clusters for UK-DALE House 2, $k=3$. The groups with “+” are considered as a single point.

CLUSTER INDEX	CHANNELS
1	1 - LAPTOP 2-MONITOR 3-SPEAKERS
2	17-MODEM 4-SERVER + 5-ROUTER
3	6-SERVER HDD + 7-KETTLE + 15-TOASTER 8-RICE COOKER 9-RUNNING MACHINE 10-LAPTOP 2 11-WASHING MACHINE 12-DISHWASHER 13-FRIDGE 14-MICROWAVE 16-PLAYSTATION 18-COOKER

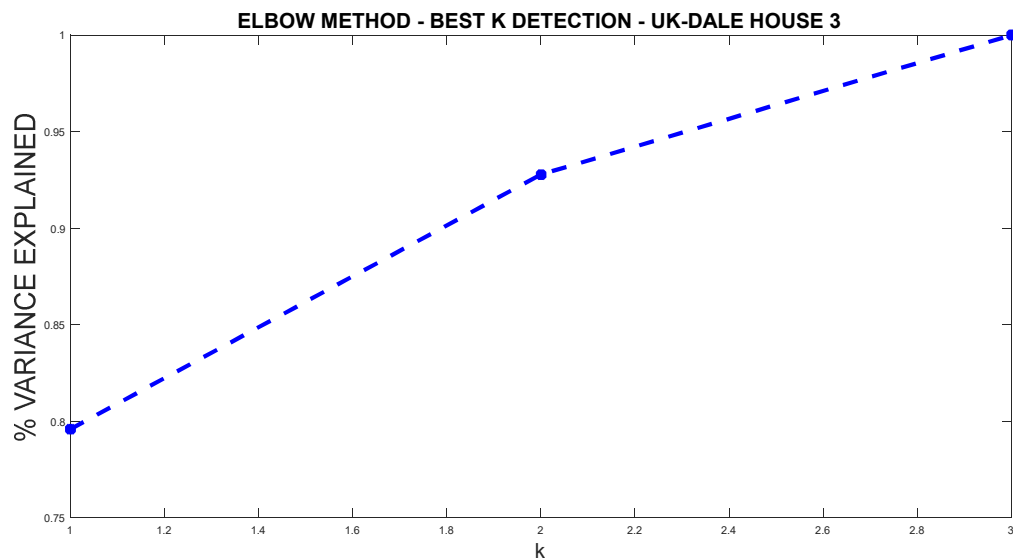


Figure 10.3: Variance Explained according to the number of clusters (k), for UK-DALE, House 3. The best k is 3.

Table 10.3: Clusters for UK-DALE House 3, $k=3$.

CLUSTER INDEX	CHANNELS
1	1 - KETTLE 2-ELECTRIC HEATER
2	3-LAPTOP
3	4-PROJECTOR

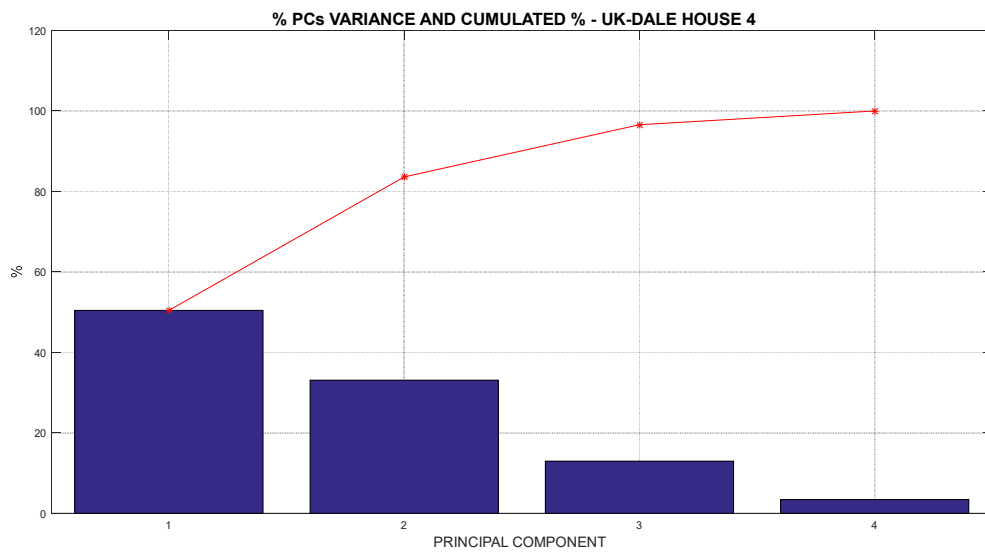


Figure 10.4: Variance Explained according to the number of clusters (k), for UK-DALE, House 4. The best k is 5.

Table 10.4: Clusters for UK-DALE House 4, $k=3$

CLUSTER INDEX	CHANNELS
1	1 – TV/DVD/DIGIBOX/LAMP 2-KETTLE/RADIO 3-GAS BOILER
2	5-WASHING MACHINE/MICROWAVE/BREADMAKER
3	4-FREEZER

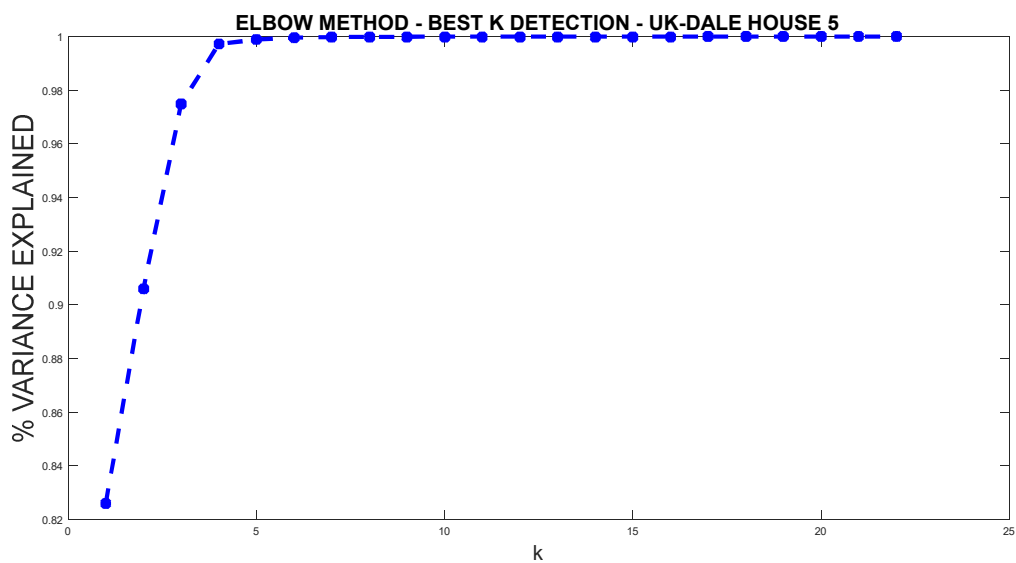


Figure 10.5: Variance Explained according to the number of clusters (k), for UK-DALE, House 5. The best k is 4.

Table 10.5: Clusters for UK-DALE House 5, $k=4$. The groups with “+” are considered as a single point

CLUSTER INDEX	CHANNELS
1	1 – STEREO SPEAKERS BEDROOM + 2-HAIRDRYER + 10- PS4 + 11- STEAM IROM + 12-NESPRESSO PIXIE + 14- TOASTER 17- KETTLE + 20- ELECTRIC HUB 6- TREADMILL 19- OVEN 21- DISHWASHER 23- WASHER DRYER 5- 24 INCH LCD BEDROOM 4- PRIMARY TV
2	2-i7 DESKTOP 7- NETWORK ATTACHED STORAGE + 8-CORE2 SERVER + 13-ATOM PC + 15-HOME THEATER AMP + 16 – SKY HD BOX
3	22- MICROWAVE
4	18- FRIDGE FREEZER 19- 24 INCH LCD

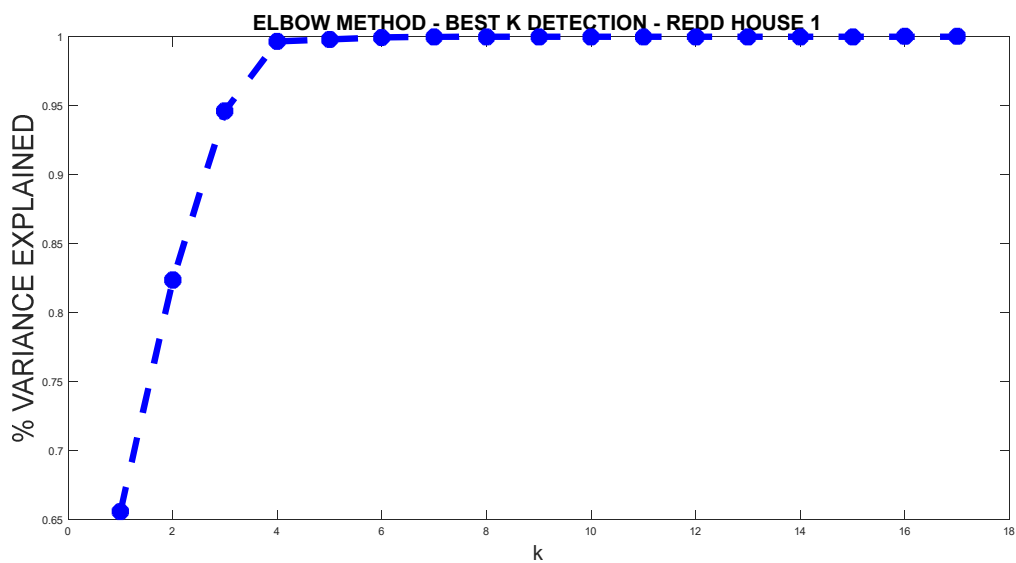


Figure 10.6: Variance Explained according to the number of clusters (k), for REDD, House 1. The best k is 4.

Table 10.6: Clusters for REDD House 1, $k=4$.

CLUSTER INDEX	CHANNELS
1	5-KITCHEN OUTLETS 1 6-KITCHEN OUTLETS 2
2	1-OVEN 1 2-OVEN 2 3-REFRIGERATOR 4-DISHWASHER 8-WASHER DRYER 1 9-MICROWAVE 10-BATHROOM GFI 11-ELECTRIC HEAT 12-STOVE 13-KITCHEN OUTLETS 1 14-KITCHEN OUTLETS 2 17-WASHER DRYER 2 18-WASHER DRYER 3
3	7-LIGHTING
4	15-LIGHTING 2 16-LIGHTING 3

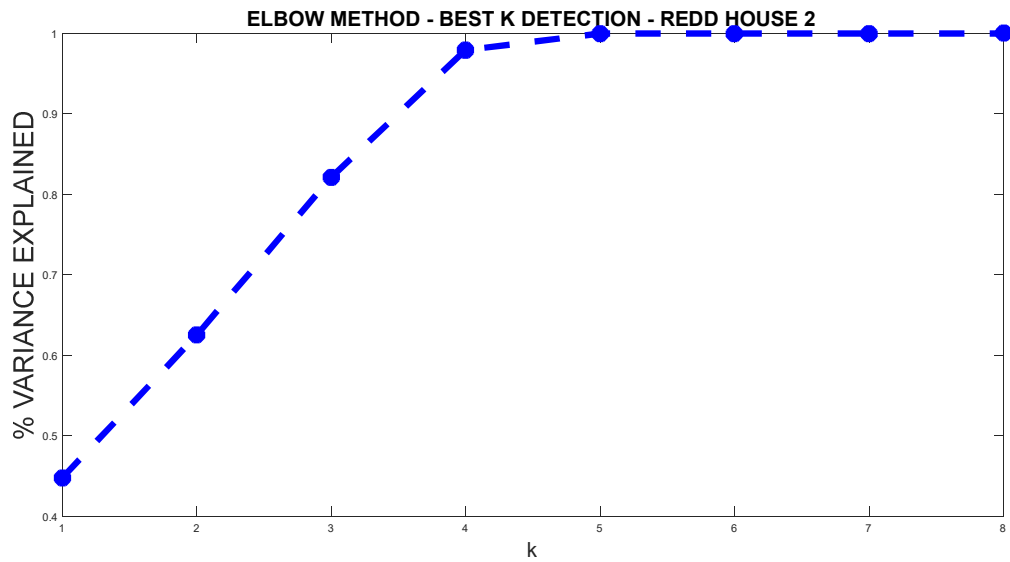


Figure 10.7: Variance Explained according to the number of clusters (k), for REDD, House 2. The best k is 5.

Table 10.7: Clusters for REDD House 2, $k=5$.

CLUSTER INDEX	CHANNELS
1	4-MICROWAVE
	5-WASHER DRYER
	6-KITCHEN OUTLETS
	8-DISHWASHER
	9-DISPOSAL
2	3-STOVE
3	1-KITCHEN OUTLETS
4	2-LIGHTING
5	7-REFRIGERATOR

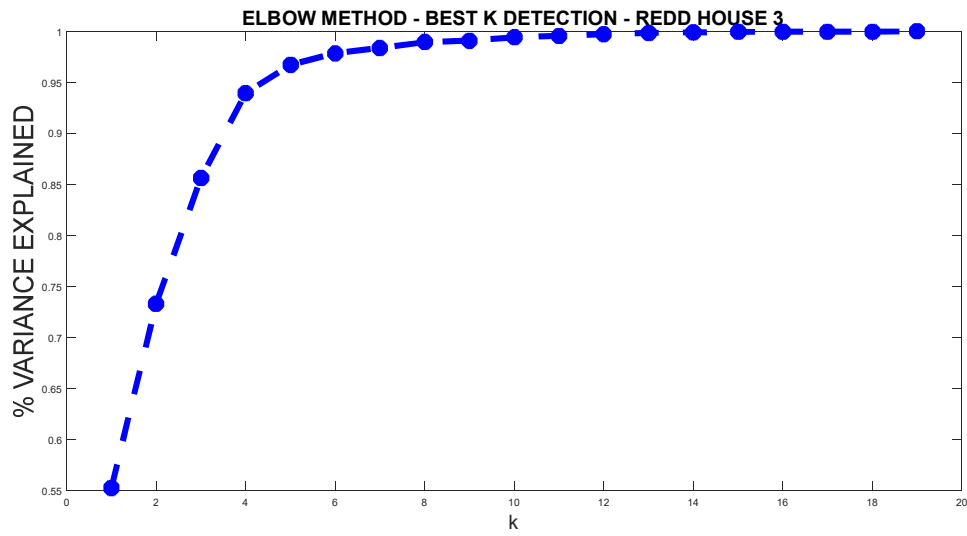


Figure 10.8: Variance Explained according to the number of clusters (k), for REDD, House 3. The best k is 5.

Table 10.8: Clusters for REDD House 3, $k=5$.

CLUSTER INDEX	CHANNELS
1	9-LIGHTING 5 10-OUTLETS UNKNOWN 13-LIGHTING 2 15-LIGHTING 3 17- LIGHTING 4 19-KITCHEN OUTLETS 1 20-KITCHEN OUTLETS 2
2	2-OUTLETS UNKNOWN 3-LIGHTING 1 6-DISPOSAL 7-DISHWASHER 8-FURANCE 11-WASHER DRYER 1 12-WASHER DRYER 2 14-MICROWAVE 18-BATHROOM GFI
3	4-ELECTRONICS 16-SMOKE ALARMS
4	5-REFRIGERATOR
5	1-OUTLETS UNKNOWN

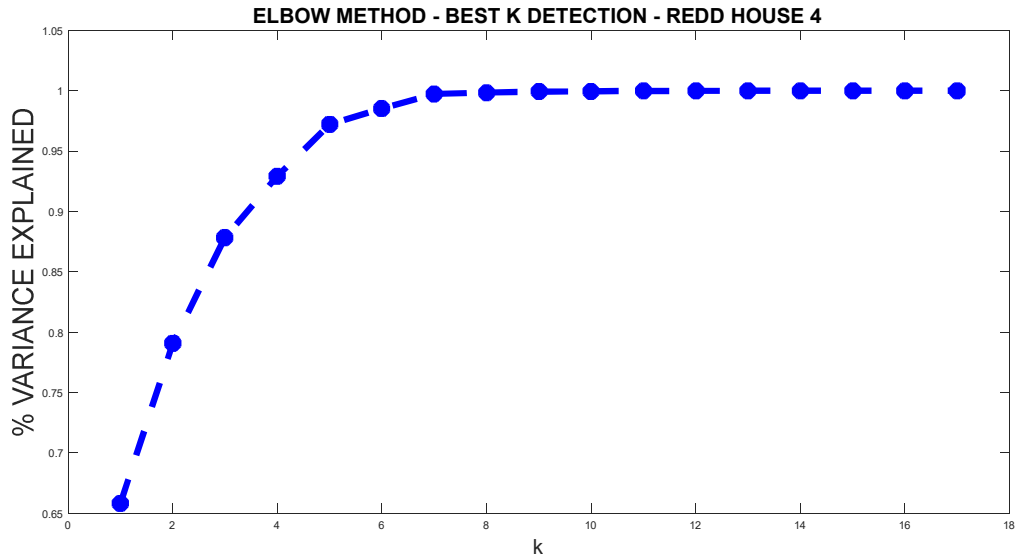


Figure 10.9: Variance Explained according to the number of clusters (k), for REDD, House 4. The best k is 5.

Table 10.9: Clusters for REDD House 4, $k=5$. The groups with “+” are considered as a single point.

CLUSTER INDEX	CHANNELS
1	14-BATHROOM GFI
2	1-LIGHTING 1 3-KITCHEN OUTLETS 10-SMOKE ALARMS 16-LIGHTING 3
3	2-FURANCE 11-LIGHTING 2
4	4-OUTLETS UNKNOWN 5-WASHER DRYER 6-STOVE 7-AIR CONDITIONING 8-AIR CONDITIONING 9-MISCELLANEOUS + 15-BATHROOM GFI 13-DISHWASHER 17-LIGHTING 4 18-AIR CONDITIONING
5	12-KITCHEN OUTLETS

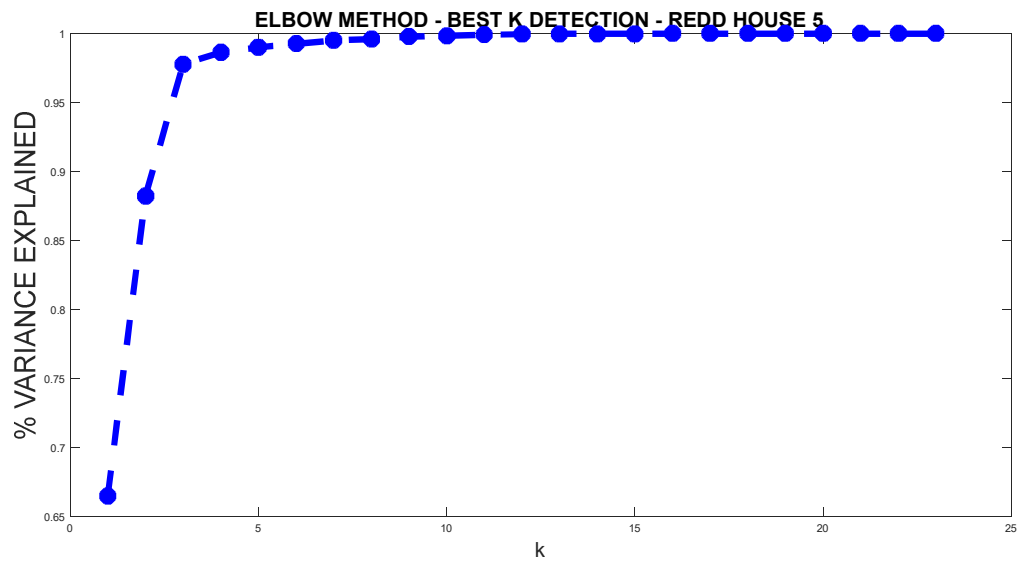


Figure 10.10: Variance Explained according to the number of clusters (k), for REDD, House 5. The best k is 4.

Table 10.10: Clusters for REDD House 5, $k=4$. . The groups with “+” are considered as a single point.

CLUSTER INDEX	CHANNELS
1	1-MICROWAVE 2-LIGHTING 1 3-OUTLETS UNKNOWN 1 7-WASHER DRYER 9-SUBPANEL 10-ELECTRIC HEAT 1 11-ELECTRIC HEAT 2 + 14-BATHROOM GFI + 15-LIGHTING 3 13-OUTLETS UNKNOWN 2 17-LIGHTING 4 18-DISHWASHER 19-DISPOSAL 20-ELECTRONICS 21-LIGHTING 5 22-KITCHEN OUTLETS 1 + 24-OUTDOOR OUTLETS 23-KITCHEN OUTLETS 2
2	12-LIGHTING 2
3	4-FURANCE 5-OUTLETS UNKNOWN 6-WASHER DRYER 8-SUBPANEL
4	16-REFRIGERATOR

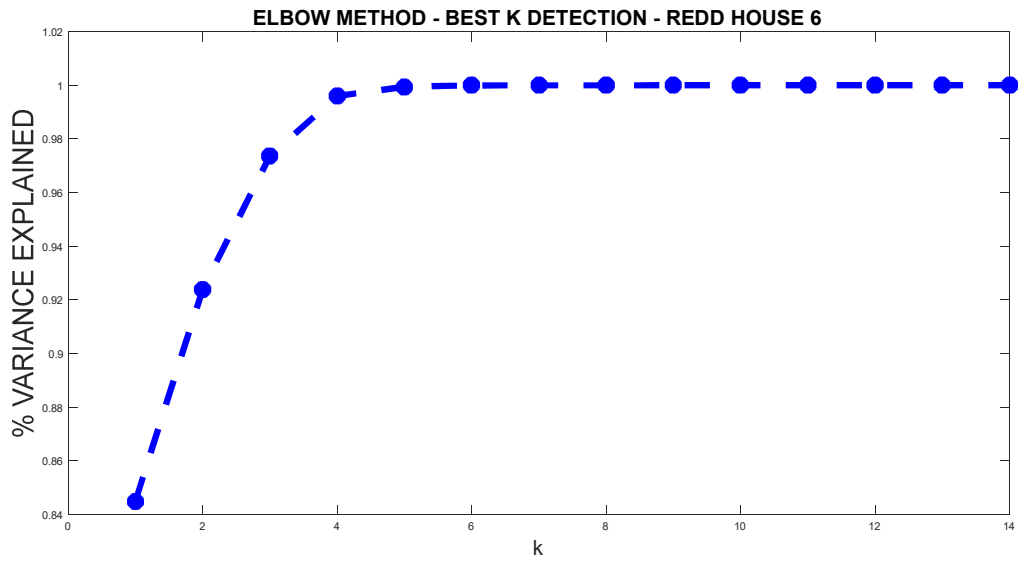


Figure 10.11: Variance Explained according to the number of clusters (k), for REDD, House 6. The best k is 3.

Table 10.11: Clusters for REDD House 6, k=3.

CLUSTER INDEX	CHANNELS
1	3-STOVE 4-ELECTRONICS 7-DISHWASHER 8-OUTLETS UNKNOWN 9-OUTLETS UNKNOWN 13-AIR CONDITIONING 14-AIR CONDITIONING 2 15-AIR CONDITIONING 3
2	1-KITCHEN OUTLETS + 2-WASHER DRYER + 10-ELECTRIC HEAT + 12-LIGHTING 5-BATHROOM GFI 11-KITCHEN OUTLETS
3	6-REFRIGERATOR

11. APPENDIX D – MINIMUM SPANNING TREE RESULTS

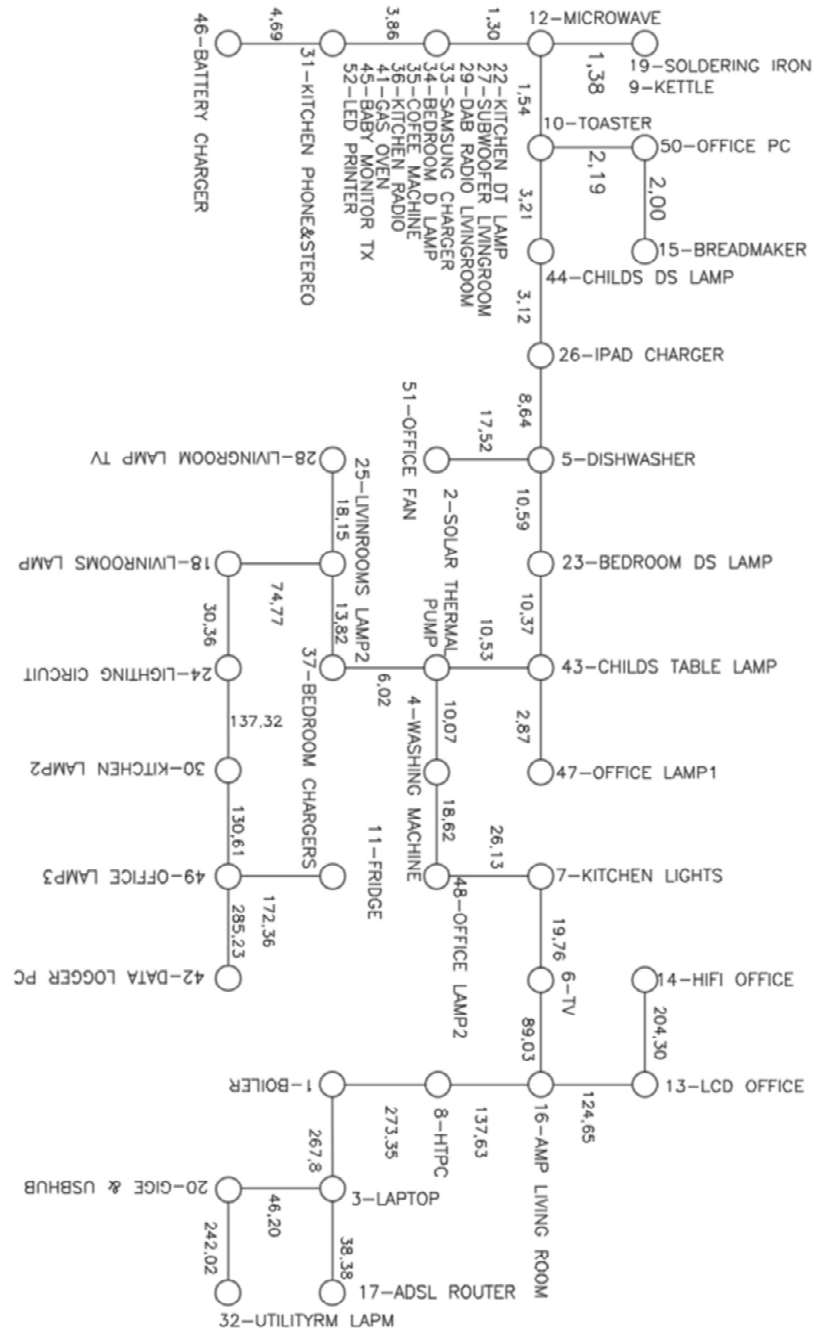


Figure 11.1: MST for UK-DALE, House 1, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.

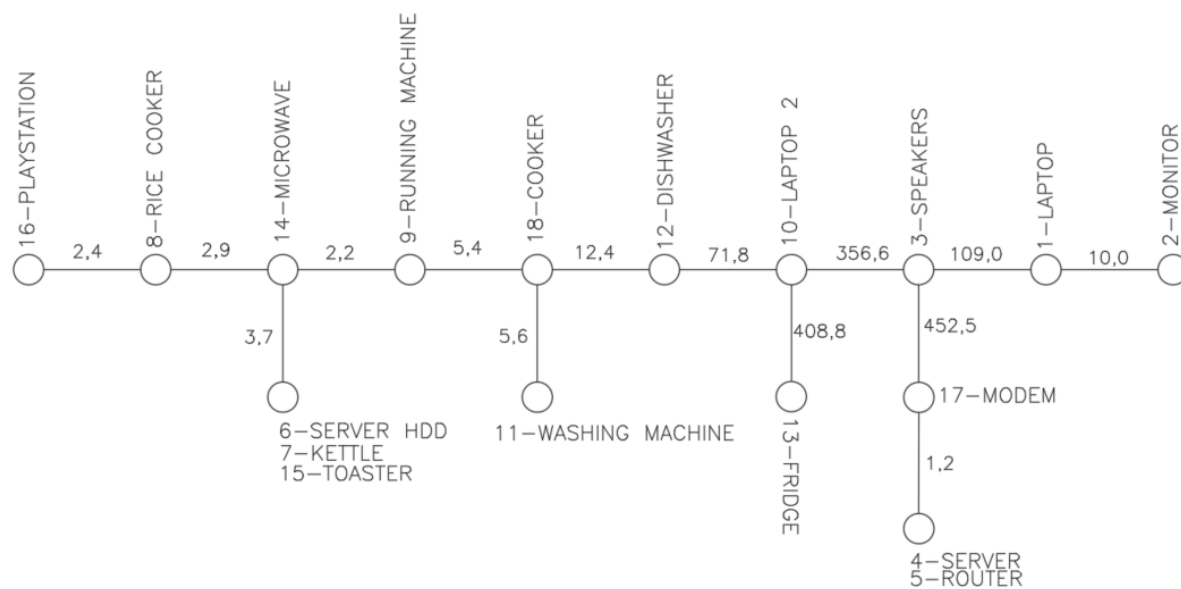


Figure 11.2: MST for UK-DALE, House 2, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.

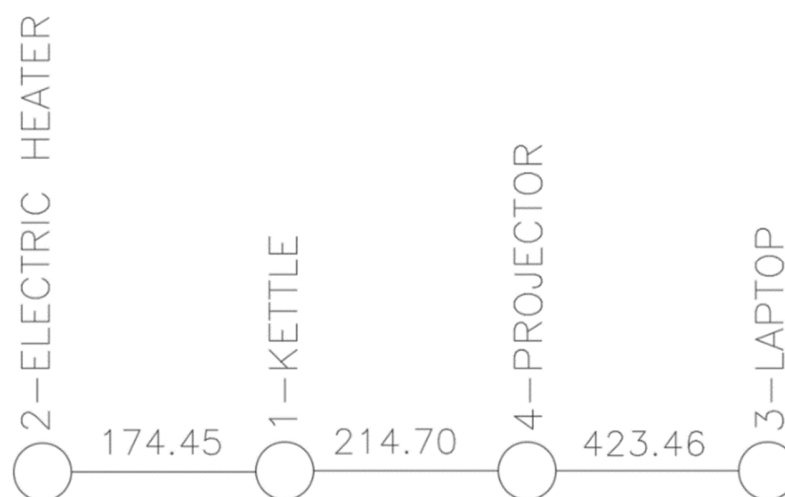


Figure 11.3: MST for UK-DALE, House 3, including the distances between nodes.

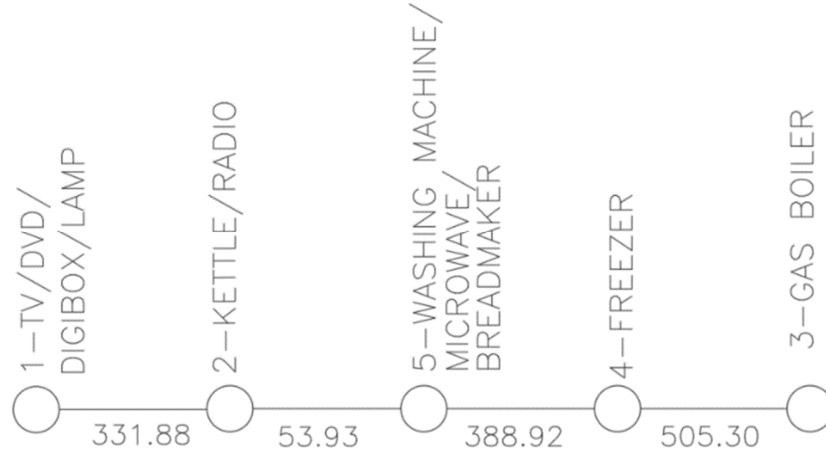


Figure 11.4: MST for UK-DALE, House 4, including the distances between nodes.

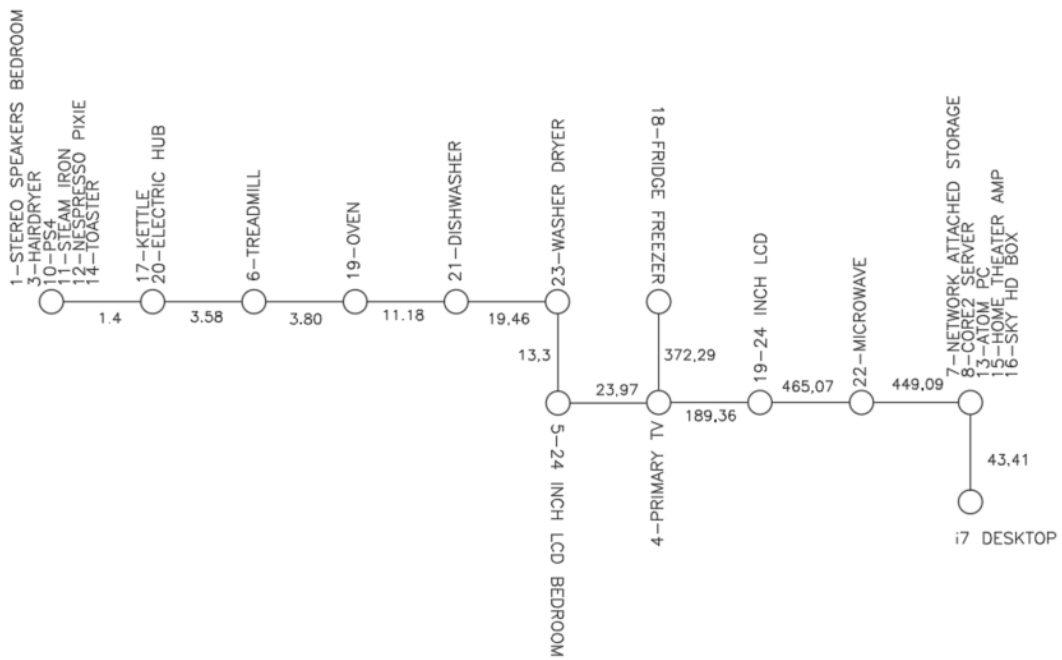


Figure 11.5: MST for UK-DALE, House 5, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.

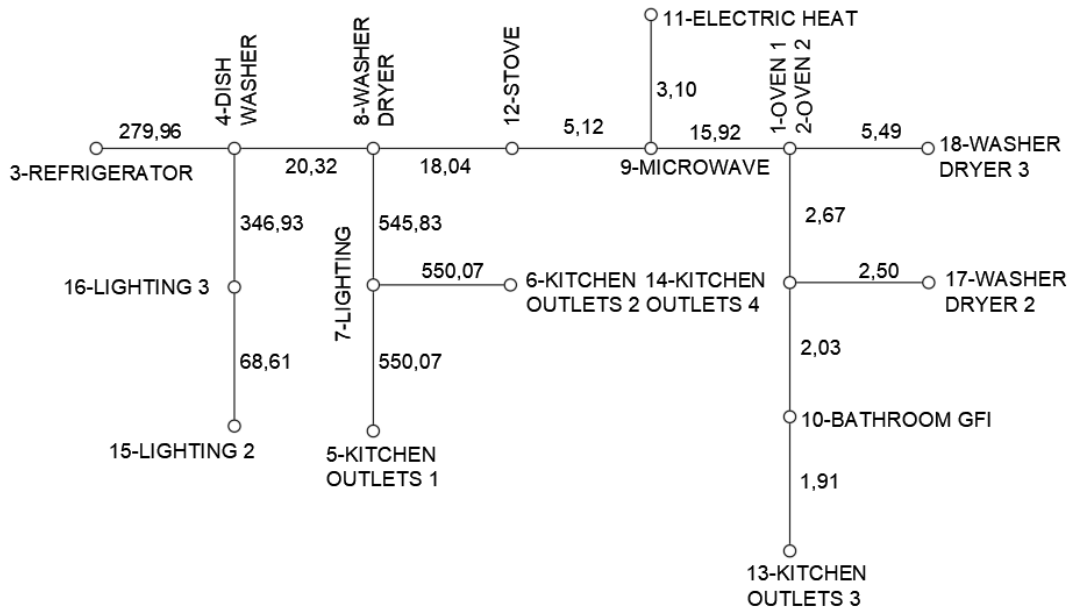


Figure 11.6: MST for REDD, House 1, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.

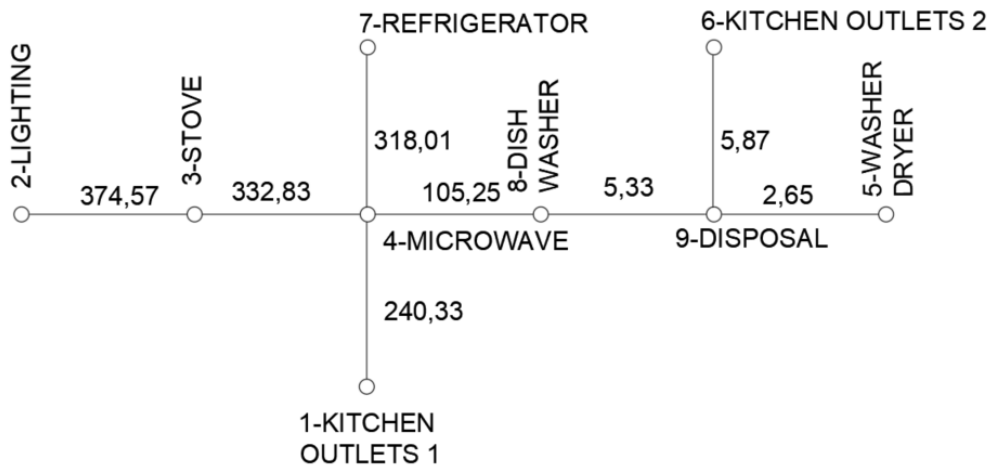


Figure 11.7: MST for REDD, House 2, including the distances between nodes.

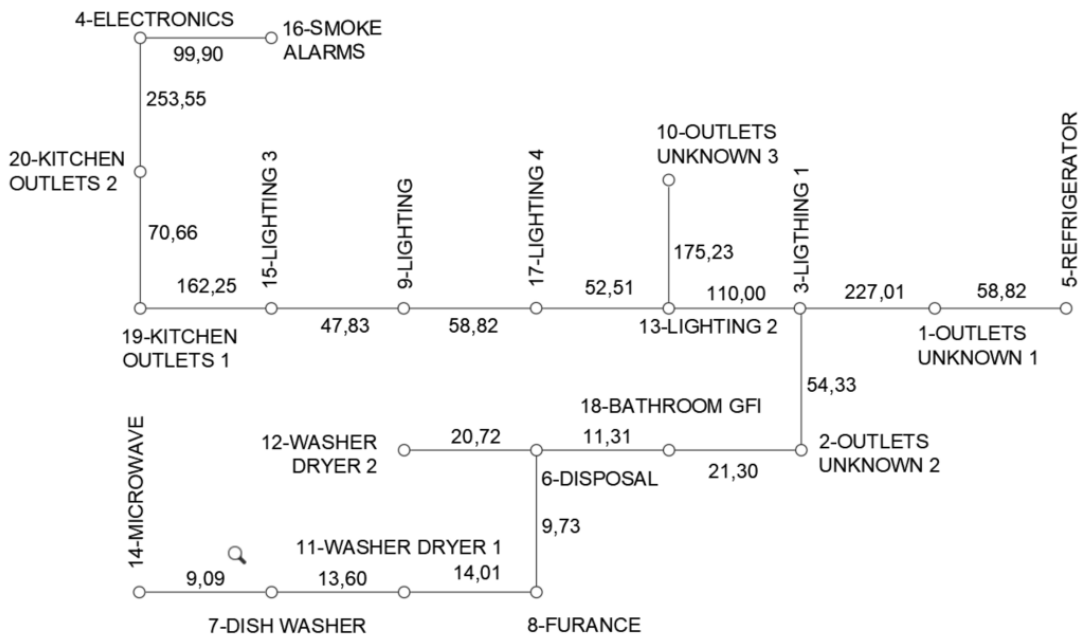


Figure 11.8: MST for REDD, House 3, including the distances between nodes.

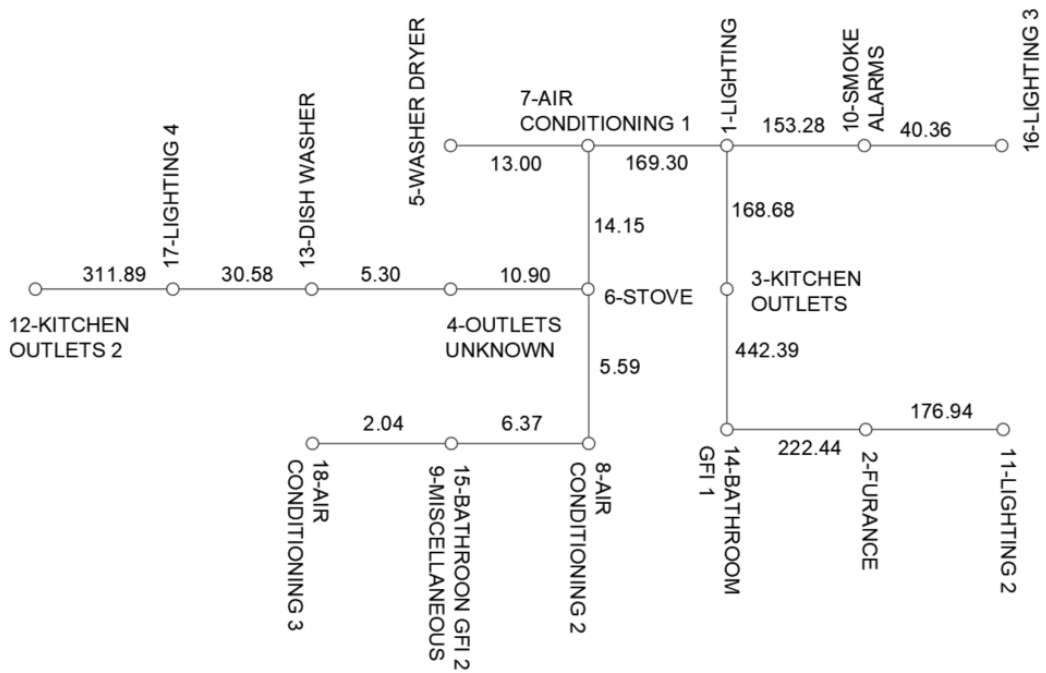


Figure 11.9: MST for REDD, House 4, including the distances between nodes.

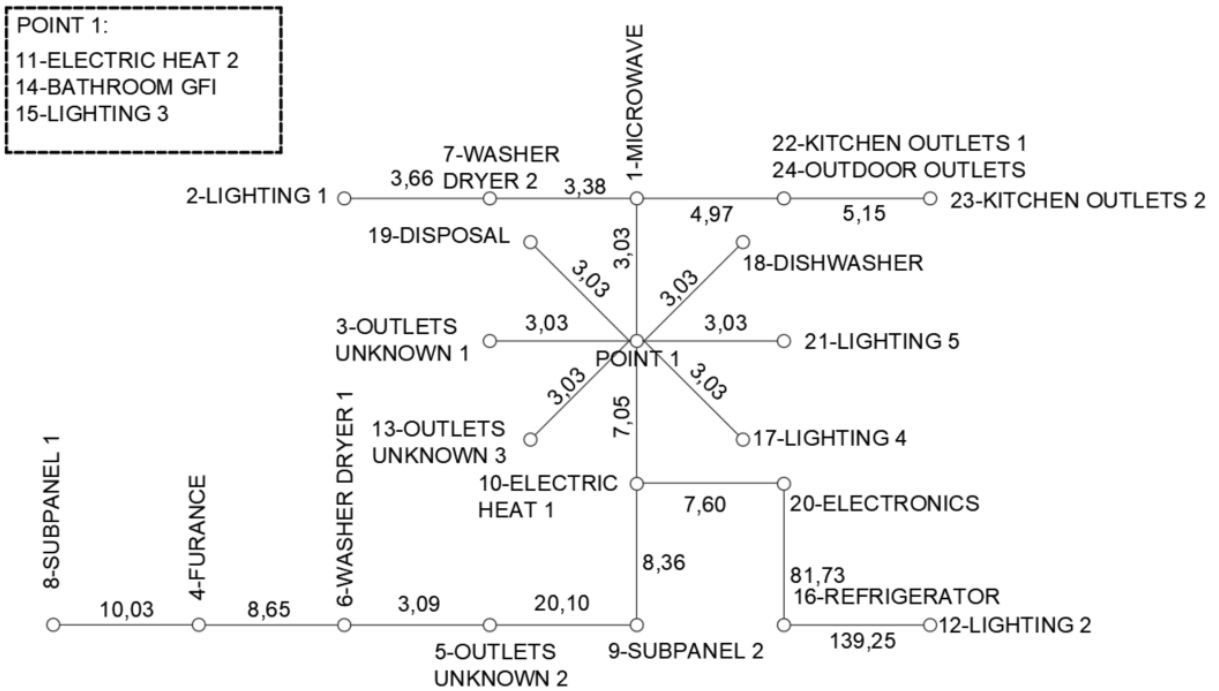


Figure 11.10: MST for REDD, House 5, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.

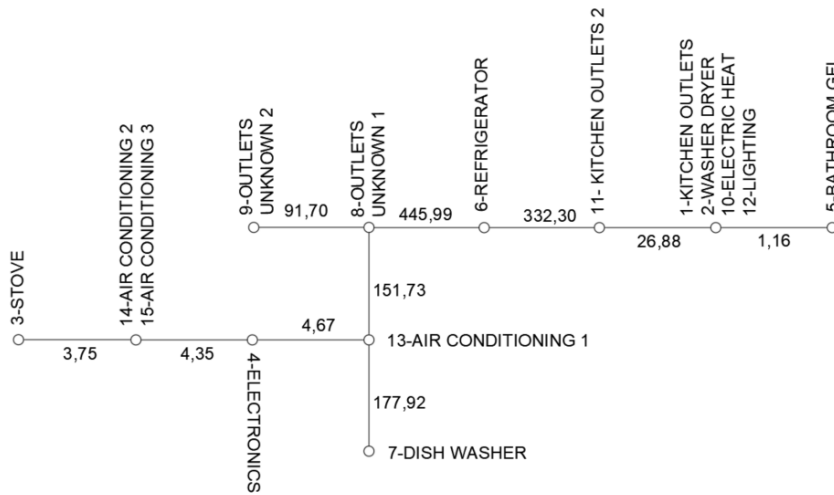


Figure 11.11: MST for REDD, House 6, including the distances between nodes. The appliances with distances smaller than 1 was considered as a single node.

12. APPENDIX E – K-MEANS VISUALIZATION OVER THE MST

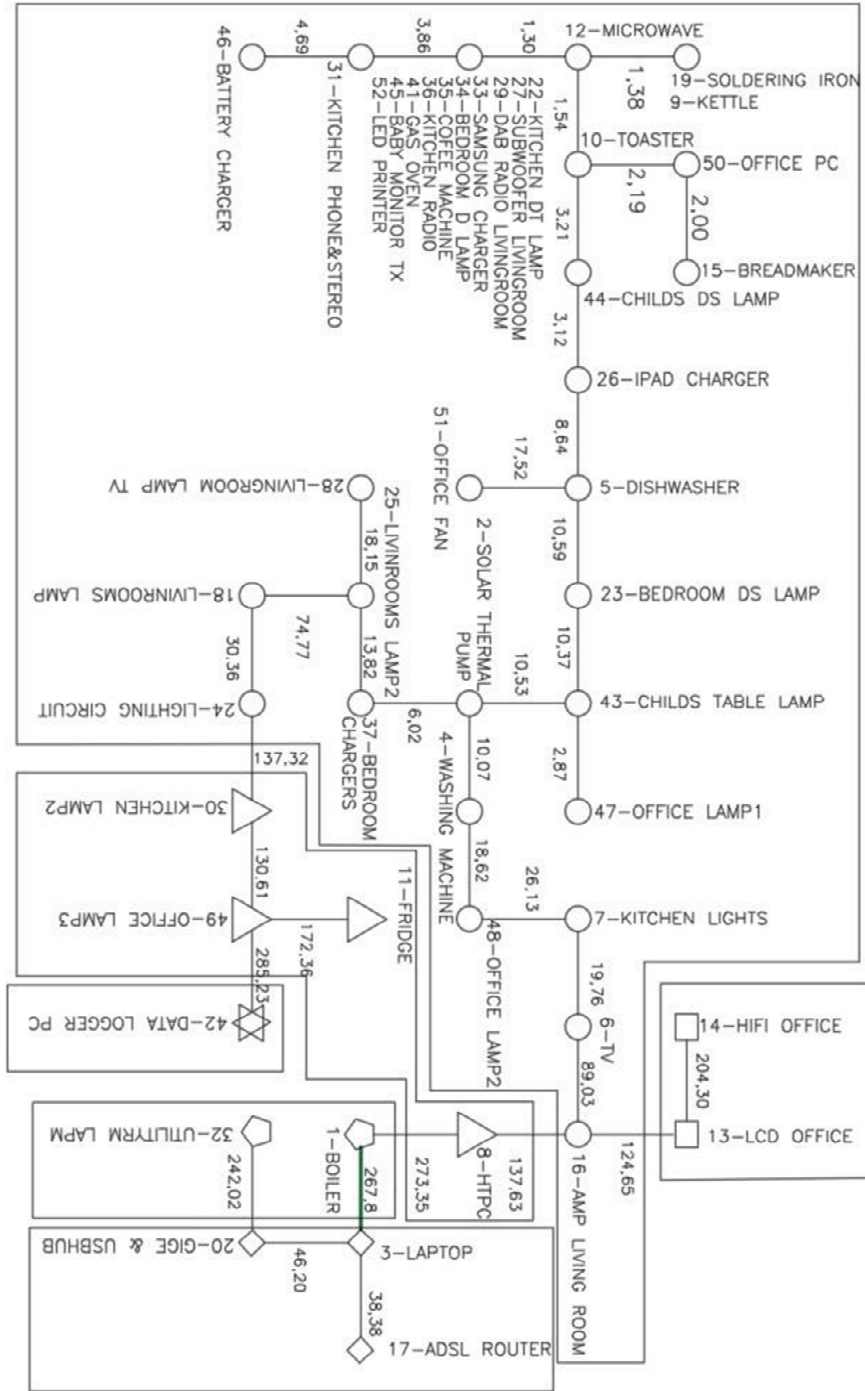


Figure 12.1: MST with k-means results represented over the MST for UK-DALE, House 1, with k=6.

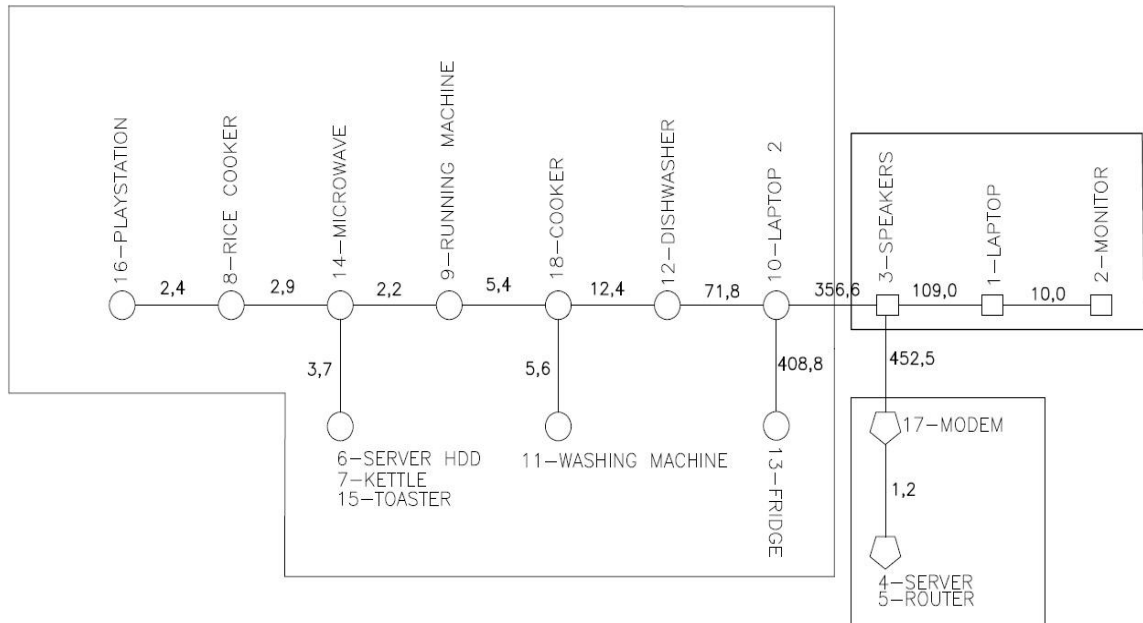


Figure 12.2: MST with k -means results represented over the MST for UK-DALE, House 2, with $k=3$.



Figure 12.3: MST with k -means results represented over the MST for UK-DALE, House 3, with $k=3$.

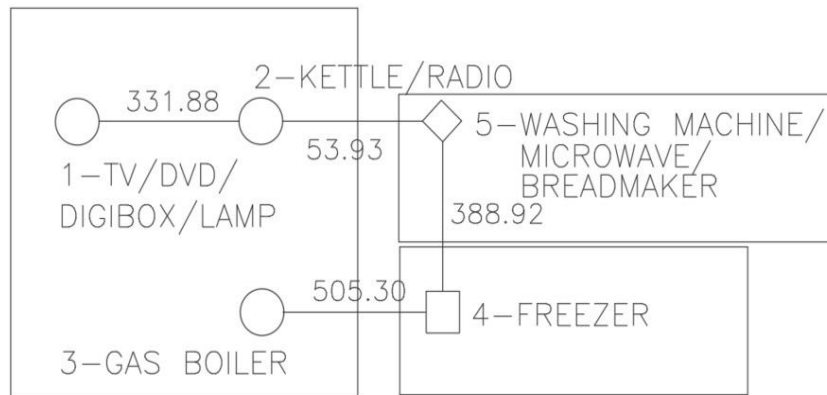


Figure 12.4: MST with k -means results represented over the MST for UK-DALE, House 5, with $k=3$.

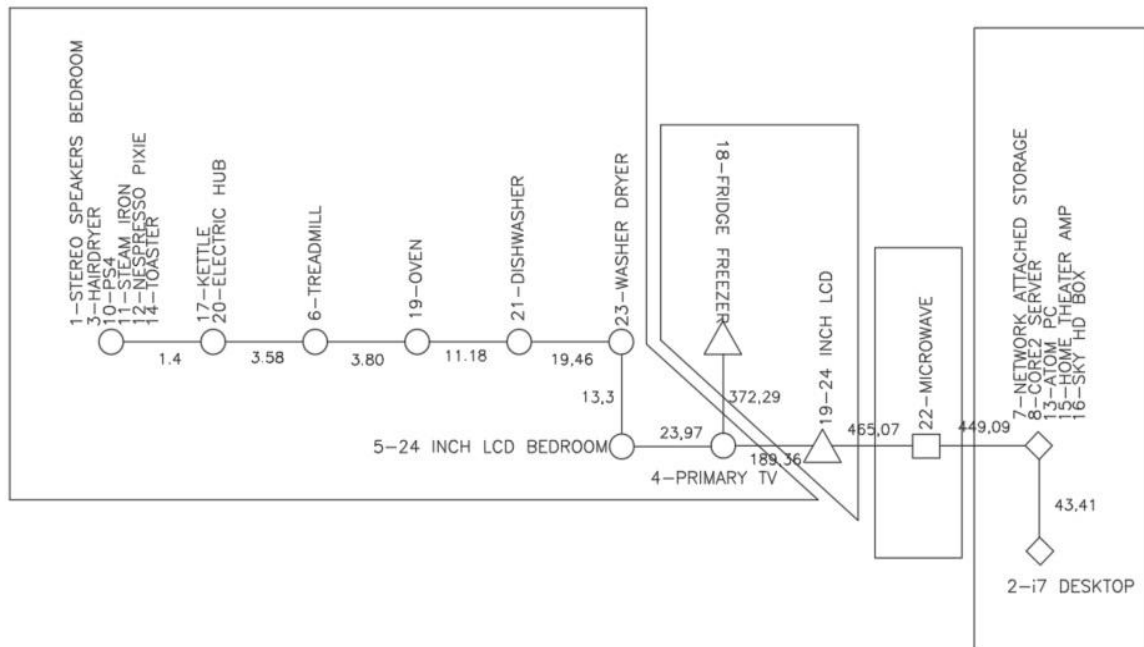


Figure 12.5: MST with k -means results represented over the MST for UK-DALE, House 5, with $k=4$.

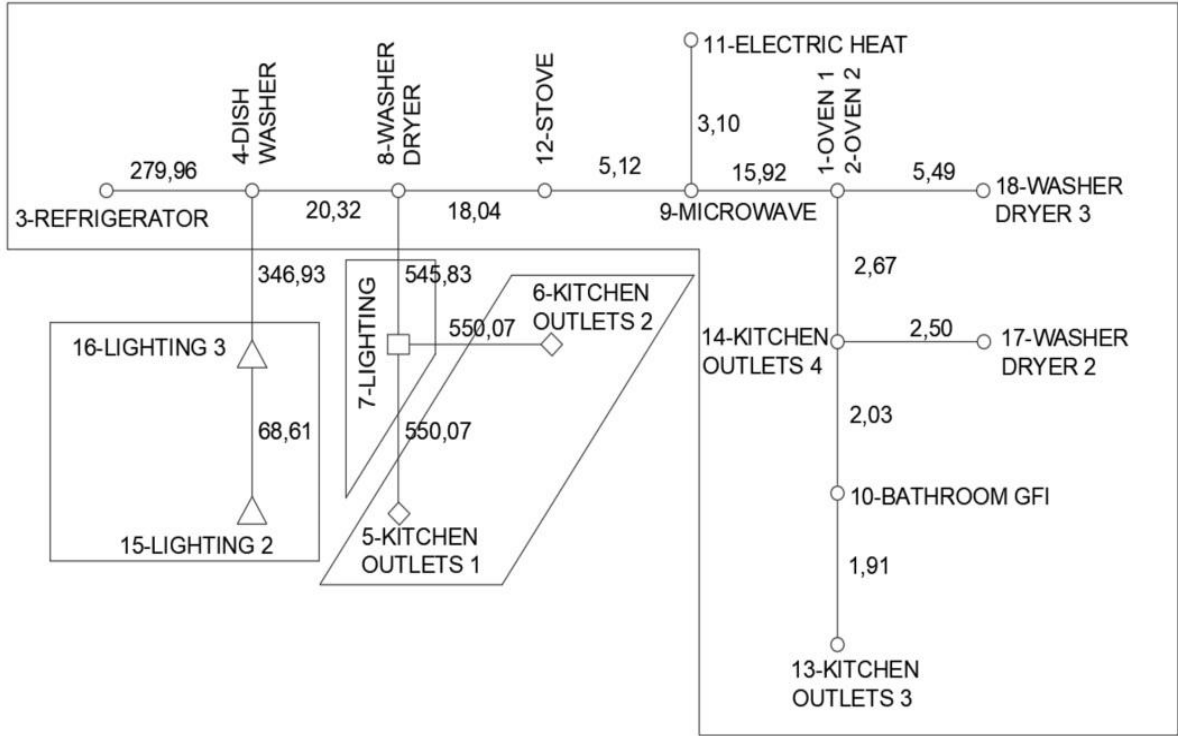


Figure 12.6: MST with k -means results represented over the MST for REDD, House 1, with $k=4$

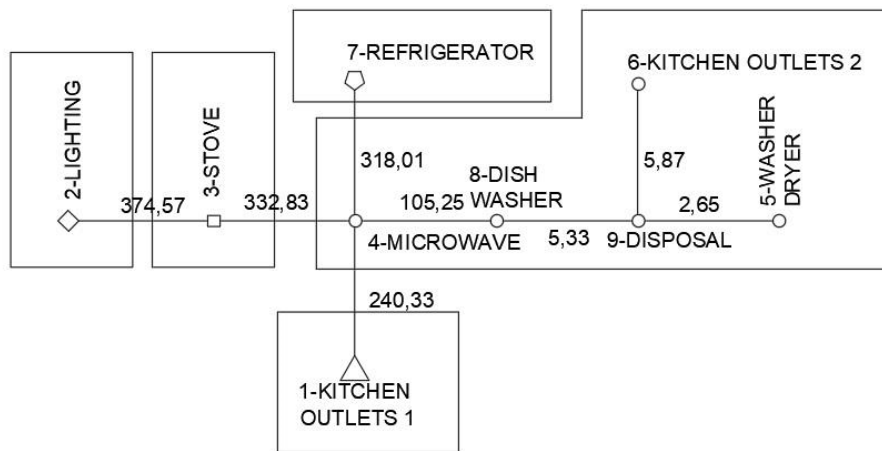


Figure 12.7: MST with k -means results represented over the MST for REDD, House 2, with $k=5$

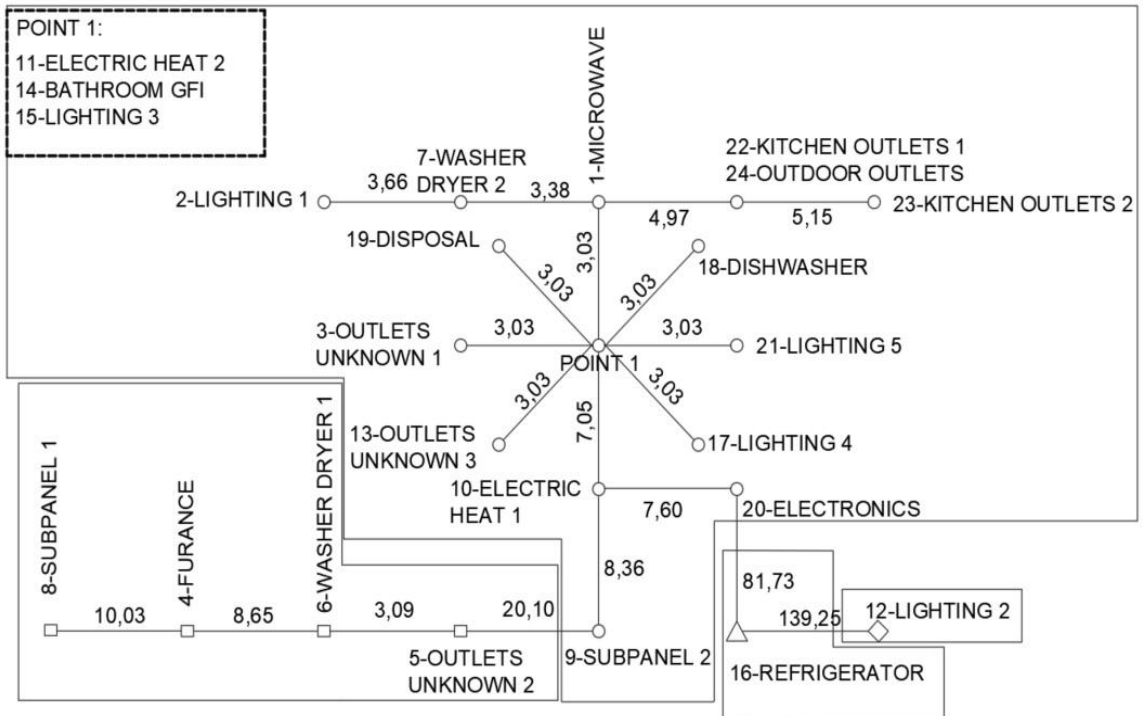


Figure 12.10: MST with k-means results represented over the MST for REDD, House 5, with k=4

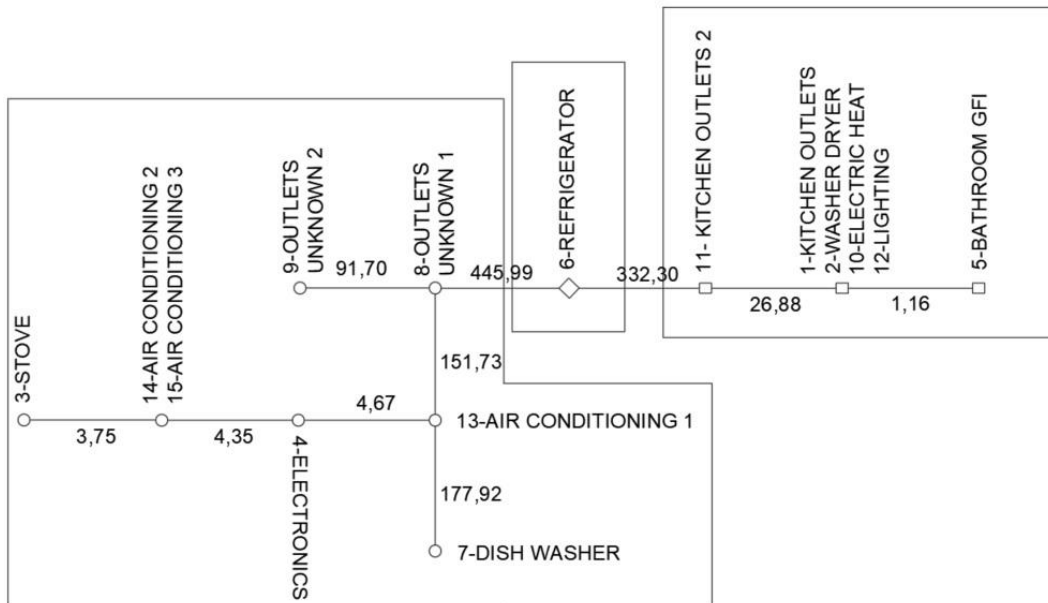


Figure 12.11: MST with k-means results represented over the MST for REDD, House 6, with k=3