Dent, Ian (2015) Deriving knowledge of household behaviour from domestic electricity usage metering. PhD thesis, University of Nottingham.

Deriving knowledge of household behaviour from domestic electricity usage metering

Ian Dent, B.Sc. (Hons)

Thesis submitted to
The University of Nottingham
for the degree of Doctor of Philosophy

July 2015

Dedicated to my father, Derrick Dent
for stressing the importance of education

Abstract

The electricity market in the UK is undergoing dramatic changes and requires a transformation of existing practices to meet the current and forthcoming challenges. One aspect of the solution is the deployment of demand side management (DSM) programmes to influence domestic behaviour patterns for the benefit of the overall network. Effective deployment of DSM requires segmentation of the population into a small number of groupings.

Using a database of electricity meter data collected at a frequency of five minutes over a year from several hundred houses, households are clustered based on the shape of the average daily electricity usage profile. A novel method, incorporating evaluation criteria beyond compactness, of evaluating the resulting groupings is defined and tested. The results indicate the potentially most useful algorithms for use with load profile clustering.

Patterns within the electricity meter data are approximated and symbolised to allow motifs (representing repeated behaviours) to be identified. Uninteresting motifs are automatically identified and discarded. The different possible parameters, including size of motif and number of symbols used in representing the data, are explored and the most appropriate values found for use with electricity meter data motif detection.

The concept of variability of regular behaviour within a household is introduced and methods of representing the variability are considered. The novel method of using variability in timing of motifs is compared to other techniques and the results tested using the previously defined evaluation criteria.

Combining the generated motif data with the meter data to produce a single set of archetypes does not produce more useful results for use with DSM. However, creating complementary sets of archetypes based on each set of data, provides a more complete understanding of the households and allows for better targeting of DSM initiatives.

Acknowledgements

Thanks are due to my supervisors, Professors Uwe Aickelin and Tom Rodden for their support and advice.

Various people at the University of Nottingham have helped to make my research time an enjoyable experience. My office colleagues (Simon Miller, Jabran Aladi, Alexandros Ladas and Robert Miles) are thanked for being interesting and thoughtful companions.

Members of the Intelligent Modelling and Analysis group at the University have always been helpful in providing ideas and background on their own work as it applies to my research. Daniel Soria, in particular, has always been available for advice when needed and Peer-Olaf Siebers has always gone out of his way to include me in related projects.

I've worked closely with members of the Horizon Digital Economy Research Institute on various projects during my time at the University. James Colley and Ben Bedwell were very supportive when I was starting my research and have continued to provide assistance. Alexa Spence was of particular help in remembering my research when talking with her contacts and was directly responsible for setting up the academic links that provided suitable data to allow the research to be completed.

Outside of the University of Nottingham, various people have provided support and advice to me. The most important of these is Tony Craig from the James Hutton Institute in Aberdeen who has shared data from his NESEMP project which has allowed the research in this thesis to proceed. Tony has also collaborated with me on a number of papers and his ideas and efforts in supplying additional information to me were key to the completion of this thesis.

Other Universities working on the Desimax project have provided support and advice and, in particular, Aristides Kiprakis, Steve McLaughlin, and Adam Collin should be thanked.

People working in related areas of research, and who have provided information, include Pavel Senin, Stephen Habin, Giancarlo Chicco and Eamonn Keogh.

Richard Mayer of the University of Derby and Nick Clark were kind enough to provide references for me to be accepted onto the PhD research project and I hope the results presented in the thesis justify their faith in me.

Outside of academia, I need to thank a number of people for their support particularly my wife, Janet, for putting up with the period of commitment to the research, and the writing up, and the resulting lack of time for household maintenance, holidays, etc.

Steve Adams is gratefully thanked for his efforts in proof reading the thesis and for his regular prompting for demonstrations of progress.

Finally I'd like to thank my examiners (Professors Bogdan Gabrys and Tony Pridmore) for their conscientious and fair assessment of the thesis and their supportive ideas for improvement.

This work is possible thanks to the Research Councils Energy Program grant reference EP/I000496/1.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

The electricity market in the UK is undergoing dramatic changes. Legal, social and political drivers for a more carbon efficient electricity network, along with the dramatically increased flow of data from households through the deployment of smart meters, requires a transformation of existing practices [2]. In particular, the change of sampling of electricity usage using smart meters alters what is possible in understanding households' behaviour [3].

One approach to addressing the pressures on the electricity network is the application of Demand Side Management (DSM) techniques to achieve changes in consumer behaviour. River [4] defines DSM as "systematic utility and government activities designed to change the amount and/or timing of the customer's use of electricity" for the collective benefit of society, the utility company, and its customers.

The investigation of household electricity load profiles is an important area of research given the centrality of such patterns in directly addressing the needs of the electricity industry, both now and in the future. Being able to identify archetypical representations of households is essential for effective implementation of demand side management (DSM) techniques which, itself, is necessary to allow the electricity industry to meet the upcoming challenges [5]. The ability to produce more useful archetypes than the industry currently uses can lead to more effective behaviour modification interventions.

To allow selection of appropriate DSM interventions, a good understanding of the existing behaviour of households is required. Firstly, knowledge is needed on an individual household that can be deduced from house-wide electricity metering. Secondly, a method is required to group large numbers of households into a manageable number of archetypes where the members display similar behaviours. This approach allows for cost effective targeting of the most appropriate subset of customers whilst allowing the company management to deal with a manageable number of archetypes [6].

One behavioural trait to use as the basis of clustering is the variability that households exhibit in their daily household activity patterns [7]. For instance, some households will be creatures of habit and will eat their evening meal at almost the same time each evening, whilst others have a much more variable activity pattern and will eat at different times. Clustering households using their degree of variability, as shown by electricity consumption, provides a way of identifying the subset of electricity users who may be most receptive to an intervention intending to influence their activity patterns. It is worth noting that this intervention may be to reward households for NOT changing their current pattern of usage if it is already as desired by the utility company.

The National Grid has particular problems, as described in Section 2.3, that can only be addressed by influencing user's behaviour to change their electricity usage patterns [8]. The classification of users into manageable groups will allow realistic targeting of appropriate DSM techniques to the most important (from an effect on the overall supply point of view) consumers.

1.2 Motivation and Research Questions

The research aim is to explore the limits of the information that can be derived from simple monitoring of the household electricity meter as will be provided by the forthcoming roll-out of smart meters [9]. This thesis investigates knowledge that can be derived from the overall shape of the usage pattern as well as analysis of how regular behaviours may vary from day to day. The goal is to define a small number of understandable groupings of households that demonstrate similar electricity usage behaviour and to which specific DSM techniques could be addressed.

The overall research question is "Can variability in behaviour within a household be identified by finding motifs within UK electricity meter data and can these motifs then be used for clustering households into a few archetypes?"

This goal raises a number of questions that the research will address.

1.2.1 Can clusters of households be found using UK electricity meter data?

This question leads to a number of sub-questions:

- Can a method of evaluating load profile clustering be defined that incorporates aspects of the effectiveness of the results for use in DSM?
- Does the sampling frequency affect the quality of the archetypes found?
- Which clustering algorithm provides the best clustering results, as defined by the evaluation method and using UK data?

This work establishes a baseline (prior to the incorporation of variability) of how clusters can be found within the data and defines a new way of evaluating the results. The questions are considered in Chapter 4 which includes:

- A description and demonstration of the current industry practice of clustering households based on average load shape.
- A description of the approach of using compactness measures to evaluate different clustering algorithms.
- The introduction of a new composite validity index that measures more aspects of the clustering results than just the compactness.

- A comparison between the composite measure and the compactness measure and an evaluation of how the different measures give different results regarding the ranking of clustering algorithms.
- An evaluation, using the composite measure, of data sampled at five minute and 60 minute intervals to test the effect of sampling frequency.

1.2.2 Is it possible to find a reasonable number of interesting motifs within the electricity meter data?

A sub question arising from this work is "Is it possible to define a set of parameters to use for the chosen motif finding technique to provide a reasonable number of motifs?"

This work establishes a method to find motifs and evaluate between different parameters and is detailed in Chapter 5 which includes:

- The definition of a motif finding approach and consideration of the possible parameter settings.
- The identification of simple non-interesting motifs and their exclusion from the analysis.
- The definition of a qualitative evaluation method to allow choosing between the differing sets of parameters. Parameter settings are particular to the problem of motifs within electricity meter data although the approach would be valid for other application domains.
- The identification of the set of parameters to use for the motif finding technique to provide a reasonable number of motifs.

1.2.3 Does extending the household attributes to include measures of variability alter the archetypical clusters obtained?

This work is detailed in Chapter 6 and addresses the overall thesis question, building on the work in previous chapters and covering the following:

- Introduction of the concept of variability of behaviour in electricity meter data.
- Determination of ways of measuring variability including the timing of short, repeated patterns (motifs).
- Clustering using various algorithms and application of a composite measure to assess the quality of the resulting clusters.
- Comparing the variability results with those from load profile clustering.
- Evaluation, using the composite measure, between combinations of original usage data, attributes based on motif variability, and attributes based on motif frequency.
- Comparing clustering on variability with clustering using the base dataset.

Differences between an archetypical household profile and that of other archetypes can be used to suggest energy usage behaviour changes to reduce overall electricity usage or to improve electrical efficiencies, possibly by time shifting particular activities. In addition, particular groups (e.g. large users during peak times) can be identified for targeting for reduction initiatives.

Interested parties to the results of the research include electricity suppliers, householders, regulators and electricity generators. The use each stakeholder could make of the research is summarised in Table 1.1.

1.3 Scope of Work and Limitations

While other studies have extensively analysed the data collected from multiple monitors (both electricity and other) within a household (e.g. Jiang et al. [10]) this research concentrates on exploring the information that can be extracted from a single monitor (at the household level) sampling at the scale of minutes.

Table 1.1: Use made of research by interested parties

Stakeholder	Possible Usage
Energy Suppliers	Better targeting of consumers for marketing and for billing or tariff offers
	Measuring effectiveness of DSM initiatives
Householders	Feedback on progress against tariffs with caps or payment tiers or on how much "greenness" households exhibit
	Comparison with neighbours or similar type of people (same house or family size)
	Feedback on when regular tasks (e.g. cooking) are undertaken
Regulators	Better understanding of usage profiles
Energy Generators	More accurate aggregation of total demand

The data predominately used for the analysis is from the 2011 and 2012 period and data from alternative periods may produce different results due to, say, different weather patterns or societal behaviour changes. For example, the UK extensively used storage heaters around 1970 and figures on electricity usage from that period will provide very different patterns of behaviour.

This thesis and much of the other published work in the area of load profiles focuses on the electrical usage information provided by an electricity meter or by a usage monitor (connected to the electricity supply). This measures the real power consumed within the household. When the appliances in the house use electric motors (e.g. washing machines) there is also a "reactive power" used. This is not measured by the electricity meter but does cause a load on the utility company as it contributes to warming of the electricity network connections. Large industrial users making use of large motors for industrial processes generally have a meter installed that measures both real power and reactive power. However, this isn't common in domestic households. Collin et al. [11] contains a more detailed analysis of the different power measurements that utilities are concerned about and which different appliances generate. While these enhancements are not considered in this thesis, the approach to analysing the electricity usage can equally be applied to other power measurements if data were available (rare for domestic households).

Limitations of the work include:

- While the approach is valid for any geographic region, the data used for analysis is from part of Scotland only. The assumptions made for "reasonable" settings for parameters are driven by knowledge of the UK situation and may not be valid elsewhere.

- The sampling rate of the meter readings is five minutes and the results and approach are valid for similar scale sampling rates only. Other work has explored very frequent sampling (e.g. millisecond) [12] and different information can be found.

- The work focuses on domestic households only and does not consider non-household domestic electricity usage (e.g. street lighting) or industrial usage.

- The load profiling clustering and motif finding algorithms have not been assessed for computational efficiency and, while usable within the scale of the data used in this study, may need modification to enable scaling to usage across much higher populations of households.

1.4 Research Contribution

At its core, Computer Science is about representing, transforming, analysing and distributing information. This information can come from very different sources and mean different things [13]. This thesis addresses the issues of analysing data arising from electricity meters and using techniques of transformation, searching, pattern identification and analysis to draw conclusions about the behaviour of the members of the households generating the data.

The genre of electricity load profile clustering is well established within the fields of Computer Science and Electrical Engineering and, as it is concerned with collecting information and then transforming, analysing and interpreting the information, is firmly part of Computer Science.

The research contribution of this work is:

- Extending the currently available cluster validity indexes for comparing clustering techniques with particular applicability to electricity usage profiles by the definition of an extendible composite validity index. This is described in Chapter 4.
- Defining archetypes of domestic users incorporating information beyond the shape of usage throughout the day.
- Providing a ranking of clustering algorithms for load profile clustering using data drawn from a region of the UK. This work offers guidance to other researchers using clustering of electricity load profiles and is described in Chapter 4.
- The application of a symbolisation approach to finding motifs that has not previously been researched using electricity data. This work is detailed in Chapter 5.
- The development of a qualitative method of evaluation for selection between different motif finding options. This work is detailed in Chapter 5.
- The definition of set of parameters to use for finding motifs which can be useful to other researchers in similar application areas. This work is detailed in Chapter 5.
- The application of the concept of "variability of behaviour" within electricity load profiling to determine useful clusters of households showing similar amounts of variability in their behaviours. This work is detailed in Chapter 6.
- Techniques making use of the meter readings from sampled households (beyond averaged profiles) by using the variability between

the samples as a basis for clustering. This work is detailed in Chapter 6.

The work demonstrates how the composite measure proposed in the thesis provides different results when assessing clustering algorithms than previous work has shown and thus provides an alternative (and arguably more useful) method for assessing the success or otherwise of different load profile clustering approaches.

While there has been extensive work on clustering averaged load profiles [14], no previous research has concentrated on using the variability within households as the basis for clustering. This work thus provides an alternative (and complementary) technique for clustering electricity load profiles.

1.5 Thesis Structure

The thesis is structured in the following way:

- This chapter briefly provides the context for the work within the UK electricity market. It also defines the scope of the work and details the research questions that will be addressed in the following chapters.
- Chapter 2 provides a survey of the literature relating to the UK electricity market and the challenges that arise, together with information on Demand Side Management as a key part of the solution to the challenges. It also explains the importance of defining and analysing load profiles as a method of targeting demand side response effectively to small groupings of receptive households. The areas of research that have not previously been addressed are identified.
- Chapter 3 gives information on the techniques used throughout the rest of the work. This includes a description of the data (and how it is prepared) together with the techniques used in the analysis.
- Chapter 4 describes the method that is used to create average load profiles from the base data and then to cluster similar profiles to-

gether to create a few archetypical households. The chapter also introduces a novel approach to evaluating the resulting clusters by calculating a composite measure. The results from this chapter provide guidelines for selection of the most appropriate clustering algorithms for clustering of electricity load profiles.

- Chapter 5 explains the usage of motifs and finds repeating patterns in data streams by making use the electricity meter data. An evaluation of the techniques for finding motifs leads to a set of parameters that is most appropriate for finding a reasonable number of useful motifs. These parameters are then used in the work in the following chapter.
- Chapter 6 introduces the concept of flexibility of regular behaviour within households and explores how this can best be evaluated and used as the basis for clustering into similar archetypical households.
- The thesis concludes with Chapter 7 which draws conclusions from the work and suggests possible areas of future research. It also includes a list of peer reviewed publications and describes how these publications map onto the topics covered in the thesis.

1.6 Summary

The demands on the electricity network require research in the area of demand side management to create feasible solutions to the challenges arising over the next few years. The forthcoming torrent of data from smart meter deployments across the UK will provide new opportunities to better understand household behaviour and to address differing types of households in differing ways depending on their behaviour patterns.

While there has been extensive research, as detailed in Section 2.4.2, in clustering similar households using the average load profiles for a household, there has been no investigation of the variability of the behaviour within the household and this thesis will address that gap.

The data collected in the NESEMP project [15] provides an excellent dataset for analysis and has not previously been analysed. Finding repeating motifs within a stream of data has been researched in various technology areas but has not been addressed in detail within domestic electricity data as measured at the household level. In addition, the symbolisation technique, while applied to many application areas, has not previously been used to find and compare motifs within domestic household meter data.

CHAPTER 2

Literature Review

2.1 Introduction

This chapter provides a description of the current state of the UK electricity market and details some of the issues needing to be addressed now and in the near future. The current period is one of rapid change within the industry which, as well as raising a number of problems, also provides opportunities for addressing the upcoming challenges in new, innovative ways, particularly with regard to the huge increase in availability of consumer data. Domestic households are a significant part of the overall demand [16] as well as being the area of the electricity network that has been least analysed in the past due to lack of data. The changes in metering that will be implemented over the next few years provide an opportunity to analyse domestic household demand in new ways.

To make a significant impact on the electricity challenges, more intensive use of demand side management techniques is necessary to "smooth" household demand [17]. Increasing the effectiveness of DSM will require more accurate targeting of the proposed techniques to change behaviour. This in turn will mean that targeting techniques will need to improve from those currently in use in the industry. The need is for the UK to move from the previous approach of designing supply to match predicted demand towards methods for modifying the demand.

Previous research, and common practice within the electricity industry, has allowed the partitioning of similar households into a few groups which are then represented by an archetypical household. Various electricity industry processes (e.g. forecasting future demand) have then made use of the archetypes.

Various methods of assessing the quality of the results have previously been defined in the field of data mining [18] and previous work [14] has made use of some of these evaluation measures to provide guidelines on the most appropriate clustering algorithms. However, it is argued that the previous work has concentrated on only two aspects of the partitions and other criteria, that can be used to judge a "good" partitioning result, have not been incorporated into the evaluation measures. Work in the field of marketing is reviewed to ascertain appropriate criteria to include in evaluation measures.

The extensive body of work relating to the detection of individual appliances by analysing the household meter data [19] is considered and it is argued that, while a useful goal for research, the current state of the art of appliance detection has too many restrictions to be useful across a large population. Instead, the identification of particular repeating activities (which may involve the use of many appliances) is proposed as a more realistic goal. However, interpreting the detected repeating activities automatically is still a challenge and it is argued that the repeating activities are more usefully used to evaluate variability in behaviour without needing to know what each activity actually represents in the household.

Symbolisation, as the term is used in this thesis, refers to the process of representing real valued data, such as electricity meter readings, by a series of letters drawn from a restricted alphabet (e.g. the letters a to e). A range of real values will match to the same symbol.

The approach of using symbolisation for detecting repeating patterns is considered and its usage with electricity meter data is reviewed. It is found that there has been no application of symbolisation techniques to domestic electricity meter data and it is argued that this is a fruitful area for study.

The investigation of variability of behaviour using electricity meter data is explored and the techniques previously applied are considered. Using a process of motif detection and then assessing the variability of the timing of the motifs is an area of study that has not previously been considered.

2.2 Scope of the Literature Review

The review includes:

- Consideration of the current state of the UK electricity market including an analysis of some of the issues and possible solutions.
- A summary of the current state of load profile clustering.
- A review of the status of motif discovery within electricity load profiling.
- An analysis of the usage of symbolisation and motifs applied to electricity data.

The review does not consider:

- General time series techniques as these have been well studied and are well understood. Instead the focus of the thesis and the literature review is on specific techniques for dealing with electricity meter data.
- The electricity industry beyond domestic households other than in discussion of how changes in domestic household usage can affect the efficiency of the wider electricity network.
- Geographical regions beyond the UK although some useful international studies applicable to the UK are included.

Relevant literature has been identified by the use of Google Scholar to identify key papers and by following citations in useful papers. In addition, various government and commercial reports have been identified which provide background on electricity industry trends and future plans. The greatest emphasis has been placed on literature from the last six years although key publications from earlier are also included.

2.3 Information on Current UK Situation and Challenges

The electricity market in the UK is under various pressures. Some are due to the history and current design of the National Grid and others are arising from worldwide trends, such as the need to reduce carbon emissions and the declining sources of hydro-carbon fuels [2]. New technologies, such as electric cars needing household charging facilities, will be much more common [20]. The information monitored in the home will grow rapidly, particularly with the roll-out of Smart Meters planned for completion in the UK by 2020 [21]. In addition, the drive to change the mix of electricity generation technologies to reduce greenhouse gas emissions, the desire to reduce carbon dioxide by changing non-electric demand such as gas central heating to the electricity network, and the impact of climate change on altering electricity demand and the greater occurrence of extreme weather events will increase the difficulties in providing a stable and cost effective supply without better modelling of the patterns of consumption within the grid.

MacLeay et al. [1] is an annual publication from the UK Department of the Environment which gives details of energy production and consumption over the previous year. This makes a number of points relating to changes in electricity generation and usage including:

- UK electricity generation (including pumped storage) in the UK fell by 1.0%, from 367 TWh in 2011 to 364 TWh in 2012. Total electricity supply (including net imports) increased by 0.6%.

- Final consumption of electricity in 2012, at 318 TWh, was broadly unchanged on 2011 and at its lowest level since 1998. Domestic consumption increased by 2.8%.

- Domestic demand accounted for 30% of the total demand during 2012.

- Of the electricity consumed by the domestic sector in 2012, 21% was reported as being purchased under some form of off-peak pricing

> structure (e.g. Economy 7). 16% of consumption was through prepayment systems, broadly unchanged from the level in 2011.

A report from the Parliamentary Office of Science and Technology [22] gives background briefing information to Members of Parliament and makes the point that legislation (which the UK has committed to) requires some challenges to be met to allow meeting of the commitments.

The Climate Change Act 2008 [23] imposes a number of legally binding commitments on the UK including the reduction of carbon emissions by at least 80% before 2050 (from 1990 levels). The Act also requires the Government to set carbon budgets for each 5 year period detailing how they intend to meet the reduction requirements by setting interim goals. The Committee on Climate Change provides advice on the appropriate level of each carbon budget so that the longer term objectives can be met. The first four carbon budgets (up to 2027) have been defined (and relevant legislation produced). Meeting the fourth carbon budget (2023-27) will require that emissions be reduced by 50% from 1990 levels by 2025 (and 32% below 2009 levels). The fourth carbon budget also includes details on the importance of the roll-out of smart meters and states that they "will be crucial in providing flexibility to respond to volatile power demand and intermittent supply" [24]. The report goes on to note that there are a number of challenges to ensure that the smart meter roll-out encourages and enables consumer engagement.

The EU Renewable Energy Directive 2009 [25] requires that 15% of the UK energy used as electricity and for heating and transport comes from renewable sources by 2020. However, while this is a minimum level of renewable generation, models produced by the Department of Energy and Climate Change (DECC) conclude that electricity supplies will need to be almost completely de-carbonised by 2030 to meet the emission targets. This implies that about 30% of electricity supply would need to be from renewable generation by 2020 (up from only 6.7% in 2009) [26]. Significant deployment of renewable generation at the local level, such as house based solar panels, will disrupt the existing model of electricity flow (from power station to user) and increase the need for effective DSM.

A report from the UK Energy Generation and Supply Knowledge Transfer Network [2] summarises the situation in the UK and makes the following main points:

- To meet the challenging goals for carbon emissions, the UK will need to almost completely de-carbonise electricity generation within reasonably short time scales (2030).
- Issues about security of supply will need to be addressed by ensuring a diverse mix of generation capabilities which, in turn, will require a smart grid to effectively manage the diverse generation resources.
- A lot of electricity network components are nearing their designed end of life and there will need to be massive investment in new assets over the next ten years.
- A smart grid is necessary to allow the electricity network to robustly cope with unexpected events such as terrorist attacks and natural disasters.
- The UK is an island and the electricity network has few linkages to international energy systems. This implies that the UK needs to be more self-sufficient in balancing the supply and demand than some other countries who can rely on neighbouring countries to meet short term shortages.
- The UK has an extensive commitment to offshore wind power which will require investment in an offshore grid to allow efficient integration into the UK network.
- A smart grid enabling full demand-side dynamic response will smooth peaks in demand considerably to match supply. Successful implementation of the smart grid will reduce the amount of backup generation capabilities that would otherwise be needed to handle the intermittent supply of large amounts of wind generation resources.

DOE [27] reports on the findings in the USA from a number of industry experts from 2003 and predicts the situation in 2030 as well detailing issues that would need to be overcome. Points raised include:

- The existing infrastructure for electricity in the USA is old and failing and major changes will be needed, in particular with regard to investment and ways of operating, to meet the challenges that will arise over the next few years.
- There are major concerns that investment may be insufficient to meet future needs.
- Information Technology has been implemented in many other industries (e.g. telecommunications) and has had a transformational impact on the industries but this has yet to occur in the electricity industry.
- The implementation of smart power systems and distributed intelligence needs to be accelerated within the power industry.

These major issues also apply to the UK.

Previously the electricity market was dominated by the energy suppliers. Before privatisation, the UK nationalised energy industry decided on the available tariffs (which were generally very simple) and concentrated on providing a continuous energy supply to their customers. In general, the customers had to accept whatever tariff they were offered as their only negotiating power was political (if price increases were perceived to be too high). After industry privatisation, a degree of choice was offered to customers to move between suppliers but, in practice, most of the suppliers offered similar tariffs and imposed lock-in contracts so domestic users still retained little power. In the future, the increasing likelihood of service interruptions, the changing profile of appliance usage, and the pressures on the energy suppliers to provide more efficient provision of supply is likely to increase the power of the individual consumer. Suppliers will have to provide a variety of tariffs to gain the end-user behaviour that they desire. This increasing consumer power and increase in complication in supplier/consumer relationships will allow much greater implementation of DSM programs.

The major trends in the UK are summarised in Table 2.1.

The UK government produces many reports on UK energy usage [1] including the chart at Figure 2.1 which shows the sources and final usage destinations for electricity in the UK during 2012.

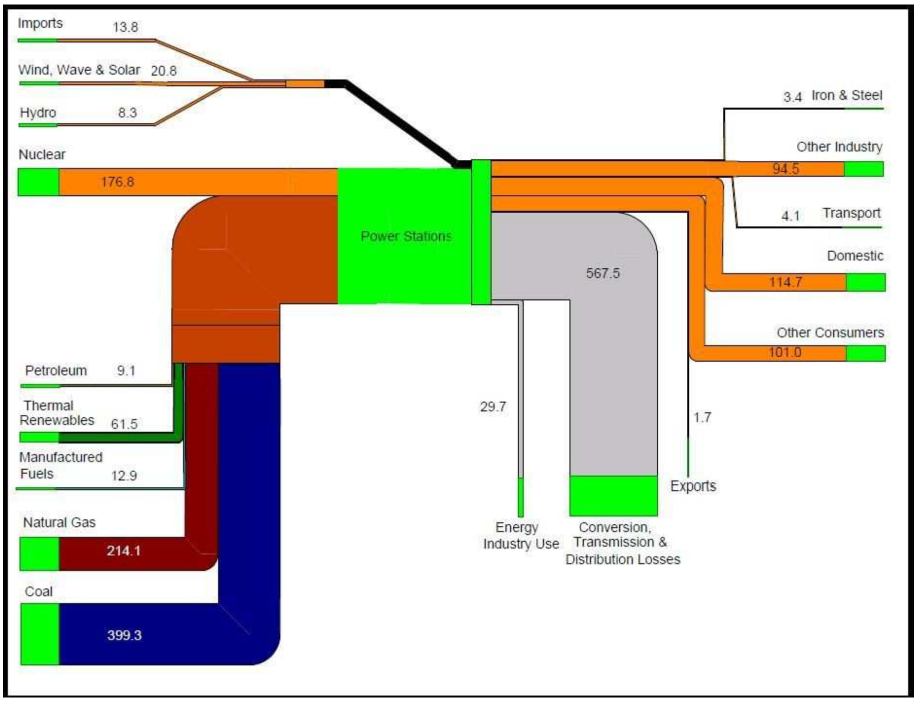


Figure 2.1: Overall Electricity Generation and Usage in UK (2012) [1]

Cap Gemini [28–31] have produced a number of reports detailing how they expect the electricity market to change. The points they make include the growth of new types of companies such as tariff brokers who will make use of the data generated from smart meters to develop a marketplace for sophisticated pricing offers which can be automatically controlled on behalf of the customer.

Industries with access to detailed user behaviour information have introduced various sophisticated customer offers and it is likely that the electricity industry will follow a similar evolution as increased information becomes available. For example, mobile phone providers in the UK offer complicated tariff packages with a certain level of usage included in the monthly fee (e.g. 300 minutes of voice calls, 300 texts, 1GB of data transfer) with additional usage over the bundled amount charged at a premium rate. This could be replicated in the electricity market with tariff offers providing a certain amount of electricity for a fixed fee with addi-

Table 2.1: Some Factors Affecting the UK Electricity Market

Factor	**Impact on Market**	**Trend**
Electric Vehicles	Extensive load needed for charging Opportunity to time shift charging	Currently very low but expected to grow
Direct Current Devices	Impact on supply quality	Growing as LED light use expands
Smart Meter Roll-out	Greatly increased data availability, reduced meter reading costs	Planned to be complete in UK by 2020
Local Generation	Possible overloading of local grid. Change from original design of Grid (power station to consumer)	Low currently but rapidly growing
Local Storage	Allows for load balancing	Very low at present
Real Time Pricing and new kinds of tariffs	Possible way of balancing load	Not currently used as requires smart meters
Energy Usage Feedback	May reduce overall usage and time shift usage	Simple monitor displays now, more sophisticated in future
Price Increases	Political influence on suppliers to minimise price increases	Simplifying of tariff offers to avoid obfuscation
Energy Reduction Incentives	Government under pressure to reduce cost to consumers of reduction initiatives	Less uptake of reduction measures (e.g. insulation). Less budget for low income households for measures.

tional usage being more expensive or, maybe for a discount, not being available (i.e. the household has a power cut once the limit is reached).

Grocery delivery provides an example of differential pricing that could be applied to the electricity industry as more information becomes available. Popular time slots or days (e.g. weekends) generally incur a higher delivery charge. A similar concept could be introduced for electricity where usage of electricity during periods of peak demand could have a different tariff rate which would either increase revenue to the electricity supplier to help meet increased generation costs (or wholesale prices), or to discourage usage during peak times for appliances which can easily be time shifted (e.g. dishwashers).

The Internet Service Provider (ISP) industry have introduced many measures to control demand to meet supply at peak times. In the UK, depending on the type of service purchased, many users have a reduced quality of service at times of peak demand for network bandwidth as "throttling back" prioritises certain traffic over peer-to-peer networking (often file downloads) and other usage. This reduction in service can be avoided by the consumer paying a premium price. A similar situation could arise in the electricity market in the future (and already exists for industrial users) where consumers paying premium rates are less likely to be impacted by service interruptions.

Home telephone packages in the UK often include off peak calls at no cost (or as part of the overall package). This can have the effect of modifying consumer behaviour in that they are more likely to make their calls during the non-business hours (outside peak time) and this allows the telephone service providers to better balance network capacity between domestic and business users. This type of offer is a good example of using price signals to modify consumer behaviour for the benefit of the overall network.

With the advances in technology, it often becomes possible to merge offerings across different markets (e.g. the bundling of TV, broadband and telephone services in single packages), and a similar situation may arise in the electricity market as electricity suppliers may bundle the costs of electricity with other services such as appliance maintenance insurance or home generation technology (e.g. wind turbines).

2.3.1 Smart Grid

To assess the degree of "smartness" within the Electricity grid, Dupont et al. [32] propose six areas of measurement that can be used to define smartness. One of these is the area of informed participation by customers which covers measurements relating to the numbers of smart meters deployed, the fraction of customers using real time pricing, the amount of smart appliances sold, the amount of customer load subject to demand side management, and a measure relating to local generation. While focusing on the overall network, the paper makes some general points but excludes a number of other factors relating to the domestic household such as the number of behaviour modification interventions (outside of the narrow definition of real time pricing).

Microsoft [33] are an important supplier within the technology industry and are often able to mould new markets to match their product offerings. They have published a smart grid architecture and expect to see the grid evolve towards a Smart Ecosystem.

The UK electricity market will allow for more targeted and more complicated tariff offers for customers to provide many benefits including maximising the efficiency of the supply process. DECC [3] shows that the provision of Smart Meters will allow greatly increased analysis of a customer's electricity usage and provide the ability to make customised offers on pricing and availability to change customer behaviour (for example, to minimise usage during peak periods) or to increase efficiencies in the electricity supply chain in meeting the predicted demand [34].

In May 2013, DECC announced that the mandated UK roll-out of smart meters will commence in late 2015 and be complete by the end of 2020 (an adjustment from the previous goal of completion by 2019) [1]. A detailed description of the roll-out approach can be seen in DECC [21].

The report from the UK Energy Generation and Supply Knowledge Trans-

[1] www.gov.uk/government/uploads/system/uploads/attachment_data/file/197794/smart_meters_programme.pdf

fer Network [2] suggests that there could be issues with customer backlash against the smart grid (in particular the implementation of smart meters) due to issues of privacy and "Big Brother". The author of the report believes that little research has been done on the social impact of the smart grid (other than the success or otherwise of usage feedback displays) and much more work needs to be done. DECC commissioned a report [35] which investigated consumer attitudes to smart meters in 2012 and, while they found that 32% of respondents supported smart meters, 20% were against the deployment and a large number (48%) were undecided. This suggests there is likely to be a sizeable minority opposed to the roll-out of smart meters and this may be a factor impacting on their success.

The Netherlands government encountered problems implementing a national smart meter program due to concerns about privacy. Cuijpers and Koops [36] describes the situation and makes the point that legislation for smart metering needs itself to be smart and that the privacy issues need to be well addressed as "Energy consumption reveals details of personal life, in the most privacy sensitive place - the home, and therefore smart metering has to strike a careful balance between detailed energy metering and privacy protection". The privacy issue has not been a major problem in the UK as of 2014 but could become more important as meter roll-out increases and consumers become aware of the information being collected.

The Energy Demand Research Project [37] details a major UK study which ran from 2007 to 2010 and comprised 60,000 households including 18,000 with smart meters. Various differing demand modification initiatives were explored with the households. One major finding from the study was that households with smart meters demonstrated a more successful response to interventions and larger reductions in energy usage. No reliable or persistent effect was found from implementing financial incentives which supports the view that a comprehensive set of possible interventions (beyond pricing incentives) needs to be explored by utilities.

The overall message from the various smart grid studies referenced is that the deployment of smart meters will not be straightforward and that successful implementation will require addressing of various non-technical concerns such as the perceived privacy implications. In addition, research shows that deployment of smart meters does not, in itself, produce large changes in usage of electricity and that programmes targeted at parts of the consumer base will be necessary to gain effective, large changes in behaviour.

2.3.2 Domestic demand

Alongside changes in electricity generation technologies (e.g. more nuclear generation), a focus on the way that the generated electricity is used is going to be an important factor in reaching the legal requirements. Understanding and influencing domestic usage will be a significant part of the solution.

When considering the need to reduce carbon emissions in electricity generation the fourth carbon budget report [24] notes that there is "power system flexibility on the supply side (e.g. fossil fuel plant can be operated flexibly), but limited demand-side flexibility (e.g. only the largest customers are able to respond to high prices when the system is operating at capacity)". New low carbon generation plants are likely to be less flexible than traditional fossil fuel generation and the need for demand side flexibility will increase to cope with weather variabilities (e.g. lack of wind) and to reduce the need for high levels of backup generating capacity previously needed to avoid power cuts.

The Committee for Climate Change commissioned a report on possible ways of ensuring power generation flexibility [20] which makes the point that "Facilitating demand-side response through the roll-out of smart technologies and tariffs could provide a key source of within-day flexibility". This will be particularly important as increasing amounts of transport energy usage (e.g. electric vehicles) and space heating (e.g. household heating by electricity in preference to gas) is put in place.

The massive increase in data provided by the roll-out of smart meters allows for much better understanding of current domestic electricity usage and provides the basis for implementing programmes to change the behaviour for a more effective overall network.

In particular, the change of sampling of electricity usage from a three monthly billing cycle to a 30 minute sampling period using smart meters, alters the degree of understanding of households' behaviour that is possible [3]. DECC [38] provides the technical specifications for smart meters in the UK which defines a sampling frequency for reporting from the household to the utility company of 30 minutes while specifying an in-house reporting of usage of ten seconds. The selection of a sampling period of 30 minutes has been made on political and economic grounds and a more frequent sampling rate is technically possible and may allow enhanced analysis.

An important factor influencing the UK electricity market is that the assumption by consumers of the availability of a practically infinite supply of electricity (albeit at a cost) is no longer valid and domestic users will have to adapt to changing ways of using electricity or suffer from increasing unreliability of the electricity supply.

2.3.3 Demand Side Management and Demand Response

Research in the area of demand side responses has been ongoing for a number of years. For example, Newborough and Augood [39], in 1999, demonstrated the ability to reduce UK household peak usage of electricity by up to 60% by an assortment of interventions including the replacement of some appliances by gas powered equivalents. Chamberlin [40] also showed that demand side management was being seriously investigated over 20 years ago.

Prior to the roll-out of smart meters, electricity suppliers were reliant on a meter reading on a three monthly cycle for feedback on usage by a given household. This provided a single reading (or possibly two readings for households with Economy7 meters) for total usage for the three monthly period and hence gave no feedback on time or day of usage (beyond the Economy7 period). Electricity suppliers were therefore unable to offer tariffs to change behaviour as there was no way of knowing the detailed prior behaviour, or the subsequent behaviour, resulting from introduction of the new tariff.

Ofgem [17] estimates that the Electricity industry in the UK will need to invest an estimated £32bn by 2020 to deliver the networks required for the low carbon economy and to maintain secure, reliable supplies. This is a near doubling of the expenditure seen over the last twenty years. Ofgem also considers shortfalls in the current UK infrastructure which will impact on addressing the challenges of the next 10-15 years. Particularly of interest is the fact that all the proposed possible packages of changes (except for the particular package suggesting a central energy buyer) include the need for an increased ability for the demand side to respond to supply signals.

Historically, electricity supply in the UK has been driven by a desire to provide sufficient supply to match the predicted demand and to avoid shortages and blackouts. The restrictions on supply due to cost, changing political opinions regarding generation technologies, and international obligations to meet carbon reduction targets, means that, in the future, the emphasis will need to change to demand more closely matching the available supply. One approach to addressing this issue is the application of demand side management (DSM) techniques to achieve changes in consumer behaviour. River [4] defines DSM as "systematic utility and government activities designed to change the amount and/or timing of the customer's use of electricity" for the collective benefit of society, the utility company, and its customers.

Tata Power provide a summary of how they define demand side management [2] including Figure 2.2 which details the steps needed in a typical DSM program. Figure 2.3 provides a summary of the varying objectives for a DSM program. For example, it might be the goal of the program to reduce the peak demand (to reduce standby generation capacity) and a "peak clipping" program may be instituted.

River [4] provides a good explanation of what is meant by demand side management as well as detailing various US-based trials of variable pricing. It provides a good historical perspective on DSM and makes the point that the current focus is on real time pricing (or dynamic pricing)

[2] http://cp.tatapower.com/cip/cpnew/dsm/demand-side-management.php

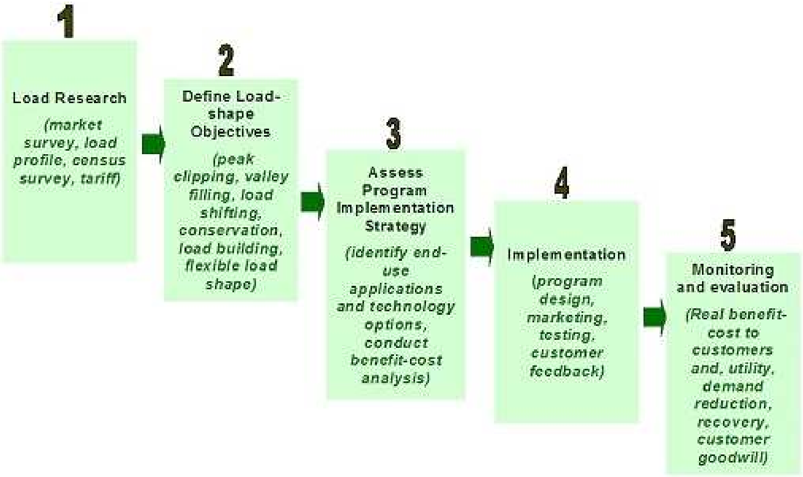

1
Load Research
(market survey, load profile, census survey, tariff)
2
Define Load-shape Objectives
(peak clipping, valley filling, load shifting, conservation, load building, flexible load shape)
3
Assess Program Implementation Strategy
(identify end-use applications and technology options, conduct benefit-cost analysis)
4
Implementation
(program design, marketing, testing, customer feedback)
5
Monitoring and evaluation
(Real benefit-cost to customers and, utility, demand reduction, recovery, customer goodwill)

Figure 2.2: Stages of Demand Side Management (from Tata Power)

rather than Time of Use (TOU) pricing. Dynamic pricing allows utilities to change prices in real time based on the current load on the electricity network and requires consumers (previously industrial customers but increasingly also domestic customers) to sense the changing price and make changes to their electricity usage automatically (e.g. by using smart appliances).

The report also states that "it is important to note that the demand-side of the market has typically been neglected. This has been true in developed as well as developing countries."

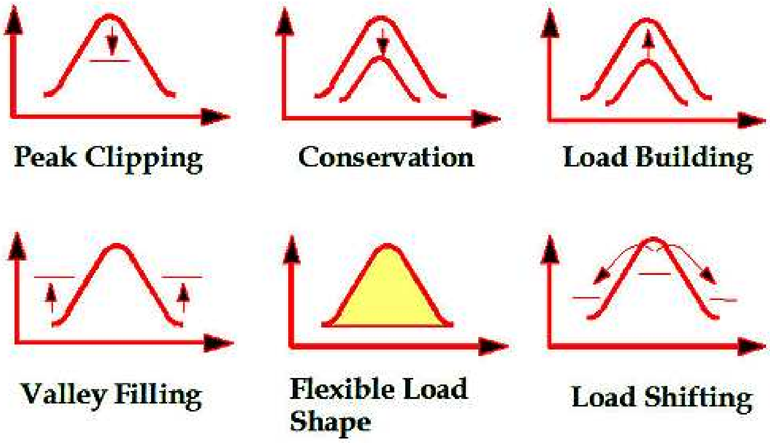


Figure 2.3: Load shape objectives (from Tata Power)

Hamidi et al. [41] has investigated the amount of the total domestic demand that could be responsive to demand side management interventions by examining a small number of households in the UK. This analysis was done by considering different classes of appliances (e.g. “wet” such as washing machines, “cold” such as fridges) and examining at which periods of the day they are used as well as considering how their usage could be varied. This work is useful as it can be applied to activities derived using motif detection and could give utility companies a measure of how responsive a particular activity may be to interventions.

Zachary et al. [42] provides a mathematical approach to assessing the value of wind generation capacity and how the ability to move peak demand can be measured in terms on impact on the cost of wind generation capabilities. This approach could be used to give a financial value to DSM programmes but is based on many assumptions and is currently only applicable to wind generation.

Darby and McKenna [43] reviews the social aspects of demand response programmes with a particular emphasis on distinguishing the UK (as a country with a temperate climate) from other countries with different climates. The study identifies short term (up to 2020) approaches to demand response and longer term (after 2020) changes. Darby and McKenna [43] makes the useful point that “From the user perspective, the extent to which people shift their consumption patterns will depend on factors such as perception of the need to do so, trust in the utility or energy service provider, incentives, and transaction costs (including cognitive costs).”

While many DSM programs have been tested, there have generally been poor results from the research with improvements in electricity usage typically being only a few per cent. Some reasons for demand side response programmes not being taken up in the UK are suggested by Torriti et al. [44]:

- Utilities are measured by the government on achieving a net reduction in electricity usage and many DSM programs do not clearly generate a net conservation (while they may usefully shift the time of peak usage).

- Demand side response technology has been available but only at a high cost relative to the electricity savings expected.

- A large electricity demand targeted in some countries is the use of air conditioning (AC) units with consumers encouraged to change the times of usage. However, as there is little domestic AC deployment in the UK, this area of behaviour modification is not available.
- Government regulation and industry oversight does not currently encourage utilities to implement DSM programs.

As well as the above points, the history of energy usage in the UK (including little air conditioning, a temperate climate, and the previous high deployment of storage heaters) makes the UK a special case. In addition, the availability of North Sea gas and cheap coal led to expectations of low energy costs and thus poor thermal qualities of a lot of the UK housing stock. Thus, a lot of research from around the world may not apply to the special circumstances of the UK situation.

Various studies have investigated the type of feedback that is most effective in changing individual's usage of electricity to meet peak shifting or overall reduction goals (for example Fischer [45]). These studies provide some useful information although it is possible that other, more fundamental, interventions (for example, by altering local travel conditions such as bus timetables) will be more successful in meeting the goals of demand side management. Neenan [46] provides a comprehensive review of various feedback studies relating to energy usage and distinguishes between six levels of feedback varying from standard monthly or quarterly billing to "real time plus" feedback (which includes real time details on usage of individual appliances).

An interesting study in Australia extended the deployment of monitors to the use of water (in addition to electricity) with the monitors providing alarms to the household occupants when usage hit certain levels. A saving of 3% of water and 2.4% of electricity was reported [47]. This shows how the need for demand side management can be extended to other resources requiring management. However, while important in some parts of the world, the need to manage water efficiently is less important in the UK and may be better addressed by improving water supply management (e.g. fixing leaks) before the need to deploy water demand management programmes. As the predicted changes in climate come to pass, the importance of water management may increase in the UK as droughts become more likely.

2.4 Generating Archetypes

2.4.1 Marketing Experience

The field of marketing has explored the segmentation of a population of individuals into archetypical groups for marketing purposes and has been applied in many different situations. In particular, the generation of clusters showing individuals with similar behaviour is used as a precursor to the creation of a specific offer to a given cluster that the marketing professionals believe will be attractive to members of that cluster (and hence taken up).

There is a body of literature in the field of marketing which includes extensive discussion of the requirements for effective cluster analysis. Most of this discussion is based on that proposed by Kotler and Keller [48] which suggests that clusters should be measurable, accessible, substantial and actionable. This list has been extended by many authors with a good summary of the state of the art provided by Dibb [49] which also considers the problem of assessing the attractiveness of each derived segment. The points raised in Dibb [49] are summarised and further modified by Sarstedt and Mooi [6] who suggest criteria by which clustering solutions should be assessed including:

1. Substantial: The partitions need to be of a reasonable size to be useful as interventions are likely to be expensive and must be addressed to a sizeable population to be effective.

2. Accessible: The partitions must be able to be effectively reached and served, which requires them to be characterised by means of observable variables.

3. Differentiable: The partitions need to be distinguishable and to respond differently to different interventions.

4. Actionable: Effective programmes can be formulated to attract and address the partitions.

5. Stable: Partitions that are stable over time are more attractive for an effective intervention strategy.

6. Parsimonious: To be meaningful to a wide audience making use of the clustering results, only a small set of substantial clusters should be identified.

7. Familiar: For wide acceptance, the cluster's composition should be comprehensible by the layman.

8. Relevant: Partitions should be relevant in respect of the organisation's competencies and objectives.

9. Compactness: Partitions should exhibit a high degree of within-segment homogeneity and between-segment heterogeneity.

10. Compatibility: Segmentation results should meet the requirements of other aspects of the organisation's business.

In general, previous work in the data mining field on defining cluster quality measures has concentrated on addressing the compactness (homogeneity and separation) criterion [50].

2.4.2 Load Profiling

The identification of typical electrical usage patterns within households is necessary as a starting point for:

- Defining the type of DSM program (e.g. peak clipping) to undertake to match the overall goals.

- Understanding the current pattern of electricity usage to allow for decisions on desired changes needed to the pattern.

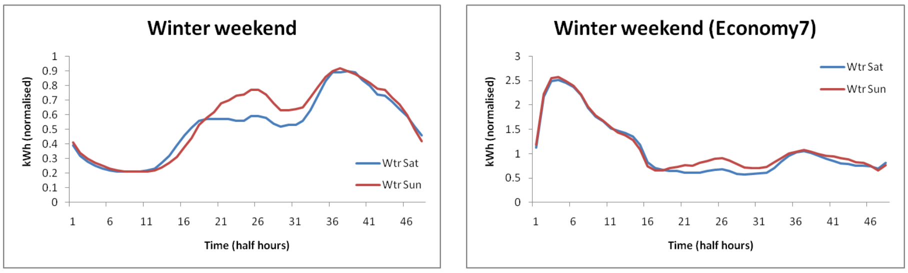
- Assessing the impact of any initiatives to reduce overall energy usage to discover the amount of overall reduction which occurs during different times of the day.

- Allowing accurate aggregation to provide a pattern of total demand over the day that is to be met by supply side generation and transmission.

The definition of load profiles has been a long standing activity within the electricity industry but the current and forthcoming avalanche of data provides for alternative ways of approaching the definition of the load profiles.

Current industry practice has focussed on commercial electricity users. Electricity Association [51] identifies a process for defining the details of eight different standard usage profiles for the UK. Of these eight, only two refer to domestic properties although the profiles take into account the season and the day of the week. As an example of the standard profiles, Figure 2.4 shows the winter profiles for Saturday and Sundays, both for Economy7 customers and non Economy7 customers, plotted as 48 half hourly readings across the day. Economy 7 is a tariff offer that provides much cheaper night time electricity (typically between 11pm and 6am) at the expense of increased day time charges.

The focus of most of the load profiling work was to develop load profiles for archetypical customers in the absence of regular customer meter readings. However, with the roll-out of smart meters and other monitoring devices, the focus of the load profiling work can move from building up a profile to exploring the details of the profiles captured by the monitoring devices. To apply future DSM techniques, a more precise splitting of the users into similar groupings is needed thereby allowing for targeting of appropriate groups.



(a) Standard users **(b)** Economy 7 users

Figure 2.4: Example UK Industry Standard Profiles

In a report to the DistribuTECH Europe DA/DSM Conference, the Load Research Group of the Electricity Association Services Ltd. [52], set forth a set of general guidelines for load profiles:

- Each profile should represent a relatively homogeneous group of customers.
- Each profile should be distinctly different from the others.
- The identifying characteristics for assigning customer load to a profile should be readily determined.
- The number of load profiles should be relatively low.
- The accuracy of estimated load profiles should be judged primarily on how well they perform over a trading period (typically one year) .

Bailey [53] provides a good summary of how load profiles can be built up and distinguishes between methods that produce different kinds of profiles:

- Dynamic profiling from collecting meter data from a subset of customers.
- Dynamic modelling where knowledge of an external factor (e.g. temperature) is used to modify the standard profile using some formula.
- Same day profiling where a similar day from history is selected to represent the current load profile for a given day (based on weather, overall loading or other criteria).
- Static profiling where a profile is derived for a given season and type of day (e.g. weekend).
- Deemed profiles which are built up from assumptions made about the detail of the load used. For example, a typical household could be assumed to use certain appliances at certain times of the day and the aggregate load profile is built up from the profiles for each appliance.

Swan and Ugursal [54] provides a review of the modelling techniques in use in load profiling. The paper distinguishes between a top down approach which is not concerned with individual household appliances but which apportions calculated usage for a geographic region to individual households, and a bottom up approach which calculates the usage of individual houses and then aggregates this information to a regional level. The top down approach is criticised as being reliant on historical data and being unable to model disruptive advances in technology (such as the take-up of electric cars). However, the data collection cost of the top down approach is much lower than the bottom up approach with its need to monitor at a much finer level.

Baker and Rylatt [55] investigated the differences between households using questionnaires and then used this information to form clusters. Analysis of the criteria for membership of particular clusters showed the importance of home working as a major influence on the overall load profiles.

Figueiredo et al. [56] worked with data from Portugal for a small number of households for which appliance level measurements had been taken. From this data, the work builds up an aggregation of usage of appliances to create aggregated load profiles. These are then used to assess the effect of simulating altering the time of appliance usage on the overall load profile for a collection of households. This is close to the requirements for effective demand side management as it allows utilities to assess the impact on their supply requirements from changing consumer's demand patterns. However it relies on detailed monitoring of a few households and would not scale well for reasons of cost and time.

Load profiles form an important part of the requirement within the electricity industry to perform short term (a few hours) load forecasting. Gross and Galiana [57] provides a review of the current approaches and, in relation to load profiles, concludes that the load generated by a given household is a function of three principal time factors: the season, the weekly/daily cycle and the occurrence of public holidays. Other factors come into play relating to school holidays, daylight saving and weather conditions which can have large short term effects on the longer term patterns.

Cancino [58] provides a review of the literature on load profiling which is defined as the application of methods that deal with customers' load diagrams with the goal of grouping customers with similar load profiles into coherent clusters.

Most of the published load profile work assumes that the load profile for a given household on a given day of the week and season is relatively constant. In practice, the actual usage on the day will be influenced by external factors such as the weather. Lin et al. [59] has studied this in China and used the similarities in how substations react to external factors as the basis for clustering. This work uses information on the peak temperature and peak relative humidity. However, this work is less applicable to the UK due to the lack of air conditioning in domestic properties and the wide usage of gas for space heating.

Zakaria and Lo [60] distinguishes between static load profiling, dynamic modelling and dynamic load profiling. Static load profiling relies on the collection of data from a large number of customers for a period of over a year. This data can then be categorised by season and type of day (e.g. weekend) and then averaged to create a load profile for the customer. The paper makes the point that climate is not included in the categorisation and this can be a significant shortcoming. Dynamic modelling makes use of the historic load shapes (as with static load profiling) but also includes a climate adjustment mechanism. The dynamic load profiling method relies on load data being read regularly (daily) with "new" load profiles being produced daily. However this approach relies on the installation of monitoring equipment in each of the households requiring large commitments of time and money.

A good example of a bottom up approach has been taken by Ihbal et al. [61]. Using the results of questionnaires and making assumptions about the occupancy of a household at different times of the day and the probability of using a particular appliance, it is possible to build up an overall load profile for the given household. This generated profile can then be aggregated over a population of different types of households (e.g. single person, retired) to create an overall usage profile for a community. The approach is valid but depends on various assumptions on parameters such as occupancy rate, or probability of using a particular appliance, and relies on extensive survey work to build up realistic estimates for these parameters. The workload (in time and money) means that this approach cannot be scaled to large populations and shows some of the drawbacks of the bottom up approach.

Capasso et al. [62] is a similar example of a bottom up approach to generating a household load profile. The paper uses information on the electricity usage profile of individual appliances and then uses probabilistic methods to build up a profile of the usage of particular kinds of appliances and thence to an overall aggregation for the whole household. The modelling includes information on household occupancy and draws on psychological theories. The aggregation is further extended over the many households in an area to generate an aggregation for the district. The model used and the predictions resulting have been validated against the aggregation of 95 households in Milan, Italy. Paatero and Lund [63] also uses a bottom up approach by defining a set of appliances that each monitored household is assumed to use. Each appliance is then assigned a usage profile (i.e. how often and for how long the appliance is used) which are then aggregated to create an overall household load profile.

Dominguez-Navarro et al. [64] uses a top down approach to take a load profile for a group of consumers and to use Tabu search to disaggregate this overall demand into that for individual households based on a number of assumptions about the type of electricity usage at different times of the day. This approach attempts to overcome the absence of detailed meter data from each household but is an approach that will be superseded by the roll-out of smart meters.

2.4.3 Clustering of Load Profiles

Electricity load profiling involves the creation of a daily shape for each household based on regular (e.g. half hourly) meter readings. The research area of load profile clustering takes the generated household load profiles and uses various clustering techniques to group the households into a small number of partitions.

A large amount of the literature on load profile clustering makes use of the periodic readings for a household across the day (e.g. 24 hourly readings or 96 x 15 minute readings) and cluster using these readings as the dimensions input to the clustering algorithm. Some work has been done to derive other measures from the periodic readings and Chicco et al. [65] has calculated a number of ratios, such as night time usage versus day time usage, before using these as the basis for clustering.

There has been extensive research on determining user load profiles to represent household's electricity usage [14, 66, 67]. In many cases, (e.g. Ramos et al. [68]), these profiles are then used as the basis for clustering "similar" households together to develop a small set of archetypical profiles which can then be used as the target for various behaviour change interventions.

Many differing approaches to finding the "similar" households have been applied in the load profiling area and various measures have been suggested to determine which provides the "best" clustering solution. Most of the previous work has used very similar ways of assessing "best" and there is a gap in the literature for a measure that gives a quantitative value for the effectiveness of a particular clustering approach in terms of how the results can best be used to change electricity usage behaviour.

Chicco et al. [69] distinguish between the contractual information retained by electricity suppliers (such as type of business, supply voltages, contractual information on outages) and the field measurements from the actual households and businesses on electricity used. The paper considers various possible indices that can be calculated from the hourly electricity readings (e.g. the ratio of night time usage to day time usage) and makes use of these indices for unsupervised clustering.

Chicco [14] provides a 2012 view of the state of the art for clustering of electricity load patterns. This paper contains a review of the differing algorithms used including kmeans, fuzzy cmeans, self-organising maps, and hierarchical clustering using various linkages. Other techniques have also been used including "follow the leader", entropy-based algorithms and neural networks. The paper also contains a review of the cluster validity indices used in various work. The conclusion is that the best validity indices are those that manage to isolate the outliers in the data. This review supersedes previous reviews of the subject by the same author [69, 70]. The clustering algorithms used within this thesis are described in more detail in Section 3.2.

Singh et al. [71] has applied Gaussian mixture models to electricity load forecasting and has found benefits in using the technique from a performance point of view. The work's emphasis is to compare the Gaussian mixture models against other possible probability density functions being used to model electricity load. The approach is applied to forecasting load in particular parts of the electricity supply network rather than within a household but demonstrates that it is a technique worth exploring further.

Chicco and Akilimali [72] proposes a novel method of clustering similar load patterns using the similarity between the centroids found using an entropy based algorithm (rather than Euclidean distance). The method effectively identifies outliers and makes use of this knowledge in creating the clusters. Various cluster validity indexes (CVIs) are used to show the benefit of the novel approach over traditional methods of clustering.

Bidoki et al. [73] evaluates various different clustering techniques including Classical Kmeans, Weighted Fuzzy Average Kmeans, Modified Follow the Leader, Self-Organizing Maps and Hierarchical algorithms using two cluster validity indexes to assess quality. The conclusion is that the best algorithm depends on what results are required from the clustering exercise (e.g. better separation or more compact clusters) with weighted fuzzy average Kmeans providing a good compromise. The work makes use of the daily load profile curves only and does not consider statistics derived from the actual data readings. The work is useful for suggesting clustering algorithms to include in analysis.

Chicco and Ilie [74] provides details of using support vector clustering as a useful way of identifying outliers within the data and then clustering using the data with the outliers excluded. The benefit of the technique, compared to other clustering approaches, is tested using various cluster validity indexes. The work uses the measured meter data collected at hourly or 15 minute intervals as the basis for the load profile and does not consider other measures derived from the data. The work provides useful guidance for selection of appropriate clustering techniques to consider as well as cluster validity indexes to use to measure the relative benefits of each algorithm. The work follows on from earlier work by the same author investigating the use of self-organising maps [75].

Gavrilas et al. [76] uses a Honey Bee Mating Optimisation algorithm to cluster the load profiles for a small sample of customers. The main benefit of the approach is to minimise the amount of parameter setting that is necessary for the clustering. This provides an interesting alternative clustering algorithm but has not been widely used in the load profile clustering community and does not appear to offer dramatic benefits.

Gerbec et al. [66, 77] use hierarchical and fuzzy cmeans algorithms to cluster small business users within Slovenia into typical clusters. However, additional information on the business area of the customers (e.g. manufacturing, service industry) was used as an attribute to combine some of the clusters to reach a small, manageable number of partitions. The work is not applicable to the UK domestic market as it relies on information not available for domestic customers. However, it provides an example of using hierarchical and Fuzzy cmeans as the clustering algorithms and concludes that fuzzy cmeans is the most useful. The study builds on previous work [78] which only made use of a hierarchical algorithm.

Figueiredo et al. [79] have extended the clustering approach to involve the classification of households into the classes created by the clustering part of the framework. This allows for new customers to the analysis to be assigned to existing archetypical groupings based on their measured load profile. This work raises the important point that membership of households and their behaviour is not static over time. For example, houses can be bought and sold resulting in completely different members of the household. Alternatively, life style changes (e.g. retirement, death of a member of the household) can occur and which impact on the behavioural patterns detected in the household.

Garamvolgyi and Varga [80] have taken a self-organised map approach to clustering households using their load profiles. This work has then been extended into an analysis of the price of energy on the spot markets with this data applied to the archetypical load profiles defined for each group to measure the differing costs that could be possible for each consumer group. This approach can form the basis of the demand side goals in that utilities are likely to wish to move households from a relatively expensive grouping to one where the demand can be addressed by buying wholesale electricity at a lower price. This movement will generally correspond to moving households from peak time to off peak usage. The work in this thesis allows the definition of the customer groupings and a similar cost analysis could be applied to the groupings to determine the best intervention to implement, as well as the best grouping to target.

Lopez et al. [81] have made use of Hopfield Artificial Neural Networks to cluster load profiles (measured at hourly intervals) making use of the MIA, CDI and DBI cluster validity indexes to assess the differences between the various dimension reduction techniques tested. The volume of data analysed (230 properties) does not require the dimension reduction techniques. However, some interesting ratios of usage derived from the hourly measures are defined.

Mori [82] provides an analysis of 42 studies applying data mining techniques to power system issues. Only a small minority of these are relating to load profiling but the overall results show a large preference for using decision tree algorithms. Little further analysis is included in the paper but it provides some good references for further information.

Gullo et al. [83] use kmeans clustering of load profiles as well as a novel approach (called TSpart) to create archetypical profiles for domestic customers. As well as making use of Euclidean distance, the authors also explore using Dynamic Time Warping (DTW) distance as the method of measuring separation of load profiles and conclude that DTW is most effective. CDI and MIA cluster validity are used to measure the benefits of each approach.

A criticism of the load profiling work is that load profiles are normalised and the households are grouped using the shapes of the profiles. This takes no account of the total amount of electricity used and a high usage household may be grouped in the same cluster as a very low usage household which happens to have the same shaped load profile. For effective

DSM, it is likely that focus should be placed on the high usage household in preference to the low usage household. Jardini et al. [84] introduces the concept of distinguishing between residential customers based on their total amount of usage before considering the load profiles and clustering. Most other studies have considered all households as "equal" and have normalised the data across all the available households.

2.5 Aspects of Clustering Techniques

2.5.1 Fuzzy Clustering Techniques

The data collected from households is necessarily noisy as it is a reflection of the human occupants of the house and their usage of the house. In addition, a lot of the terminology of domestic electricity usage (e.g. high/low user, green) is vague and there is likely to be a place for using fuzzy techniques to model this vagueness.

The clustering techniques explored in the literature include the use of fuzzy techniques to incorporate the inexactness of the daily electricity usage pattern of a given household. Previous work on this area can be found in [85] which details the application of Fuzzy Cmeans to a set of meter data from Milton Keynes (93 households), demonstrates how these households can be clustered. The paper shows how the fuzzy membership function can be used to allow personalised marketing offers to be made to each household while the utility company only deals with a few, archetypical households.

Chang and Lu [86] provides a method of using Fuzzy Cmeans clustering to assign domestic customers to already known load profiles based on their monthly usage statistics only. This precludes the need to install a meter in each household to provide hourly readings. While useful in the past, this approach is of little benefit as smart meters are rolled out and the main benefit of the technique (the determination of load profiles without the need for monitoring) becomes redundant.

2.5.2 Cluster Validity Indexes

When clustering a population into a number of groupings there are many choices to make including, amongst others, the number of clusters, the clustering algorithm to use, the parameters to define within the chosen algorithm, and the attributes of the members of the population to use for the clustering. Thus, many possible partitions can be created using differing choices and a method of selecting between these partitions in necessary to choose the "best" solution.

To provide an objective evaluation of differing partitioning schemes, various cluster validity indexes have been developed. A recent review of the state of the art regarding clustering, including a discussion on validity indexes, has been published by Jain [50].

Validity indexes can be grouped into three major categories; internal, external and relative [87]. External indexes are those used to compare the generated clusters with previously known information that has not been included in the clustering exercise and which is referred to as the "ground truth". Internal indexes are concerned solely with the internal representation of the generated clusters. An example may be an index calculated from the "tightness" of the members of a given cluster. Relative indexes provide comparisons between different clustering solutions built using different input parameters.

Chicco et al. [69] introduces two cluster adequacy measures to allow comparison between differing clustering approaches. The MIA (Mean Index Adequacy) gives a value which relies on the amount by which each cluster is compact - i.e. if the members are the cluster are close together the MIA is low. The CDI (Cluster Dispersion Indicator) depends on the distance between the members of the same cluster (as for the MIA) but also includes the distances between the representative load diagrams for each cluster. This therefore measures both the compactness of the clusters and the amount by which each cluster differs from the others. These measures are used extensively in further papers on the subject of electricity load profiling. The adequacy of the clustering measures has been assessed by consideration of 471 non-residential customers of the Romanian electricity supplier to determine clusters and to match the generated clusters against the marketing information used by the electricity supplier.

2.6 Behavioural Patterns

Davito et al. [88] provides some background information on Smart Grids and the impact of Demand Side Management. The authors identify six factors that influence the behaviour of customers and determine that organisations implementing behaviour change interventions need to make use of one or more of the factors. The factors are rates, incentives, access to information, technology and controls, education and marketing, and customer insight and verification. This report supports the belief that differing approaches are needed for successful behaviour modification initiatives for differing groups of households and hence emphasises the importance of accurately identifying the archetypical households.

Raw and Ross [37] provides feedback on a large scale (60,000 household) study on the effects of differing approaches to reducing energy usage. Reporting of a household's usage against benchmarks of similar households was identified as a successful approach and quotes "Although a small effect, this is one of the clearest pieces of evidence for an effect of benchmarking". This shows that to affect consumers, there is a need to report back on their usage and how it could be modified (either for reasons of supply efficiency or for personal benefits such as reduced costs). The paper develops a simple theoretical framework based on the means, motive and opportunity for householders to change their behaviour (i.e. for householders to reduce energy demand, they must know what to do, have a reason for doing it and have the resources to do it).

Abrahamse et al. [89] provides a psychological approach to reviewing various studies assessing the success of interventions intended to reduce electricity usage. This makes the useful point that most studies have intended that participants make conscious decisions on their electricity usage (e.g. by the use of feedback monitors) and little work has been done on changing the environment to allow for unconscious changes. Most studies have concentrated on measuring the amount of desired change in behaviour rather than studying the reasons for the measured behavioural change.

The Energy Saving Trust reported on a study comprising a few 100 UK households [90] in which they explored in great detail, using diaries, interviews and similar approaches, the day to day behaviour of the sampled households. This gives useful information which can be compared with groupings of similar households generated by considering (only) the meter data.

2.6.1 Appliance Detection

Previous research into detailed household behaviour has generally concentrated on working with a small number of households which are well understood, which include many different monitoring devices, and where the householder is supportive of the research and is prepared to dedicate time and effort to correct labelling of devices and following researcher defined procedures. There remain a large number of households where there is not the commitment to "green issues" and where detailed monitoring will not be possible either due to lack of support from the householder or for financial or time reasons.

The detailed monitoring of households (with monitors on each circuit and each plug) is very cost and time intensive and is likely to mean sample sizes for individual studies are small. Monitoring of the whole household at the meter level is much easier and allows for larger sample sizes although the detail of data collected is much reduced. To provide advice on electricity usage behaviour it is necessary to discover appliance and task level (e.g. cooking) detail from the coarse household data.

Zeifman and Roth [19] reviews the current state of Non Intrusive Appliance Load Modelling (NIALM) which is the technology to identify individual appliances from the overall electricity usage. The conclusion from this review is that, currently, no approaches are suitable for detecting all kinds of appliances.

Chang et al. [91] is a typical study investigating the detection of appliances making use of the overall power usage only - without the need for intrusive appliance level monitoring. This study focuses on the signature of the turning on of the appliance rather than on the total appliance usage load. Like a lot of NIALM studies, some success has been demonstrated in detecting particular appliances but often only in a laboratory environment where the researchers know the actual appliances they are searching for from a short list and/or by making use of intensive monitoring. This monitoring may be at a very high frequency (sub-second) or by measuring multiple items (such as reactive power). This detailed information will not be available from smart meters rolled out across the UK and, while interesting, are seen as of little benefit when considering the wide population.

Firth et al. [92] has taken an intermediate approach to identifying appliances in that they define four classes of appliances: continuous such as clocks which cause a continuous electricity load, standby appliances which always draw some load when plugged in but also increase their load when actively in use, cold appliances such as fridges which are continually on but cycle through periods of cooling which draw a load and then idleness, and active appliances such as kettles that draw a load only when in active use. The work uses household level monitoring at a five minute interval as the basis for disaggregating the overall electricity usage into the four classes defined. Conclusions drawn from the study include the observation that overall electricity use increased over the two year monitoring period and this emphasises that a household's electricity usage cannot be seen to be constant (with added cyclical variations and noise) over a period but is likely to alter due to household changes such as number of members and due to changing sets of appliances (e.g. purchase or disposal of appliances).

Raine [93] have created the graph at Figure 2.5 showing appliance usage variation across the day for UK households using data from Stamminger et al. [94]. The data is extracted from a European wide study and then applied to various countries with the UK information displayed. Some of the classes of appliance conflict with other studies (e.g. the usage of air conditioning in domestic homes which is relatively rare in the UK) and suggests that the extrapolation of European wide figures to specific countries has made use of some invalid assumptions and that particular countries (e.g. the UK) may be grossly different from the European average. However, the study provides a view of what may be possible in changing the total electricity usage at certain times of the day by moving the use of certain types of appliances from the peak time to other times of the day.

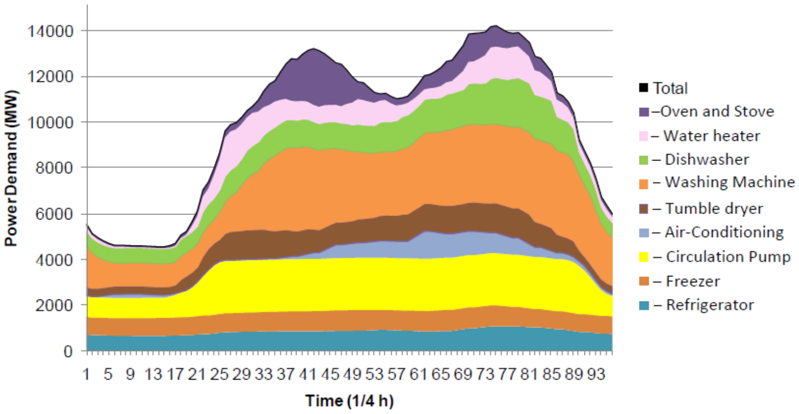


Figure 2.5: Average usage of appliances in UK across the day

An alternative approach is to identify individual appliances in use based on the characteristics of the stream of meter readings and then to combine the profiles of the appliances into an overall house profile. Successful implementation of such an approach often requires extensive time for training of the data mining algorithm to learn the attributes of the various appliances and normally requires the input of household members to label particular appliances within the meter data stream. These requirements can restrict the size of the population analysed. The approach is bottom up and complements the method detailed in Chapter 4 which is applicable to large volumes of households and can be seen as top-down. Lines et al. [95] details a UK based implementation of an appliance detection (drawn from a set of known appliances) approach.

This area of research has attracted extensive interest and forms the basis of a 2013 Kaggle data mining competition with a prize of $25,000 [3]. Most previous work has concentrated on analysing the power usage patterns while the competition includes the novel approach of making use of the electromagnetic interference that each domestic appliance generates and relies on the assumption that the same appliance will always produce the same (or very similar) electromagnetic interference pattern.

[3] http://www.kaggle.com/c/belkin-energy-disaggregation-competition

2.6.2 Pricing Signals

The impact of price signals (e.g. Time of Use pricing) on household behaviour has been researched in a number of studies and the results seem to differ from study to study. For example, Allcott [96] showed that US households reacted to increases in peak time prices by reducing their electricity usage. The study also found that this was, in general, an absolute reduction in usage, and not a shift in time of usage from peak to off-peak. However, other studies have shown little or no response from consumers to price signals.

Kirschen [97] takes an economic theory approach to the application of variable pricing as a means of altering household's electricity demand and considers various methods of increasing the price elasticity of demand for electricity. This work suffers from the common economist approach of considering all consumers as rational and with sufficient time available to make the appropriate rational decisions based on the price information provided. In practice, most householders are time poor with many calls on their attention and the most efficient selection of electricity usage is low down on their list of priorities. It is likely that any successful price driven interventions for demand side management will be directed at intermediaries (e.g. an automated agent working on behalf of the household) rather than directly at the householders.

Mahmoudi-Kohan et al. [98] proposes clustering households into groupings which are then offered different pricing models to maximise the overall profit for the utility. While not related to demand side management (no modification of behaviour by the customers is considered), the approach is likely to be used by utilities making use of the results from this thesis and similar work.

Chicco et al. [99] takes a conventional approach of clustering households into a few archetypical groups but extends the results to consider offering differing tariffs for customers and suggests a possible charging approach that could be used by utilities. This demonstrates a use of the clustering of households and could be adapted to apply to the clusters found using the techniques in this thesis.

An approach to changing behaviour by using social norm theory is reported by Allcott [100] which provided households with details of their neighbour's electricity consumption. One conclusion is that sending energy usage reports provides a similar change in behaviour as would result from a long term price increase of 5%. This work shows the importance of considering alternative interventions to the variable pricing model and also fits with the work in this thesis by emphasising the importance of efficient selection of groups to target for particular interventions. For example, providing energy reports relating to neighbours who are similar to a given household (e.g. retired couple) is likely to be more effective in changing behaviour than sending reports from a random selection of geographically close neighbours.

There have been many studies in the field of Economics which show that the behaviours of individuals in real life often doesn't follow the general economic assumption that people will maximise their individual utility when making choices. Various alternative theories have been suggested including that of Prospect Theory [101] which considers that people make different decisions under conditions of risk (which may explain the attractiveness of gambling and insurance) than they might do when purely considering utility. This has application to effectively influencing individuals to undertake activities for the benefit of the electricity network. However, it is likely that a lot of the "failures" of price incentive and similar studies to achieve a significant change in behaviour are due to the relative unimportance that individuals place on using their electricity effectively when compared to the other pressures that they may be under to make decisions (e.g. TV schedules, family activities, personal relationships).

2.7 Motif Detection

The electricity meter data stream from a household can be considered as a graph of usage against time and regular activities (e.g. cooking) can be seen as similar shaped usage patterns. Short patterns that repeat within the data are defined as "motifs" and detection of these motifs, and their timing, can lead to understanding of behaviour within the household.

A number of different problems have been addressed in the literature relating to motif detection. These include:

- Finding particular known motifs within a long time series
- Finding unknown motifs that repeat within the time series
- Determining the most common motifs within the time series
- Considering the warping of motifs in different ways so that two patterns may be assessed as the same motif even though one may be “stretched” in length or amplitude.
- Finding motifs within a large dataset where there are restrictions on CPU time or computer memory that is available.

The problem was initially stated by Lin et al. [102] who distinguished between the problem of efficiently finding defined patterns within a dataset (which the authors felt was generally solved) and the problem of finding repeating patterns within the data which were not previously known and which the authors named as motifs.

Previous work from Das et al. [103] had addressed a related problem of finding rules that relate patterns in one time series to patterns in another. [104] considers the problem of finding temporal relationships between primitive patterns in time series in a generalised way. The term “primitive patterns” can be defined as motifs.

A significant amount of work has been done in the area of DNA pattern detection and within textual analysis which can be applied to the motif finding problem. If real valued data (as from electricity meters) can be represented by a series of discrete characters (letters) then the techniques of DNA and textual analysis can be applied [105].

The SAX (Symbolic Aggregate approXimation) technique provides for symbolic representation of time series data and thus provides access to bioinformatics and text mining techniques [106]. The original work has been extended to iSAX [107] which provides support for large volumes of data. One of the techniques available through SAX is the detection of motifs in a data stream. The “holy grail” of electricity meter analysis is to detect individual appliances from the time series data and, while there has been some success by researchers around the world, in general there is a need for detailed monitoring (e.g. very frequent) or training of the software to accurately detect appliances. The focus in this thesis is on finding interesting, repeating patterns of behaviour (e.g. cooking or washing) rather than individual appliances.

No applications of the SAX technique to domestic electricity data have been found in analysis of the literature although work on wholesale energy price time series has been published by Mori and Umezawa [108].

Much of the motif-related literature has focused on finding efficient ways of finding motifs given computing time and memory space restrictions. These have not been directly addressed in this work but may become important when considering the large volumes of data provided by smart meters.

2.8 Clustering using Behavioural Traits

The degree to which households demonstrate varying behaviour could be an indicator as to how receptive they may be to modifying their behaviour to take advantage of some incentive offered to them by the utility company or similar bodies.

Eagle and Pentland [109] has introduced the concept of "reality mining" which is defined as the sensing of complex social systems and using various monitoring tools to detect social patterns within routines that are detected. These apply at various time-scales including daily (e.g. getting up, eating lunch), weekly (Saturday sports) and annual (Christmas holiday family visits). While Eagle and Pentland [109] makes use of mobile phones as sensors to detect the routine activities, the concept equally applies to making use of the household meter data readings to detect routines within the household. The paper discusses the amount of entropy in people's lives where "people who live disorganised lives tend to be more variable and harder to predict".

Some researchers have explored "concept drift" which detects the way in which memberships of clusters change over time [110]. While there is likely to be some change of behaviour over time as the household members change (e.g. retirement, selling of house to other occupants) the assumption made in this thesis is that the “real” clusters within the data, in each of the time periods, are the same. This approach is different to that taken with the concept drift work which measures the degree of change over time.

Meo et al. [111] is an example of using the frequency of patterns within a sample to derive classification rules. The regular patterns are used to generate a probabilistic model which provides a probability that a particular pattern will occur within a random example of the given class. This approach allows for all patterns to be used at the same time in the classification. However, the time of occurrence of the patterns and the variability of that timing being used as a feature in its own right is not considered in this work.

2.9 Analysis and Gap Detection

There is extensive existing work on creating average daily load profiles and then using a large selection of different clustering algorithms to form groupings of similar households. This work has been applied to various geographic samples of households with differing results on the best clustering algorithms to use. The evaluation between the results from the differing algorithms has generally focussed on how close the members of a given cluster are to each other (from a daily load profile point of view) and how different the archetype for each cluster is from the other archetypes.

There are various factors that apply to specific countries and, in many cases, the results from one country may not apply to another country. In particular, the infrequent usage of air conditioning units in the UK and the extensive usage of gas for space heating, means that the results from other European countries (particularly southern European) do not map well onto the UK situation. There have been fewer studies of load profile clustering using UK data and, as the results from other countries cannot easily be applied to the UK, more studies using UK data are necessary.

When clustering similar households no studies have considered the results in terms of what would be useful from a marketing point of view. Good separation between clusters and good similarity within a cluster have been well studied and CVIs that measure these statistics have been defined specifically for load profile clustering (e.g. CDI and MIA CVIs). However, more extensive evaluation of the clustering results is needed to measure the usefulness of the results as appropriate for driving a behaviour modification programme.

The detection of motifs within time series data has been applied to many different application domains but there has been little application of motif detection to electricity meter data from domestic households. What work that has been done has concentrated on activities in the electricity network at a higher level than that of domestic households (e.g. at the local substation level).

Previous data mining work on clustering domestic households on the basis of their variability in behaviour has only concentrated on calculating ratios of usage between different parts of the day (e.g. night versus daytime usage) and has not considered changes in behaviour from day to day. While there has been some work on understanding the variability of behaviour using face to face interviewing and questionnaire responses, there has been no application of data mining techniques to the meter data stream to find variability measures of household behaviour.

Areas of work identified and then addressed in this thesis include the following:

2.9.1 Load Profile Clustering Evaluation Measures

The incorporation of additional criteria into the evaluation measure for comparing between differing clustering approaches has not previously been addressed and may produce results that are more useful for electricity industry professionals. The successful definition of evaluation measures that produce results that are more suitable for behaviour modification exercises will allow for more effective selection and deployment of DSM programmes and, hence, more effective modification of electricity usage patterns for the benefit of the electricity network. Successful DSM programs can have major impacts on addressing the upcoming challenges to the electricity industry. This area of research is addressed in Chapter 4.

2.9.2 Assessing Load Profile Clustering Algorithms in a UK Environment

While there has been extensive work on daily average load profiling clustering across the world, a lot of this work is not directly relevant to the UK due to specific geographic and historical reasons (e.g. the UK weather, the availability of North Sea gas, little usage of air conditioning). Testing the available clustering algorithms using UK data is useful for selection of the most appropriate clustering algorithms for future UK based work. In addition, evaluating the clustering of the UK using a composite measure (as suggested above) provides useful guidance for selecting the most effective algorithm and set of parameters. Chapter 4 uses UK data and the proposed composite measure to provide guidelines on the most effective clustering algorithms.

2.9.3 Finding Motifs using Symbolisation

A method of effectively finding repeating activities within a household provides industry experts with very useful information to allow them to design effective DSM interventions that remove or change the demand into a pattern that is more efficient for the overall network. For example, finding a particular motif and encouraging the household to undertake the activity at a different time (e.g. by moving the use of a dishwasher to overnight) can smooth the overall electricity demand.

Using symbolisation techniques to represent the electricity meter data and hence to identify similar shapes (motifs) within the data has not previously been applied to domestic electricity meter data. The approach to finding motifs is tested in Chapter 5.

2.9.4 Using Variability of Behaviour for Clustering

Little work has used variability of behaviour as a basis for grouping households together and no work has made use of the variability in timing of motifs as the basis for clustering similar households together. Finding clusters of households with differing behavioural characteristics (and, in particular, differences in timing of regular activities) can allow specific DSM initiatives to be defined that address households with similar degrees of variability. For example, being able to target the households showing the most variability with an incentive that encourages them to make their activity a lot more regular (e.g. particular pricing incentives) may allow network managers to modify behaviour to better smooth overall demand. The ability to group households by their degree of variability provides increased information about the households and will allow better targeting of initiatives to modify behaviour. Chapter 6 evaluates the use of variability in timing of motifs as the basis for clustering of similar households.

2.10 Summary

This chapter has introduced the issues affecting the electricity industry, both across the world and specifically in the UK. While there are many future challenges in meeting the legal, political and environmental requirements, there is the opportunity to make use of the newly available stream of data from smart meters to help address the challenges. The domestic usage of electricity in the UK is 30% of the total usage and thus represents a significant proportion of the total demand.

Industry commentators are agreed that the implementation of incentives that change domestic behaviour patterns is a key part of the solution needed to meet the challenges that the electricity industry faces. Previous behaviour modification initiatives have demonstrated some success but, so far, not sufficient (typically only a few % improvement) to provide a major impact on the problems. To gain more success, better targeting of the incentives is required and thus better knowledge of the households and their behaviour is needed.

This chapter has considered the approach taken in the field of marketing and has determined that the guidelines for good segmentation of customers can be applied to the field of electricity customer clustering.

The history and current state of the art of clustering similar households on the basis of their electricity usage is considered and areas where better guidelines on appropriate algorithms and parameters are needed have been identified. The particular requirements of the UK and the need to test possible approaches using UK data is important.

The approach to be taken in the remainder of this thesis is as follows:

- Using UK data, apply the clustering algorithms used in previous work to determine a small number of archetypes for UK households based on the shape of their daily usage patterns.
- Apply the segmentation guidelines from the marketing literature to define a novel composite measure that evaluates the archetypes using criteria wider than those used in previous work. The composite measure is then used to determine the most effective algorithms to use when clustering UK households using daily average load profiles.
- Use a symbolisation technique to find motifs that occur regularly within a household. Apply an evaluation method to assess the effectiveness of differing approaches to defining motifs with the goal of finding a reasonable number of motifs for each household (not too few, not too many). This work then leads to a set of guidelines for the most effective method of finding motifs and the appropriate parameters for those methods.
- Using the findings from the motif detection investigation and the clustering algorithm investigations, discover how the timings of motifs vary from day to day within a household and use this degree of variability as the basis for clustering the households together to produce new or complementary archetypes which are then evaluated using the marketing inspired composite measures.

This approach builds on the existing published work on daily average load profile clustering but extends the work by using an enhanced evaluation method and tests the results using UK data. This initial clustering on load profile shape is then extended by developing a method of finding repeating activities within a household and using this added information to produce archetypes.

CHAPTER 3

Methodology

3.1 Research Steps

This section details the steps that will be taken to address the research questions posed in Chapter 1.

To investigate the behaviour of households using their meter data, it is first necessary to collect an adequate quantity of meter data that can be used in the analysis. This thesis makes use of the North East Scotland Energy Monitoring Project (NESEMP) data which have been collected over a period exceeding a year and contains data from 380 households, although not all households have a complete set of readings.

The data used for the study is collected from Scotland and samples households where one of the household members is employed by the local government. It is therefore not a completely representative sample of the UK population and other behaviours may be found in households with unemployed or retired household members. The approach to finding regular behaviour within the meter stream still applies to other kinds of households.

The research approach is to first define a baseline of performance by clustering households on the shape of their daily usage profiles. This clustering exercise has previously been done by numerous researchers using datasets from various countries and using an assortment of different clustering algorithms. The approach is well established and will be repeated using the NESEMP dataset and a selection of clustering algorithms.

To evaluate the baseline performance, a novel method of measuring the quality of the partitions found will be defined. This novel method produces a single statistic that is built up from a number of component statistics, each of which are widely used approaches to measuring partitioning results. The novel component is to combine the composite statistics into a single measure that provides an arguably wider view of the quality and usefulness of the partitions. The composite measure is compared with well known CVIs widely used in the area of electricity load profile clustering to compare and contrast the novel approach with the "traditional" method.

Next the research considers how best to find motifs within the stream of meter data. A symbolisation approach (SAX) widely used in other application domains is selected. While applied to many different problems, the SAX approach has not previously been used with domestic electricity data.

Various parameters can be adjusted (e.g. length of motif, size of alphabet used in the symbolisation) and an inspection approach is defined to allow selection between the parameter settings. The criteria used for selection of "best" settings could be debated and, to accommodate alternative views on appropriate evaluation criteria, the method of selecting the criteria and the way in which the results are compared is laid out in detail. Another researcher (with different criteria) can apply the same approach to selecting parameters and obtain the optimum set that matches their criteria.

3.1.1 Load Profiles

When data is collected at a high frequency, this produces a large number of dimensions for the clustering exercise (i.e. the number of samples per day). For the five minute sampling period, the data consists of 288 attributes which are used as the basis of clustering of full day average profiles.

While a lot of previous work has made use of the absolute values of the meter readings (often normalised so that the readings fit within a 0-1 range), the analysis of the differences in meter readings is of more interest as this reflects the changes in usage resulting from turning an appliance on or off. As the DSM interventions are intended to influence appliance turning on/off behaviour, the usage of the difference data as the basis for the load profiles is more useful and is used in the analysis.

3.1.2 Finding Motifs

Using the differenced meter data (i.e. the differences between successive readings) short motifs (of up to an hour in length) are found. Rather than defining the motifs as representing a particular activity (e.g. usage of a toaster) and then searching for that motif within the data stream, the approach is to find patterns in the data that are unknown before the analysis. The interpretation of what the motif represents is not possible in the absence of monitoring of the household, such as by usage of an activity diary. It may be surmised that certain patterns correspond to regular activities but the analysis does not make use of these assumptions and treats the motifs as unknown repeating patterns.

The motif finding approach uses a process of approximation to find motifs which are similar to each other and treats these similar motifs as representative of the same household activity. The size of motifs ranges from 20 minutes to one hour in length.

A number of motifs are detected which do not represent interesting behaviour. For example, a motif showing no activity in the household (such as may be seen when the house is empty) is of little interest to utilities intending to implement DSM interventions and is excluded from the analysis. An automated process for assessing whether detected motifs are “interesting” is implemented and non-interesting motifs discarded from the further analysis.

3.1.3 Variability

One aspect of the data that is lost during the creation of household load profiles is the variability of a household’s usage of electricity from day to day over the period of interest.

Once a significant number of motifs are detected, the timing of the motifs is examined and variations in timing of the occurrence of a particular motif from day to day within a household are taken as a measure of the variability of behaviour within that household. The most frequently occurring interesting motifs are examined for variability in timing and the measure of variability is used as input to clustering exercises.

3.1.4 Evaluation

The analysis described above leads to the generation of various statistics that describe aspects of the household's behaviour and which are used as the attributes input to differing clustering algorithms.

The information collected through the NESEMP questionnaires and the related demographic information is compared with the results from the clustering on variability of motifs. Any areas where the variability of motifs results match to a particular questionnaire response or demographic characteristic are identified.

The clustering algorithms, and the parameters for the algorithms, are assessed using cluster validity indexes that have been used previously in electricity load profile clustering. In addition, a measure based on criteria from marketing theory is defined and the results assessed using those criteria.

3.2 Clustering

The load profile data from a 3 month period are processed so that an average load profile for each household is generated.

The analysis focuses on the full day period. The analysis techniques can easily be applied to shorter time periods of interest (e.g. the UK peak time of 4pm to 8pm) when relevant for a particular intervention and are transferable to different subsets of the data such as differing days of the week, seasons, or time periods of the day. The approach also allows for additional data to be incorporated (e.g. external data on weather) which can then be used to filter the data analysed.

Using a defined number of clusters, the clustering is undertaken using the following algorithms, selected based on a review of literature detailed in Section 2.4.3:

- kmeans
- Fuzzy cmeans
- Self-organising maps
- Hierarchical clustering
- Random Forests
- Gaussian Mixture models

More details on each of the algorithms is provided below:

Kmeans

Kmeans is a well known algorithm that is used in a number of electricity load profiling studies. The algorithm requires a number of clusters as an input parameter (k) and works by randomly selecting an initial k locations for the centres of the clusters. Each household is then assigned to one of the centre locations by selecting the centre nearest to that household's average profile. Once all the households are assigned, each collection is considered, the new centre of the allocated households is calculated, and the centre for that cluster is reassigned. The households are then reallocated to their new nearest centre and the algorithm continues as before until no changes are made to the allocations of households for an iteration [87].

Fuzzy Cmeans

Fuzzy cmeans provides an extension of the kmeans algorithm by allowing partial membership to more than one cluster. The algorithm provides additional output showing the degree of membership that each household has of each of the derived clusters [112]. For the analysis in this thesis, each household is assigned to the cluster for which they have the highest degree of membership. A fuzziness factor of 2 is used.

Self-Organising Maps

The Self-Organising Map (SOM) is a neural network algorithm that can be used to map a high dimensional set of data into a lower dimensional representation. In this thesis, the mapping is to a two dimensional set of representations which are arranged in a hexagonal map. Each sample (i.e., the average load profile for a given household) is assigned to a position in the map depending on the closeness of the sample to the existing households assigned to each position (using a Euclidean measure of distance).

The SOM algorithm allows for a continuous input space (e.g. the set of average load profiles for the households) to be mapped to a discrete output space. This discrete output space can be viewed as a set of clusters where the households assigned to each of the discrete positions are the members of that cluster. By setting the output space to have eight locations, the results can be compared to those from other clustering algorithms (e.g. kmeans) which form eight clusters. The SOM algorithm consists of a number of phases as follows:

- Initialisation. Each node in the output space is assigned a random sample from the dataset.
- Sampling. A random household is selected.
- Matching. All the nodes in the output space are compared to the sample. The node with the weight that is closest to the sample is selected as the "winner".
- Updating. The winning node's weights are updated with the effect that it moves nearer to the sample randomly selected in the previous steps. In addition the other nodes also have their weight updated with those closest to the sample updated more than those further away (using Euclidean distance).
- Iteration. A further random household is selected and the process returns to the sampling phase. This is repeated until the output space doesn't change between iterations.

Initially the nodes are assigned at random but, over time, the map produces an arrangement where similar load profiles are placed closely together and dissimilar load profiles are placed far apart [113]. The application of the SOM to electricity load profiles is described in [75].

Hierarchical Clustering

Most of the published load profiling work uses hierarchical clustering and this approach has the benefit of providing easily understood rules for cluster membership. The algorithm uses a dissimilarity matrix for the households and, starting initially with each household in its own cluster, proceeds by joining clusters which are most similar. Thus, at any point in the process, there are a set number of clusters defined and the process can be stopped to provide the desired number of clusters [114].

The euclidean distance is used when creating the dissimilarity matrix. Various agglomeration methods are available with the “average” and “Ward” methods being most commonly used in electricity load profiling [115]. The agglomeration method provides the way in which clusters are combined. The average linkage assesses the average distance between all points in the clusters and selects the smallest average distance when deciding on households to merge with the existing clusters. The Ward method minimises the sum of squares of possible clusters when selecting households to combine.

Gaussian Finite Mixture Models

Finite mixture models assume that each of the clusters in the population is represented by a probability distribution (a Gaussian distribution for Gaussian models). Selecting a suitable number of clusters and clustering algorithm can then be done from amongst the models providing possible choices of cluster numbers and parameters [116].

The algorithm generates a number of different possible models of fitting Gaussian distributions to the data and then uses the Bayes Information Criterion (BIC) to select the most appropriate fit amongst the models. The BIC is the value of the maximised log-likelihood with a penalty on the number of model parameters and allows comparison of models with differing sets of parameters [117]. In general, the number of clusters (number of distributions) is one of the model parameters that varies such that the most appropriate combination of number of clusters and model is selected from the differing BIC values. To allow comparison with the other algorithms, the number of clusters is constrained and the algorithm selects between different models. This algorithm is implemented using the R package mclust [118].

The approach relies on the assumption that the population of interest (e.g. the average load profiles) consists of a number of different subpopulations, each of which can be described by a Gaussian probability distribution. Various different sets of parameters can be used to describe the distributions and can be chosen to maximise the likelihood of a given set of data fitting the possible sets of parameters [119].

Random Forests

Breiman [120] proposed that a number of decision trees should be created using differing random starting points. Generally each node of the decision tree can be split such that the best split is calculated using all the variables. However, the random forest approach selects a random subset of the variables at each node and then makes a split based on the best result using just the variables selected.

Random forests are collections of decision trees originally defined for classification tasks. They can be extended to an unsupervised (clustering) task by generating extra synthetic data which is then combined with the original data with each being labelled differently (e.g. original data as A, synthetic data as B). The random forest procedure can then be run to classify the combined data. One output from the procedure is the proximity matrix which gives a measure of how often real data points are found in the same terminal node of the decision trees. A higher value shows better proximity (and hence, more similarity between the records). This algorithm is implemented using the R package randomForest [121].

The proximity matrix can then be interpreted as a similarity matrix between the different household data (e.g. load profiles) and a clustering algorithm such as Partitioning Around Medoids (PAM) can make use of the similarity matrix to determine the desired number of clusters. PAM is similar to the kmeans algorithm but makes use of a similarity matrix (based on any criteria) rather than a Euclidean distance measure (as with kmeans) and selects records from within the population as representatives for each of the defined clusters rather than generating a new "centre" (as with kmeans) that may not exist within the population [122].

Cluster Algorithm Tuning

The tuning of each of the cluster algorithms to give the best possible performance (however this is defined) is not a goal of this study. The thesis uses the differing clustering algorithms to provide comparisons between the alternate ways of clustering the households (e.g. using load profiles or motif variability) and the relative, rather than absolute, performance of the algorithms is valid for comparing the approaches. However, there is the possibility that a badly mistuned algorithm can give misleading results and therefore lead to an invalid comparison between the approaches and, thus, some tuning has been applied to avoid badly mistuned application of algorithms.

A common issue with a number of clustering algorithms is the randomness inherent in the process. For example, the kmeans algorithm selects a random centre for each of the k clusters and then works forward from this random starting position. Different random starting situations can lead to different final sets of partitions (possibly due to the algorithm reaching a local optimum rather than the global optimum). To avoid this situation, algorithms such as kmeans and fuzzy cmeans are repeated many times with the best overall result being taken as the final partitioning.

Extensive research has focused on finding the best starting points for the kmeans cluster algorithm (e.g. [123]) but this level of tuning has not been implemented in this thesis with the many repetitions of random starting points being considered sufficient to avoid badly mistuned application of the algorithm.

When considering the motif variability data, three datasets are compared using the six selected clustering algorithms. The optimal tuning for each algorithm is likely to be different for each of the datasets but, to allow comparison between the datasets, the clustering parameters are kept consistent across the datasets. Thus, the partitions found for a given dataset and clustering algorithm could possibly be improved (as measured using a conventional CVI) but, for comparison reasons, is not done.

Ensemble Clustering

Ensemble learning is the process of using a number of different clustering algorithms (or set of parameters for the same algorithm) and then combining the clustering results to get an overall result. Various methods of combining clustering results have been researched (e.g. [124]).

Consensus clustering takes the information in a number of different partitions of the same instances and combines this information to produce an overall summary set of partitions that best represents the information in the underlying clusterings. The method of doing this combination and the measurement of "best" provide a number of different ensemble clustering approaches.

This thesis makes use of the R package "clue" [125] which provides a number of different approaches. The method selected is "HE" which provides for hard partitions to be created based on Euclidean distances between the component clusterings. The dissimilarity between the various component clusterings is minimised to find the overall "best" consensus clustering.

Attribute Selection

Various socio-economic data collected via questionnaires from the households under study is available for comparison with the clustering results presented in this thesis.

For each item of socio-economic data, it is useful to understand which of the various attributes used for the clustering best splits the households into partitions that match the questionnaire responses for that item. The attribute found can be seen as the most important in influencing the split of households into the groups suggested by the questionnaire response (e.g. household size).

To find the most "important" attribute, a random forest approach is taken. For the data set used for the clustering (e.g. the variability of motif data) a number of classifications are undertaken using a hierarchical algorithm. For each of the classifications, a random selection of attributes are selected. This results in a number of different trees and the associated error rate of each given tree. By considering each of the attributes and comparing the average error rate for the trees including the attribute against those omitting the attribute, a value for the average "improvement" in error rate when including the attribute can be calculated. After considering all attributes, the one with the greatest improvement in error rate is taken as the most important in predicting the given questionnaire response.

Being able to predict questionnaire responses from membership of a household in a particular archetypical cluster would be a powerful way of understanding more information about a household. If it was possible to accurately infer various characteristics of the household from the cluster in which they fall based solely on their electricity meter usage, then DSM initiatives aimed more accurately, and using the inferred knowledge, would be likely to be more successful.

3.2.1 Cluster Validity

To assess the quality of the clusters derived from one clustering exercise compared to those derived from a different exercise, the widely used CDI and MIA measures (as defined by Chicco [14]) are used. Lower values for the CDI and MIA measure denote "better" solutions.

These are defined by Chicco et al. [65] using the following description:

The data to be clustered consists of M records numbered as m=1,..M. Each record has H features numbered as h=1,..H.

The data is clustered into K clusters (numbered as k=1,..,K). Each cluster has R_k members where $r_{(k)}$ is the rth record assigned to cluster k and $C_{(k)}$ is the calculated centre of the cluster k.

The distance (d) between two profiles is defined as:

$$d(m_i, m_j) = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(m_i(h) - m_j(h))^2} \tag{3.2.1}$$

where $m_i(h)$ and $m_j(h)$ are the hth attributes for two records, m_i and m_j.

The "within set distance" $\hat{d}(S)$ of the members of a set, S with N members (s_j where j=1,..,N) is defined as:

$$\hat{d}(S) = \sqrt{\frac{1}{2N}\sum_{n=1}^{N}\sum_{p=1}^{N} d^2(s_n, s_p)} \tag{3.2.2}$$

The MIA gives a value which relies on the amount by which the members in the cluster are close together. It is calculated by using the distance of each member of the cluster from the representative profile for the cluster. The representative profile is defined differently depending on the clustering algorithm used. A lower value for MIA suggests a better clustering solution.

$$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\sum_{r} d^2(r_{(k)}, C_{(k)})} \tag{3.2.3}$$

The CDI depends on the distance between the members of the same cluster (as for the MIA) but also includes the distances between the representative load profiles for each cluster. This therefore measures both the homogeneity of the clusters and the separation of each cluster from the others. The possible values for the CDI measure range from 0 upwards with no limit to the maximum value. A lower value for CDI suggests a better clustering solution.

$$CDI = \frac{1}{\hat{d}(C)}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\hat{d}^2(R_k)} \tag{3.2.4}$$

where C is the set of cluster centres and R_k is the set of members of the kth cluster.

The CDI and MIA measures have been widely used in electricity load profile work and will be used as baseline measures against which the results in this thesis can be compared. The composite measure that will be developed and used to assess the differing algorithms is detailed in Chapter 4.

To assess the consistency of clustering solutions, the differing arrangements of households into partitions are compared.

The Rand index compares the different pairs of samples (i.e. each possible pair of households) and assesses the number in which each pair are in the same partition in the two differing clustering solutions, the number where each member of the pair are in different partitions in both solutions, and the case where the members are in the same partition in one solution but differing partitions in the other solution. This information is used to generate a measure ranging from 1 (complete agreement) to 0 (chance distribution). Thus, the higher the value for the Rand index, the better the two clustering solutions agree.

The corrected Rand index ([126] builds on the original work but adjusts the calculated value for the expected matching that would occur in a random arrangement. The corrected Rand index takes values from 1 to -1. The higher the value for the corrected Rand index, the better the two clustering partitions agree. The consistency of the clusters obtained from the differing clustering algorithms is used as a measure of the quality of the results with more consistency between the results suggesting a more useful method of identifying the clusters.

When comparing between differing clustering algorithms, the approach with the higher values for the adjusted Rand index provides the more consistent set of cluster memberships and is judged to be the better solution.

The clustering approach is assessed in terms of marketing theory to understand the usefulness of the results. This means that clusters should be assessed using more criteria than the generally used "tightness" of the clusters and "separation" between clusters with the inclusion of other criteria such as "understandability" (whether the clusters make sense in real world terms).

3.3 Motifs

To find the motifs within the data, each period of interest within the day for each household is examined by taking a moving window over the period. The subset of the meter readings within the moving window is then converted into a string using the SAX method and the resulting string is stored in a MySQL database. Next, the window moves on by one reading and the conversion into a SAX representation string is repeated. Thus, using a SAX alphabet size of six and analysing the 4pm to 8pm period (a total of 48 x five minute readings) will result in 43 motifs stored for each day for each household. See Figure 3.1 for an example of how the SAX representations are defined. The top graph shows the five minute readings for a four hour peak period. A sliding window of six readings (30 minutes) is taken across the peak period with the first two windows and the last one shown. Each window is normalised (using Z normalisation) within the values in the window and then translated into the SAX alphabet representations (using an alphabet size of five) as shown at the bottom of the diagram. The example shows a 4 hour peak period but the approach applies to any size for the period of analysis. The analysis in this thesis uses the full day as the period of interest.

The example uses a SAX alphabet of five (i.e. the letters "a" to "e") to represent the motifs. Five is selected as a reasonable compromise between having too few letters and thus not detecting changes in meter readings and having too many and thus generating too many patterns which may not repeat. This parameter is explored in Chapter 5 to assess the sensitivity of the analysis to different settings for alphabet size.

The motif size used for the example is six which corresponds to a 30 minute period. The UK electricity settlement market uses a 30 minute period and thus six is a useful starting point for investigation. 30 minutes is also a reasonable estimate for a time period that will include most household activities such as breakfast or showering. The appropriate setting for this parameter is explored in Chapter 5.

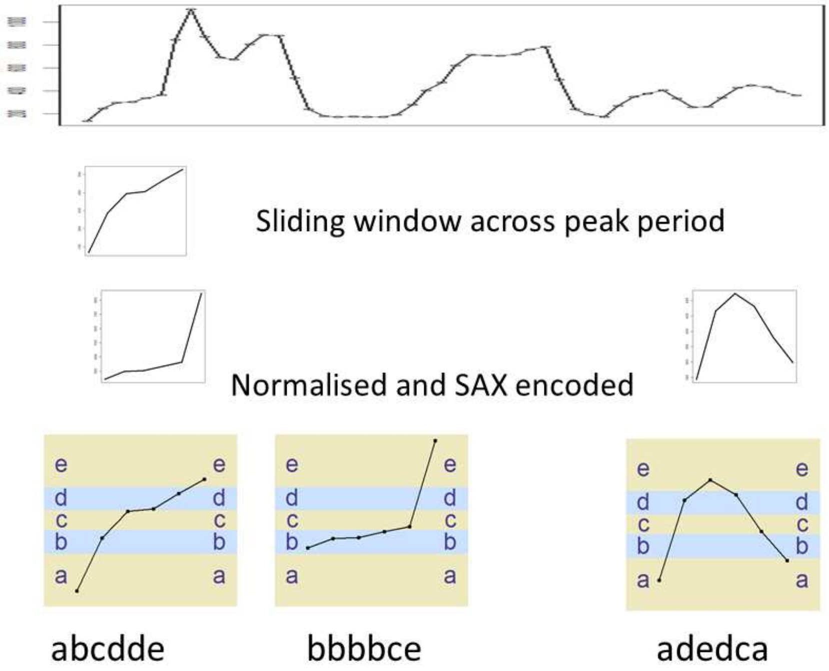


Figure 3.1: Process for creating SAX representative strings

Automatically detecting uninteresting motifs

Certain motifs are not interesting for analysis. For example, the motif showing no activity (i.e. a repeating constant usage) will be very common and is unlikely to be interesting for understanding how best to apply interventions to change behaviour. Therefore, an automated process to exclude uninteresting motifs from further analysis is defined in Section 5.3.

3.3.1 Variability

The averaging of all the daily load profiles for a household into a single representative profile for each household necessarily loses some information that could be an important differentiator between households. The variability between the household daily load profiles can give an indication of how much a household's behaviour varies from day to day and can be a useful indicator of the appropriateness of a household for a particular behaviour change intervention.

There are many ways of measuring the variability of households from day to day and this thesis explores the presence of motifs (and how the motifs vary in timing during the day) as an indicator of the degree of variability.

This is implemented by first finding all the motifs (matching certain restrictions such as length) within a given household's meter data stream. The top motif (the one that occurs most often) is further examined for the times when the motif occurs. The number of times it occurs, and the standard deviation of the times around the mean time, are calculated for each household. An example to explain this can be seen at Figure 3.2 where the meter readings during the peak period (4pm to 8pm) for one household for five weekdays are shown. The motifs are highlighted and the timing of occurrences of the motifs vary between approximately 6pm and 7:30pm. While this example shows just one week, the analysis uses all the available working days within the selected three month period.

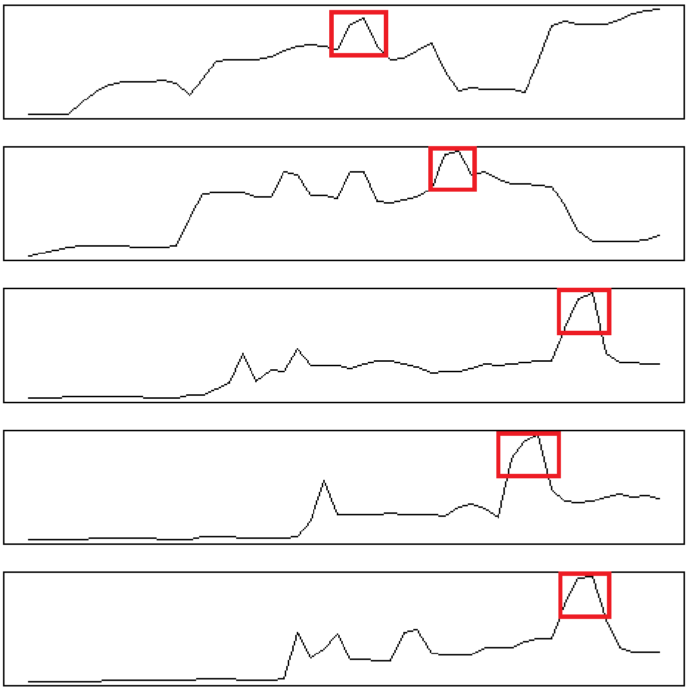


Figure 3.2: An example of one week's peak data showing motifs and variability in their timing

3.4 Data Sources

3.4.1 North East Scotland Energy Monitoring Project

The ongoing North East Scotland energy monitoring project (NESEMP) is examining the relationship between different types of energy feedback and psycho-social measures including individual environmental attitudes, household characteristics, and everyday behaviours. As part of this ongoing project, several hundred households are being monitored with the electricity usage recorded every five minutes using CurrentCost monitors [15]. There are various issues with the data including long periods of non-collection for some households and some households leaving the study during the monitoring period.

As part of the NESEMP, various demographic and attitudinal information was collected using questionnaires from each household. Some of this data is incomplete and some of the households did not provide full permission for usage of their data and, while meter readings exist for these households, their data has been excluded from the analysis.

Data Collection

Data is collected by a CurrentCost Envi [1] monitor by a clamp connected to the household electricity meter which communicates over a wi-fi network with the monitor. The sampling happens approximately every 6 seconds and provides a value for the current usage of electricity within the house, as measured by the clamp, and the current temperature in the room housing the Envi monitor.

The Envi monitor is connected to a CurrentCost Bridge [2]. Each reception of a sample from the meter clamp generates an input to the Bridge consisting of the meter reading and the room temperature. The meter reading data consists of the rate of usage at the time of the reading (i.e. an instantaneous sample). The information provided to the Bridge consists of this instantaneous reading and the instantaneous temperature reading from

[1] http://www.currentcost.com/cc128/xml.htm

[2] http://www.currentcost.com/product-bridge.html

the monitor.

The Bridge creates a summary of the electricity and temperature for a five minute period by averaging all the readings received during the five minute period and calculating the usage for five minutes that this averaged consumption rate would imply which is uploaded to Pachube. Any missing readings (e.g. due to transmission problems from the clamp to the monitor) are ignored (i.e. the average is calculated just on the transmissions received during the five minute period).

The Bridge communicates with Pachube across the internet using the household's broadband router approximately every five minutes and provides a summarised electricity usage figure and temperature. The summarised figure consists of the average of the meter readings collected during the previous five minute period. The posting of the data to Pachube takes an additional one to two seconds which is not included in the total usage figure.

An automatically scheduled job runs daily and downloads the data stored within Pachube to a .csv file on servers at The James Hutton Institute in Aberdeen, UK.

Pachube is an internet service that allows the creation and monitoring of the "Internet of Things". In this deployment, it allows for electricity monitoring devices to be connected to the internet and for data collected by the monitoring devices to be uploaded and stored on a database maintained by Pachube. Since the beginning of the NESEMP project the Pachube service has been taken over by Xively.com who continue to provide the service described.

Alongside the continual collection of meter data, various questionnaires have been completed by the households participating in the in the study. Some of the interesting information collected is used in the analysis in Section 6.5.

Data Cleaning

To allow analysis of the data, various data cleaning procedures are required which are summarised below:

- The data readings were nominally taken at five minute intervals but the technology used meant that the actual reading times slowly drifted over time with each reading being about one or two seconds more than five minutes apart. To allow comparison of households and days within a household, the data was aligned precisely to five minute times (e.g. 12 noon, 12:05pm, 12:10pm, etc.) using a process of interpolation between the actual readings.

- The technology used for collection of the meter data was sometimes faulty (e.g. the household broadband connection was down) and readings were missed for some five minute periods. These have been reconstructed from the actual readings using interpolation.

- Any partial days of readings at the beginning or end of the dataset are deleted to ensure that all days of readings in the data consist of a full set of readings.

- Daylight saving causes either an additional hour of readings (in Autumn) or a missing hour of readings (in Spring). These have been corrected for by interpolating the missing hour in Spring and by averaging the readings for the repeated hour in Autumn.

The Current Cost Bridge software makes an assumption about the maximum electricity load that resulted in readings of greater than 32767W being stored as negative numbers. To correct this problem a further step of replacing all the negative numbers with the correct positive value was necessary.

After removing data for households with insufficient readings, the data is loaded into a MySQL database and the readings are aligned with exact five minute boundaries (e.g. 1pm, 1.05pm, etc.) by interpolation between the actual readings. This is achieved by calculating the reading at an exact five minute point (e.g. 1.05pm) by considering the actual readings before and after that time and by calculating the reading such that the total usage over a longer period (e.g. an hour) is the same whether the interpolated readings or the original actual readings are used [127]. This results in a set of 288 readings (one for every five minute period in the day) for each day for each of the households in the database. Each day of sampling is labelled in a number of ways such as "working day" or "summer".

The sum effect of the above modifications is to create a collection of data for each household where the data consists of full days of 288 readings per day with the meter reading timings aligned on the exact five minute times.

The process for data manipulation is described in the following diagrams. Initially the data points in the time series are as per the stars on Figure 3.3. Each star represents the usage for the period for the five minutes up to that time-stamp.

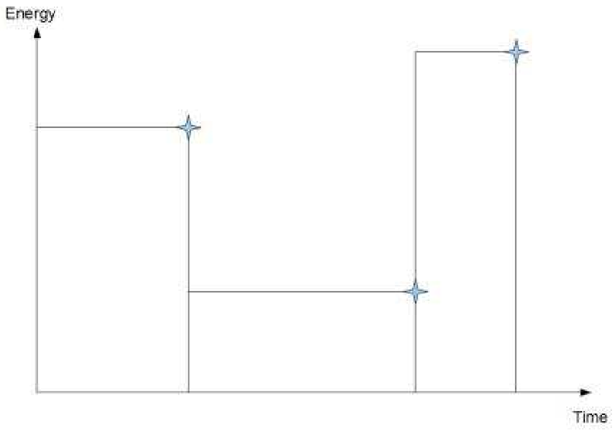


Figure 3.3: Original data points

The first step is to assume that readings are best represented by a time-stamp within the middle of each sampling period as per Figure 3.4. This is done by subtracting 2.5 minutes (150 seconds) from each time-stamp and producing the information shown in Figure 3.4. The justification for this step is that the five minute readings actually represent an average of the readings taken every few seconds over the previous five minute period and, thus, best represent the usage halfway through the five minute period.

Estimates next need to be made for the periods between the five point stars so that the points for the 12 point stars can be determined (i.e. the value for b can be estimated). See Figure 3.5.

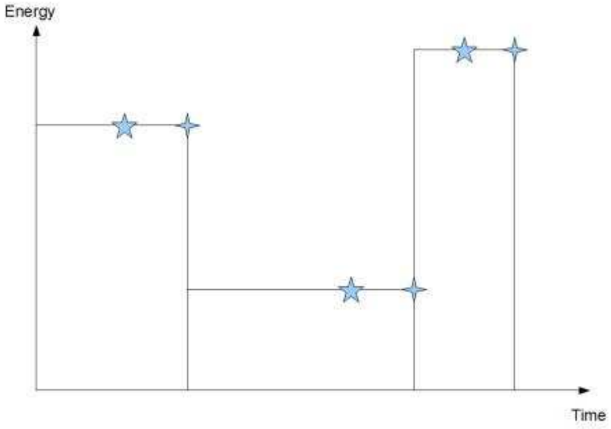


Figure 3.4: Times represented by the readings

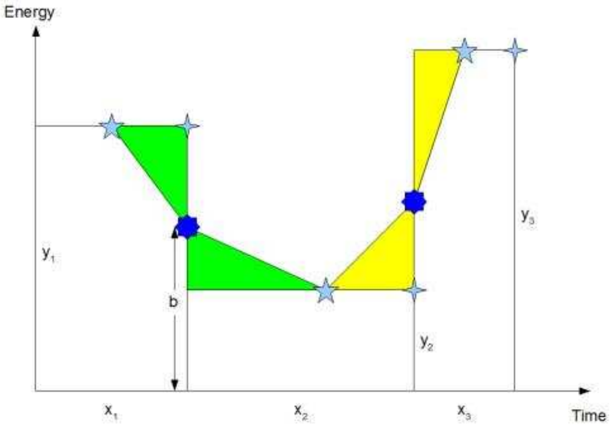


Figure 3.5: Values to be estimated

As the total usage over a long period using the interpolated data should be the same as the original data (as summing over long periods, such as an hour, should be consistent) it is important that the estimated points should be set such that the green areas and the yellow areas have equal size. In this way the total usage remains the same even though the rate of usage has been estimated at all intermediate points. The top green area on Figure 3.5 is calculated as

$$\frac{x1}{2}.\frac{y1-b}{2} \tag{3.4.1}$$

The bottom green area is calculated as

$$\frac{b - y2}{2} \cdot \frac{x2}{2} \tag{3.4.2}$$

As these areas are identical,

$$\frac{x1(y1 - b)}{4} = \frac{x2(b - y2)}{4} \tag{3.4.3}$$

$$x1.y1 - b.x1 = b.x2 - x2.y2 \tag{3.4.4}$$

$$x1.y1 + x2.y2 = b(x2 + x1) \tag{3.4.5}$$

$$b = \frac{x1.y1 + x2.y2}{x2 + x1} \tag{3.4.6}$$

This formula will allow all the 12 point star readings to be calculated from the two samples either side of the time-stamp and allow for a series of readings to be created consisting of the five point and 12 point stars and the corresponding time-stamps.

Finally, it is necessary to estimate readings for time-stamps that follow a strict five minute boundary. These can be seen in Figure 3.6 and are represented by the black arrows on the time axis. The corresponding energy usage can be read from the generated thick black lines and are represented by the horizontal red arrows.

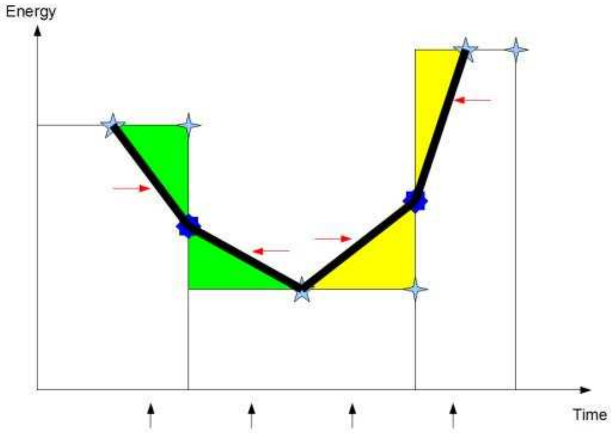


Figure 3.6: Final data layout

Issues arising from this approach include:

- Total usage summed from the generated five minute time points may not correspond to the original total usage. However, as the number of data points grows, these values will converge. The original data may underestimate the total usage for a long period (e.g. hour plus) as the original data contains periods for which no usage is collected (the one or two seconds taken to upload to Pachube and the missing time-stamps due to upload errors to Pachube).

- Long periods of missing data (e.g. hours) are represented by a sequence of very similar readings and hold little useful information. A rule is introduced to distinguish between missing data (long periods) and variation in the time-stamps of readings caused by the CurrentCost Bridge (short periods). The above algorithm is applied for the time-stamp variation with long periods (days) with missing data being deleted from the database.

The algorithm for implementing the above procedure is:

- Original data series consisting of time-stamp, meter reading and temperature.
- The mid-sample period point of each interval is calculated and a new series created with modified time-stamps (by subtracting 150 seconds).
- New value for each of the old time-stamp points (i.e. "b") calculated from the formula above.
- The two series are merged into time-stamp order. Any duplicate timestamps should have identical values and can be ignored.
- All five minute time-stamps within the specified time interval (first, just before first and last point in series) are calculated.
- Before and after time-stamps for each five minute interval are found in the merged series and the corresponding value for energy usage and temperature calculated.

Daylight saving issues are addressed by creating interpolated data for the "non-existing" 1am to 2am period on the Spring daylight saving day and by averaging the 1am to 2am values (of which there are two for each time-stamp) in Autumn. For each household, the readings (for electricity and temperature) at 12.55am and at 2am are found. Then each of the five minute times (i.e. 1:00, 1:05, etc.) are interpolated assuming a straight line between the readings at 12.55am and 2am. These calculated readings are added to the database to ensure that Spring daylight saving days have 288 daily readings per household and no missing values. The sum effect of this is that each day (including the days with daylight saving changes) have 288 readings.

The readings on the Autumn daylight saving days are found for the period of 1am to 2am. These readings are averaged (i.e. mean of the two readings is calculated) and the original duplicated readings are deleted from the database. The new averaged readings are then inserted to ensure that the Autumn daylight saving days have 288 readings per day per household.

The households that have been identified as "accepted" (i.e. have completed all necessary forms and produced a reasonable amount of data) are marked.

Data Storage

After cleaning of the data, it is stored in a MySQL database using a star schema [128] to ease the querying of the data. Appropriate indexes are added to the stored tables to improve query performance.

The roll-out of electricity smart meters across the UK collecting readings at half hourly, or more frequent intervals, will lead to a massive growth in data available for analysis by utility companies and other interested bodies. The handling of the large data volumes in ways that allow easy and quick analysis will be essential for efficient use of the data. De Silva et al. [129] has investigated suitable methods of storing the meter data to allow for rapid determination of trends. The basis of this work is the use of a data warehouse schema as used for the NESEMP data.

The data schema is shown in Figure 3.7. In addition to the shown tables, specific tables for the motifs found for each combination of alphabet size and motif size are also created. There is also a base data table which

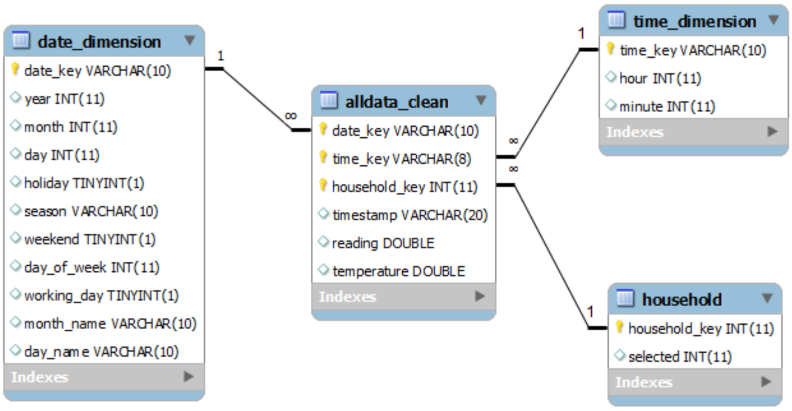


Figure 3.7: Main data schema

contains the original data prior to the cleaning process and the creation of the alldata_clean table.

The data collection project is ongoing and it is planned that, periodically, the data collected is uploaded to the database shown. At the time of the analysis described in this thesis, the data covered the period of 3rd November 2010 to 21st October 2012. The project started with an initial phase with fewer households so not all households have data for the full period.

The data volumes are shown in Table 3.1.

Table 3.1: Data Volumes

Table	Number of records
ALLDATA_CLEAN	31073472
HOUSEHOLD	380
TIME_DIMENSION	288
DATE_DIMENSION	1461

3.5 Experimental Environment

The research uses R Studio version 0.98.501 using R 3.1.0 [130] running on a Windows 7 64 bit system and accessing the data stored within a MySQL v5.5.31 database.

R software packages used include:

- e1071 [131] to provide the cmeans algorithm.
- reshape [132] to provide melt and cast functions to rearrange the data structures.
- kohonen [133] to provide the Kohonen self-organised map functions.
- fpc [134] to provide cluster similarity measures.
- randomForest [121] to provide Random Forest functions
- RMySQL [135] to provide the interface from R to MySQL.
- matlab [136] to provide emulation of some Matlab packages.
- clue [125] to provide cluster ensemble processing.
- clv [137] to provide cluster validity indexes.
- varSelRF [138] to allow selection of the most important attributes when comparing with questionnaire data using a random forest approach.
- StatMatch [139] to make use of alternative distance measures including Mahalanobis distance.

3.6 Summary

This chapter provides information on the techniques and algorithms used throughout the rest of the thesis.

The data used in the analysis, the method of collection, the data quality issues and how they are addressed and the way in which the data is stored to make it available for analysis are described.

The key clustering algorithms identified in the literature review (Chapter 2) as being commonly used in the area of electricity load profile clustering are identified and described.

The cluster validity indexes used in previous electricity load profile clustering work are defined. Additional measures are combined with the compactness measure based on the CDI CVI to provide a more “complete” measure of the quality of the clustering partitions. This composite measure is defined in Chapter 4 and is used throughout the thesis as the basis for comparing different methods of clustering, different datasets or parameter settings.

The broad approach to addressing the research questions laid out in Chapter 1 is described. Further details and full results are provided in Chapters 4, 5 and 6.

To allow others to repeat the work detailed in the thesis, the hardware and software environment used for the analysis is described.

CHAPTER 4

Obtaining Archetypical households using Load Profiles

This chapter addresses the research question laid out in Chapter 1 of whether it is possible to find clusters of households based purely on the electricity meter readings. This will be tested using the data from the NESEMP project and assessed using an appropriate cluster validity index.

Various cluster algorithms are applied and the resulting cluster solutions are compared using appropriate validity indexes (CVIs). It is argued that the existing widely used CVIs concentrate on only a few of the aspects of what makes a cluster useful and a new, extended composite measure is required. A novel measure is defined and used to assess the effectiveness of the differing algorithms when applied to electricity load profiling.

A standard marketing approach is to categorise a large population into a few, representative entities which are archetypical of the large numbers of the population assigned to that archetype. This approach allows for consideration of a few differing archetypes while allowing for interventions addressing a large proportion of the population. This chapter makes use of this approach and builds measures to reflect the characteristics of what makes a “good” cluster.

This chapter proposes that, for the clusters obtained from the load profiles to be effective other measures should be included in the assessment of possible partitions. The focus is on providing a measure of the effectiveness of the cluster solution to the person designing and implementing a programme to influence the behaviour of the population, rather than on assessing how well the clusters found match the "real" clusters within the data (which is an area already well researched).

4.1 Background

Many different arrangements of a population into clusters are possible using a variety of clustering algorithms. The question arises as how to choose between these possible arrangements. Much research has been done on defining cluster validity indexes (CVIs) [87] which can be applied to the defined clusters to produce a measure that allows the comparison of differing cluster algorithms or sets of input parameters.

The field of data mining includes many ways of measuring the "quality" of a particular clustering solution [50]. In most cases the data mining literature evaluates a good solution as one in which the members of a cluster are close together (similar to each other) and the archetypes, representing each cluster, are far apart (well separated from each other). Much work [18, 140] has been published on defining "close" and "separated" and most of the existing measures of quality focus on measuring compactness and separation.

The field of marketing has explored the requirements for effective cluster analysis [49] and the clustering of individuals into archetypical groups for marketing purposes has been widely applied [6]. In particular, the generation of clusters showing individuals with similar behaviour is used as a precursor to the creation of a specific offer to a given cluster that the marketing professionals believe will be attractive to members of that cluster (and hence taken up).

This chapter addresses the question as to whether it is possible to find a few clusters of households using UK data. Furthermore, can a composite measure, based on sub-measures, be defined that will allow choices to be made between different clustering algorithms. Using the composite measure, is it possible to give guidelines on the most appropriate clustering algorithm to use for load profile clustering.

4.2 Approach

4.2.1 Data Selection

The analysis uses data collected at five minute intervals for 123 households from the Spring 2011 period (also compared with Spring 2012 for the consistency measure) filtered to include just working days (weekdays which are not public holidays). This results in 2,653,056 readings providing a total of 9212 days of readings across all households (not all households have the same number of days of data). The data for each household are averaged by time of day so that each of the 123 households is represented by a single load profile consisting of 288 readings at five minute intervals across the day.

A second dataset is created using the five minute data described above but aggregated into hourly readings for each household. After averaging over the households, this results in a dataset consisting of 123 household representative load profiles, each with 24 hourly readings.

The average load profiles for each household are normalised such that the values for electricity usage lie in the 0-1 range.

The normalised profiles are then used as input to clustering tests using a variety of clustering algorithms.

The analysis is repeated for 3 further datasets drawn from the readings at Spring weekends, Summer weekdays and Summer weekends to assess whether the results from each different time period are substantially different.

4.2.2 Clustering Algorithms

Based on the review by Chicco [14] and other work [71], the following clustering algorithms are selected as the most commonly used for electricity load profile clustering:

- Kmeans
- Fuzzy Cmeans

- Self Organised Maps
- Hierarchical clustering (using average and Ward linkage)
- Gaussian mixture model
- Random Forests

These algorithms are described in more detail in Section 2.5.1.

4.2.3 Selecting the Appropriate Number of Clusters

A common issue with clustering is the appropriate setting for the number of clusters. To match common practice within the electricity industry, an optimum number of eight clusters is selected. See Section 4.3.1 for a discussion of this decision.

As the composite measure discounts small clusters when calculating the sub-measures, it may be more effective to run the clustering algorithms with numbers of clusters (k) greater than the optimum number as some of the small, outlier clusters are discounted. To explore this question, kmeans clustering using the five minute data is repeated a number of times for different values of k (from 6 to 12).

The optimum number of clusters is varied from 6 to 9 for each of the values for k and the composite measure is calculated. From these results the best setting of k can be determined to achieve the highest value for the composite measure for the desired optimum.

4.3 Cluster Quality Evaluation

The field of Data Mining includes many ways of measuring the “quality” of a particular clustering solution. In most cases the data mining literature focuses on the two criteria of compactness within a cluster and separation between clusters. A good solution is one in which the members of a cluster are close together (similar to each other) and the archetypes, representing each cluster, are far apart (well separated from each other). Much work has been done on defining “close” and “separated” with most of the proposed measures of quality focusing on measuring the compactness and separation.

The criteria that makes a good marketing segmentation are described in Section 2.4.1 and an effective cluster validity index should evaluate partitions using these criteria. To build an extended cluster validity index that incorporates some of the marketing related criteria it is necessary to:

1. Define a way of producing a measure for each of the criteria. To allow comparison between, and combination of, the criteria, the measure should vary over the range of 0 to 1 where 1 represents the best solution.
2. Determine a method of combining the criteria to give a composite single measure.
3. Provide a way of weighting each of the sub-measures such that adjustments can be made to the weightings depending on the intended use of the results.

The work presented in this chapter builds on the extensive prior research into cluster validity indexes and proposes an extension that allows possible partitions to be assessed for implementing DSM techniques. The selection of appropriate clusters is sometimes described as being more of an art than a science. This work provides an objective basis for decisions on the effectiveness of cluster solutions that previously may have been made subjectively.

4.3.1 Component Measures

Compactness

The data mining literature provides many possible cluster validity indexes that could be used to give a good indication of the degree of compactness within a cluster and the separation between clusters. The Silhouette measure [141] is often used in clustering across many domains. However, as it has been extensively used in previous work with electricity load profiles, the CDI (Cluster Dispersion Indicator, see Section

3.2.1) measure is selected. Applying the proposed approach of creating a composite measure to a different application domain may mean that the Silhouette measure (or, indeed, another CVI) may be more appropriate and can be included in the proposed composite measure in preference to the CDI.

The CDI can give values larger than 1 with the lowest values considered the best. Therefore, for comparison with the other sub-measures, it must be adapted to fall within the 0-1 range with 1 being the best solution. To achieve this, a value for the CDI for a simulated, random set of load profiles of the same size as the data under investigation is calculated. The random profiles consist of M records, each consisting of H values drawn from a uniform 0-1 random distribution. A clustering algorithm is applied to the random records and the CDI (CDI_{random}) calculated for the resulting partitions.

The compactness measure is calculated as

$$Measure_c = 1 - \frac{CDI}{CDI_{random}} \tag{4.3.1}$$

If the CDI generated from the random set of data is found to have a smaller value (i.e., better) than the real data CDI then the measure is set as 0. Thus, the measure is an indication of how the clustering solution compares to that resulting from a random set of data with higher results being better and a result of 0 denoting no benefit over the random solution.

A different set of random observations can lead to a different (maybe better) random CDI and thus the value of the sub-measure can change. This has been addressed by repeating the calculation of the CDI for 11 sets of random data and then selecting the set with the median value for CDI. 11 has been chosen as a large enough number to ensure that the random dataset selected generates a CDI that is likely to be very close to the average across all possible random datasets.

To allow for comparison between the differing clustering algorithms, the same random CDI can be used for all analysis using the same dataset. For example, in this chapter, the kmeans algorithm is taken as the base method and all the alternative clustering algorithms use the kmeans random CDI measure when calculating the compactness measure. As the compactness measure provides results relative to the base measure, any of the clustering algorithms could have been selected for the random calculation without impacting on the relative values of the measures.

Parsimonious

A useful clustering solution will consist of a reasonable number of clusters such that the understanding and addressing of the clusters is not too complicated. On the other hand, too few clusters is likely to provide insufficient differentiation between the groups of interest. The optimum number of clusters for each industry is dependent on standard practice within that industry.

Within the electricity industry, 8 is generally accepted as a good number for the optimum number of clusters. Ramos et al. [67], Figueiredo et al. [79], Rodrigues et al. [142] report on Portuguese electricity market experts who recommend a number within the 6-9 range. Tsekouras et al. [143] analyse various numbers of clusters using Greek data and conclude that most clustering algorithms examined perform best when generating 8-10 clusters. The UK electricity market defines eight generic profiles [51] for each season and weekday/weekend designation although only two of these profiles are for domestic users (the others are for industrial users). Defining around eight profiles for domestic household usage would provide more flexibility for the designers of behaviour modification interventions while still limiting the number to a level that can be easily managed.

A clustering solution may include clusters with very few members and which represent outliers within the data. These will not be addressed by any DSM program as they are too small for any cost effective intervention and thus will be excluded when calculating the parsimonious measure. A figure of 50% of the mean cluster size is taken as the cut-off for designating clusters as outliers.

The gamma probability density function provides a suitable shape for measuring how close the number of useful clusters is to the industry optimum. Small differences from the optimum incur small penalties while a figure far from the optimum number of clusters incurs a relatively higher penalty reflecting the lack of usefulness. The shape of the function can be scaled to provide a value for the measure in the range of 0-1. The gamma density function is given by

$$P(x,\alpha) = \frac{1}{\Gamma(\alpha)} \int_0^x e^{-t} t^{\alpha-1}\, dt$$

for $x \geq 0$ and $\alpha > 0$ and where the gamma function is defined by the integral

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt, \qquad z > 0$$

and α is the scale factor which is set as $optimum + 1$.

Figure 4.1 shows the density function (scaled to give values between 0 and 1) for an application where the industry standard requirement is for eight effective clusters (i.e., optimum=8 and hence shape=9).

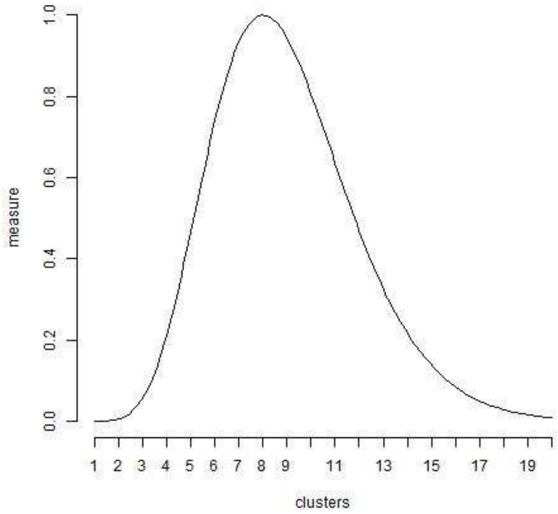


Figure 4.1: Example modified gamma density function with shape=9

The proposed parsimonious sub-measure provides a value for the degree to which the clustering solution provides a small number of substantial clusters where the optimum number of substantial clusters (for example, eight) depends on the application domain.

Substantial

To allow useful application of the clustering results, the cluster sizes should be large enough to make them worth addressing. While there is an argument that good results can be obtained by addressing the "long tail" (i.e., the clusters which have few members) this is an enhancement that can be applied once the reasonable sized clusters have been addressed successfully.

In the same manner as for the parsimonious sub-measure, any small clusters representing outliers are excluded. Once these small clusters are removed, an ideal solution for ease of developing interventions would be to have clusters of equal size. In practice, equal sized clusters are very unlikely and may not represent the "real" natural clusters within the data. However, this sub-measure is intended to reflect the usefulness of the partitioning solution and the match of the cluster solution to the "real" clusters is addressed by other sub-measures (in particular the compactness measure).

It can be argued that not all the clusters need to be of reasonable size so long as the ones representing the subset of the population that is to be addressed are. The clusters that won't be addressed can be very small or large without impacting on the usefulness of the partitioning. However, as there is no way to automatically know the clusters that will be addressed, it is necessary that the sub-measure considers all clusters as being equally likely to be addressed.

While the substantial sub-measure is related to the parsimonious sub-measure (fewer clusters will tend to mean each cluster is of a reasonable size), the relationship isn't strong and separate sub-measures are useful to represent the different aspects of the solution.

The data to be clustered consists of M records numbered as m=1,..M. Each record has H features numbered as h=1,..H.

The data is clustered into K clusters (numbered as k=1,..,K). Each cluster has R_k members where $r_{(k)}$ is the rth record assigned to cluster k and $C_{(k)}$ is the calculated centre of the cluster k.

L is the number of clusters of size greater than 50% of $\frac{M}{K}$.

$$Measure_s = 1 - \frac{1}{2M} \sum_{l=1}^{L} |\frac{M}{L} - R_l| \qquad (4.3.2)$$

The substantial sub-measure thus gives a value that depends on how close each of the cluster sizes is from the size resulting from partitioning into clusters of equal sizes. The closer the cluster sizes are to the equal sizes, the closer the sub-measure is to 1.

Consistent

The approach to measuring the consistency of a clustering algorithm is to determine two comparable time periods, and then to apply the algorithm to the data from each time period, before comparing the memberships for each partitioning to check for correspondence. When households are clustered with the same households in each time period, the solution is regarded as more consistent.

A relative cluster validity index is used to compare the results from the two time periods. The Rand index compares the different pairs of samples (i.e. each possible pair of households) and considers the number of pairs where the members are in the same partition in both time periods (good), those where the members are in different partitions in each period (good), and the case where they are together in one period but separated in the other (bad). The corrected Rand index [126] builds on the original Rand index but adjusts the calculated value for the expected matching that would occur in a random arrangement. The corrected Rand index takes values from 1 to -1 with the higher the value for the corrected Rand index, the better the two partitions for the 2 time periods agree.

The different behaviour of the Rand index and the corrected Rand index are explored in Figure 4.2 which is produced by considering a simulated set of 100 samples split into 8 clusters. The cluster sizes are set as 34, 24, 16, 10, 7, 4, 3, and 2 to approximately match the sizes of clusters found in the meter data analysis. The first 34 samples are assigned to cluster 1, the next 24 samples are assigned to cluster 2, and so on. A second synthetic dataset is created but with the samples assigned at random to the various clusters using the same distribution of cluster sizes. This results in two datasets which are very different from each other. 100 sets of cluster assignments are then generated with the probability of the cluster assignments matching the first synthetic dataset varying from 100% to 1%. The effect of this is to create a set of datasets that progressively vary from being very close to (or exactly) the first dataset to being very close to the second dataset.

The Rand index and the corrected Rand index for each of the 100 datasets are then calculated (comparing to the first synthetic dataset) and the results shown on the figure. It is apparent that the corrected Rand index reduces quickly as the cluster assignments move away from the original synthetic dataset. This quick reduction makes the corrected Rand index less useful for the composite measure as the measure quickly nears 0 for relatively small differences in partition membership. Thus the original Rand index is selected for use in the composite measure.

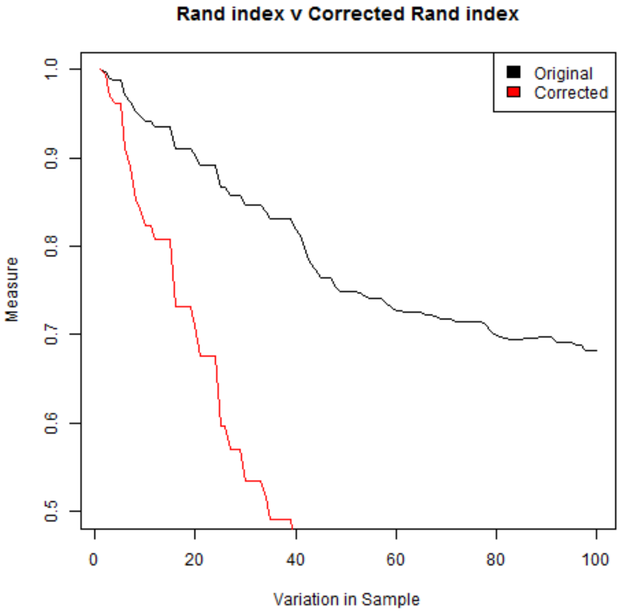


Figure 4.2: Comparing the Rand index and the modified Rand index

$$Measure_{con} = Rand(data_1, data_2) \tag{4.3.3}$$

where $data_1$ is the partitioning for the period of analysis and $data_2$ is the partitioning for an alternative, comparable period.

When considering electricity load profiles, a partitioning based on a set of profiles from one time period may be compared with a partitioning from a separate time period so long as comparable time periods are used - for example, load profiles based on Spring working days in 2011 should be compared with Spring working days in 2010 or 2012.

Building the overall measure

Each of the four sub-measures provide a value within the range of 0-1. To create a single measure, the sub-measures can be averaged in a number of ways including the arithmetic mean, the geometric mean and the harmonic mean. The harmonic mean is often used when combining a series of rates (e.g. rate of spending). The geometric mean allows the combination of sub-measures that have different ranges where the arithmetic mean would give undue importance to those sub-measures with higher ranges. Both the harmonic and geometric means give emphasis to sub-measures with small values (or zero).

To explore the differing combination methods, sub-measures are calculated for each of the clustering algorithms (see Section 4.4) and each of the means are calculated with the results shown in Table 4.1. The desired behaviour is for the combination method to give equal emphasis to each of the sub-measures without a bias towards smaller sub-measures. It can be seen that the various means give the same rankings of the results and, as there is no need for the geometric mean feature of dealing with differing ranges, the arithmetic mean is selected.

If professionals making use of the results consider some of the criteria are more important then a set of weights can be defined which bias the overall measure towards the criteria judged to be the most important for the behaviour change intervention planned. The overall measure can be calculated from the following formula:

$$Measure_{all} = W_c * Measure_c + W_p * Measure_p + \\ W_s * Measure_s + W_{con} * Measure_{con} \quad (4.3.4)$$

where $\sum W = 1$.

Table 4.1: Results using different means

	Arithmetic		Geometric		Harmonic	
	value	rank	value	rank	value	rank
kmeans	0.814	4	0.81	4	0.808	4
fuzzy	0.662	6	0.644	6	0.626	6
SOM	0.793	5	0.79	5	0.788	5
Hr-aver	0.566	7	0.389	7	0.188	7
Hr-ward	0.822	3	0.819	3	0.816	3
mix	0.841	2	0.838	2	0.834	2
RF	0.874	1	0.871	1	0.868	1

Within the remainder of this chapter, equal weights for each of the sub-measures are assumed.

Other criteria

The remaining marketing criteria of Compatible, Familiar, Accessible, Differentiable, Actionable, and Relevant discussed in Section 2.4.1 require subjective input from an expert to assess the quality of the clusters using external knowledge of the environment and are not addressed in the composite measure. With the addition of a ground truth set of information, some of these criteria are addressable.

Zhang et al. [144] suggests criteria that could be used to define archetypical households as being their property energy efficiency levels, the greenness of household behaviour of using energy, and the duration of daytime occupancy. With information collected from the population relating to these criteria, the resulting groupings can be compared to the partitioning solution from the clustering exercise with the closeness of the match leading to measures of the Familiar and Accessible criteria which could then be added to the composite measure described above.

4.4 Results

Each of the seven clustering algorithms are used with the five minute data with the sub-measures calculated shown in Table 4.2. The composite measure and its components are also shown graphically in Figure 4.3a.

Table 4.2: Results for five minute data

	Composite	Compact	Pars.	Subst.	Consist.	Cluster sizes
kmeans	0.814	0.744	0.934	0.798	0.778	3,8,8,13,15,22,23,31
Fuzzy	0.662	0.578	0.468	0.88	0.723	1,12,22,27,29,32
SOM	0.793	0.742	0.74	0.907	0.784	3,4,11,16,20,20,22,27
Heir - average	0.566	0.779	0.058	0.679	0.749	1,1,1,3,3,14,24,76
Heir - Ward	0.822	0.741	0.934	0.815	0.797	3,9,10,12,15,21,24,29
Gaussian	0.841	0.742	0.934	0.795	0.894	3,8,9,9,20,20,26,28
RF	0.874	0.8	1	0.879	0.818	9,12,12,14,15,19,20,22

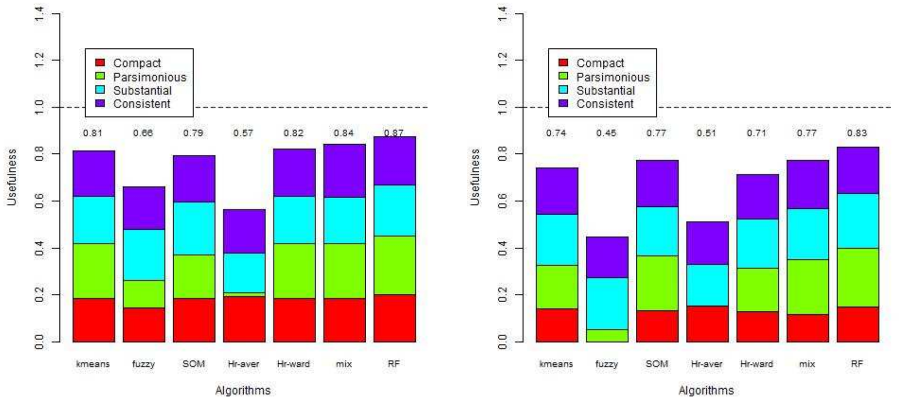
The analysis with the same seven clustering algorithms is repeated using the hourly data with the results shown in Table 4.3 and graphically in Figure 4.3b.

Table 4.3: Results for hourly data

	Composite	Compact	Pars.	Subst.	Consist.	Cluster sizes
kmeans	0.74	0.572	0.74	0.87	0.78	3,6,11,14,16,21,23,29
Fuzzy	0.449	0	0.213	0.882	0.699	22,25,35,41
SOM	0.774	0.532	0.934	0.834	0.798	5,8,12,14,15,16,22,31
Heir - average	0.51	0.614	0.006	0.695	0.725	1,1,1,2,3,6,17,92
Heir - Ward	0.714	0.523	0.74	0.833	0.759	6,7,11,13,14,17,26,29
Gaussian	0.775	0.476	0.934	0.867	0.822	7,11,12,13,15,17,21,27
RF	0.831	0.602	1	0.923	0.801	12,12,14,14,16,17,18,20

For comparison purposes with other load profiling work, the values for CDI and MIA for each clustering algorithm are shown graphically at Figure 4.4 and in Table 4.4. Each of the values of CDI and MIA have been normalised into the 0-1 range for display purposes in Figure 4.4. A lower value for the CDI and MIA suggests a better partitioning.

Figure 4.4 shows that the random forest algorithm provides the lowest CDI value. However, the MIA value for the random forest algorithm is much higher (worse) than all the other algorithms. This is a result of the use of the pam algorithm which selects a member of the cluster as the representative load profile by maximising the separation between the clusters. This provides a good value for the CDI (which includes separa-



(a) Five minute data

(b) hourly data

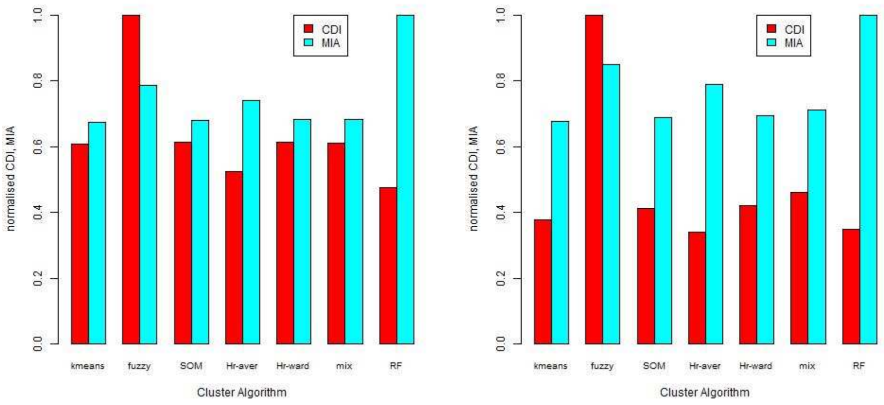
Figure 4.3: Overall composite measure showing components for each algorithm

Table 4.4: CDI and MIA results

	Five minute data		hourly data	
	CDI	MIA	CDI	MIA
kmeans	1.059	0.447	1.031	0.467
Fuzzy	1.744	0.522	2.743	0.586
SOM	1.068	0.451	1.128	0.474
Heir - average	0.915	0.49	0.931	0.543
Heir - Ward	1.069	0.454	1.149	0.479
Gaussian	1.065	0.452	1.263	0.491
RF	0.827	0.664	0.958	0.69

tion between clusters) but not for the MIA (which measures closeness of the cluster members to the representative profile).

The composite measure results (Figure 4.3a) show that the random forest algorithm produces the highest (best) values for the composite measure for both the five minute and hourly data. The composite measure is 0.87 (within a possible range of 0 to 1) which can be interpreted as being a partitioning that is 87% of the "perfect" partitioning (with equal sized clusters, no outliers, extremely good compactness within, and separation between, the clusters). The measure makes most sense when comparing between different algorithms rather than giving a definitive "size" of the quality of the solution.



(a) Five minute data

(b) hourly data

Figure 4.4: Normalised MIA and CDI measures for each algorithm

The make up of each of the clusters for the five minute data using the Random Forest algorithm (the result with highest value for the composite measure) is shown at Figure 4.5 where the coloured line shows the representative profile for each cluster and the black lines show the average profile for each household allocated to that cluster. The individual household load profile allocations to the 8 clusters shows, subjectively, a good fit to the representative load profiles for each cluster (shown on the last graph). Each of the representative profiles is reasonably distinguishable and will allow an appropriate intervention to be directed at the individual clusters. For example, the purple (22 houses) cluster has relatively little usage during the evening peak period and would not be a good candidate for incentives for load shifting away from the peak period. However, the cluster shown in orange (20 houses) has a significant proportion of usage in the peak period and may be the best households to address for a evening peak shift intervention program.

Considering the composite measure, the hierarchical solution (using average linkage) is penalised for the greatly varying sizes of clusters in the solution. The hierarchical clustering (average linkage) scores relatively highly on the consistency component of the composite measure but this is due to the poor distribution of sizes of clusters with few clusters of reasonable size and one large one. This large size means households are more likely to fall into the same clusters in the 2011 and 2012 data (as there are so few reasonably sized alternative clusters). While the sizes

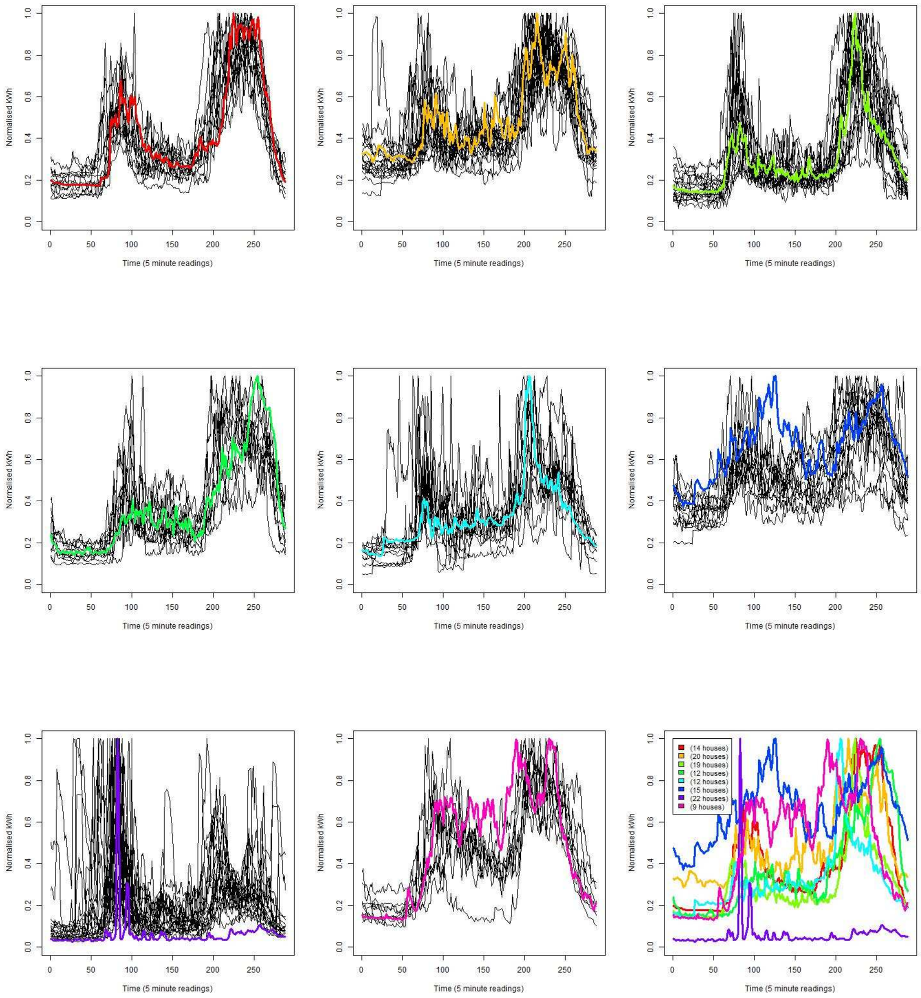


Figure 4.5: Representative Load Profiles - Random Forest, five minute data (k=8)

may match the "true" distribution within the data, it is not a useful range of sizes for developing and implementing a suitable intervention and hence is awarded a low overall composite measure.

The graphs shown in Figure 4.6 show the archetypal load profiles found for each of the eight clusters using each of the clustering algorithms. Each coloured line shows the representative usage profile for a single archetype and denotes the electricity usage over the day with the x axis showing the 288 five minute readings during each day. The hierarchical clustering with average linkage and the fuzzy cmeans results are omitted as the composite measure for these algorithms was relatively low.

From inspection of the graphs it is clear that differing clustering algorithms have grouped the same households together in some archetypes (as would be expected) while some archetypes have been produced differently. In particular, the random forests algorithm has produced archetypes to represent the members of the cluster that are as different as possible. For example, the purple archetype has been created by selecting from amongst the households allocated to that cluster with the selected household being the one most different from the other archetypes.

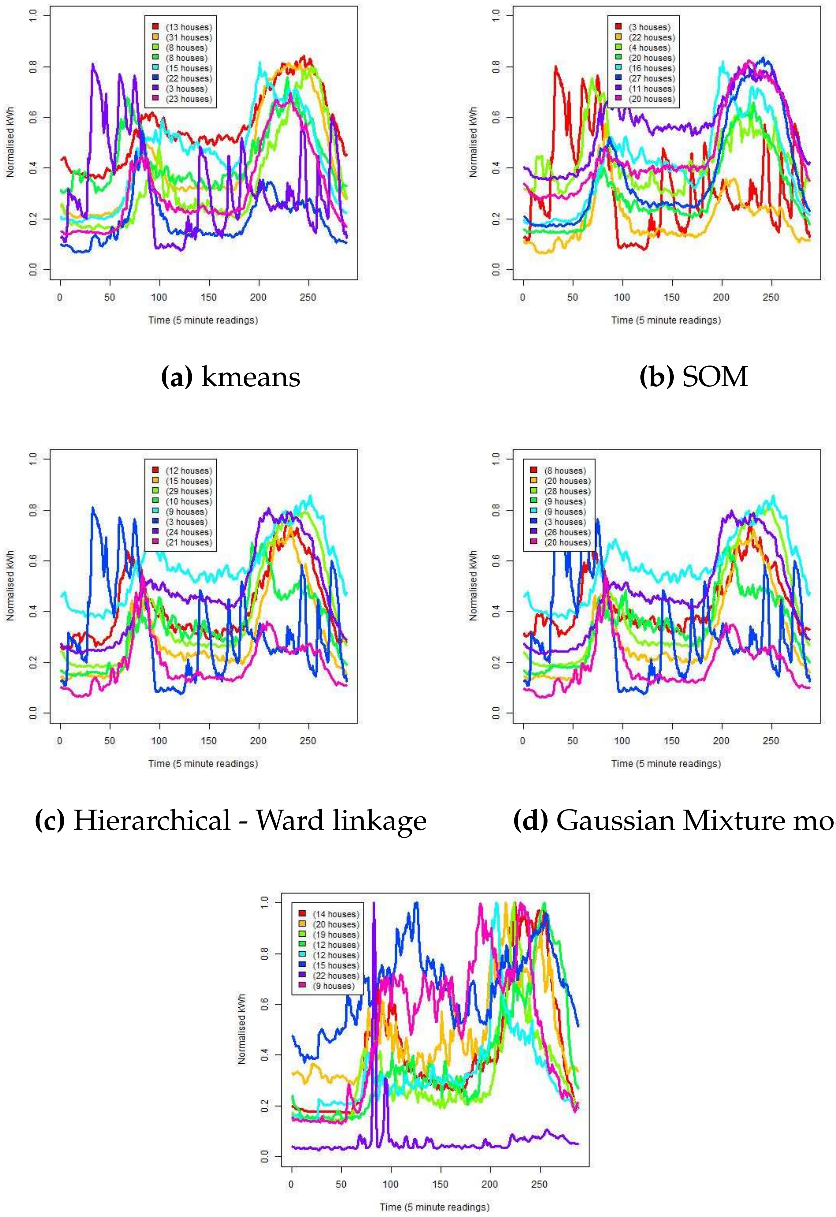
One benefit of the Fuzzy cmeans approach is the cluster membership function that the algorithm produces which has not been assessed in this analysis. In an application of the results, this membership function could be used to modify any intervention to provide customised incentives to a household while allowing the utility company to only have to deal with a small number (e.g. eight) of archetypical households [85].

4.4.1 Selecting the Appropriate Number of Clusters

The results from using differing values for the number of clusters (k) used with the kmeans algorithm against different values for the industry optimum number of clusters are shown in Figure 4.7. It is clear that selecting a value for k that is greater than the optimum number of useful clusters required often provides a higher value for the composite measure. For example, if eight useful clusters is the industry optimum then, using the kmeans algorithm, the third graph in Figure 4.7 suggests a value of k=10 would be appropriate to maximise the composite measure.

4.4.2 Evaluating Frequency of Data Sampling

Whereas the CDI and MIA results for the five minute and hourly data cannot be compared due to the differing numbers of attributes, the compactness sub-measure uses the ratio between the calculated CDI and that for the same sized random set of data. Hence the composite measure is comparable between differently sized datasets. The results (Figure 4.3 and Tables 4.2 and 4.3) show that the five minute data provides a partitioning with a higher composite measure than that for the hourly data



(a) kmeans

(b) SOM

(c) Hierarchical - Ward linkage

(d) Gaussian Mixture model

(e) Random Forests

Figure 4.6: Representative Load Profiles, five minute data (k=8)

across all the clustering algorithms.

This result matches the intuition that the five minute data will include more detail and thus allow for more accurate clustering of similar households. The components of the composite measure show that, while the hourly data tends to produce more equally sized clusters (and hence a high substantial sub-measure), the other components are lower for the hourly data. In particular, the compactness sub-measure is a lot lower.

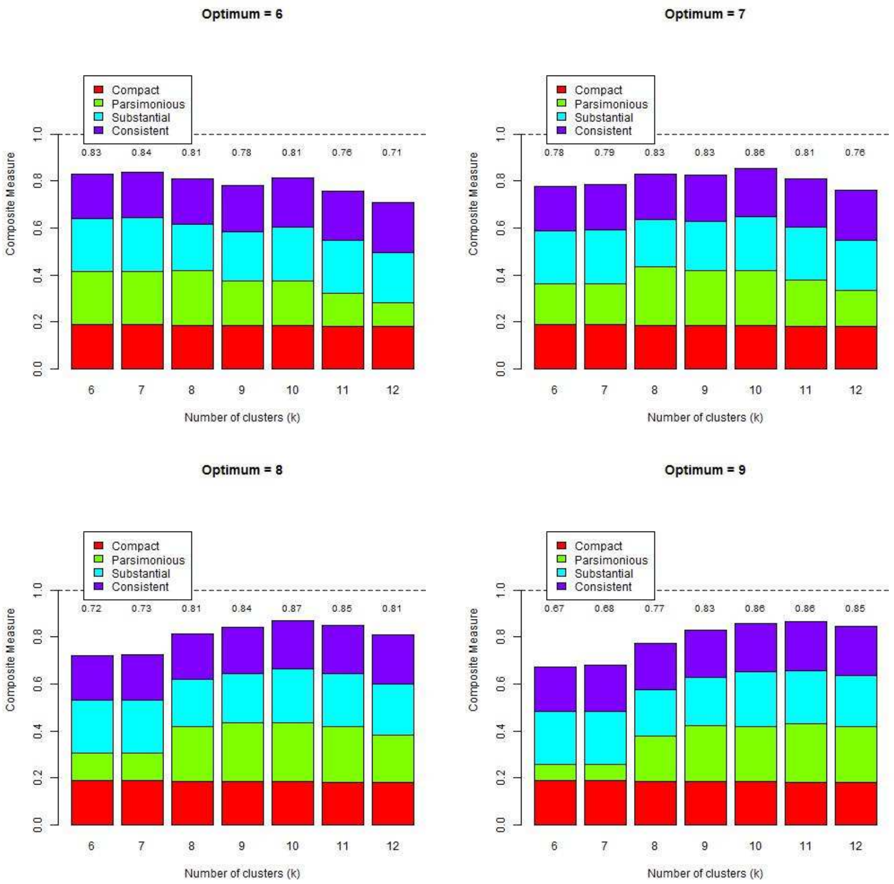


Figure 4.7: Analysis of number of clusters (k) v optimum using kmeans

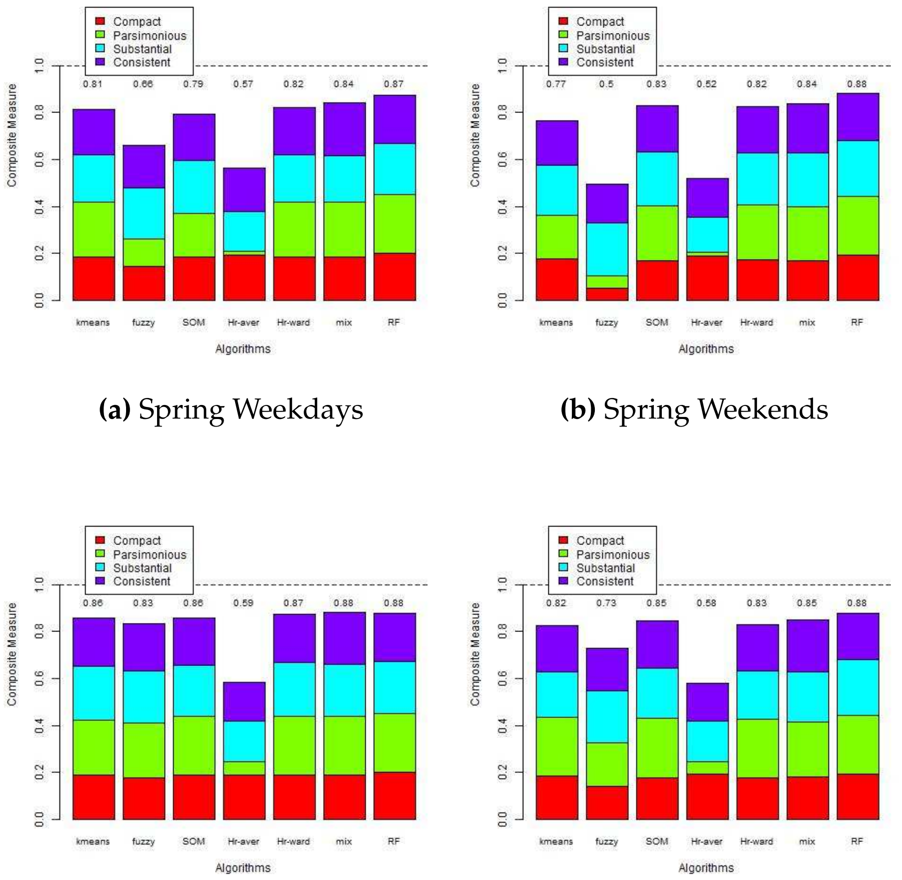
4.4.3 Comparison of results from multiple seasons

The approach detailed above is also applied to 3 additional datasets. Each data set has slightly different numbers of households included due to the need to have sufficient readings for each household within the 2011 and 2012 periods of interest. As some households left the study part way through or experienced problems with the technical set up of the monitoring equipment, not all households have a full set of readings.

The numbers of households in each data set are:

- Summer 2011 weekdays - 114 households
- Spring 2011 weekends - 113 households
- Summer 2011 weekends - 108 households

The clustering algorithm results for each dataset, as measured using the composite measure, are listed in Figure 4.8.



(a) Spring Weekdays

(b) Spring Weekends

(c) Summer Weekdays

(d) Summer Weekends

Figure 4.8: Composite results for alternate data periods

The rankings for each of the clustering algorithms, as assessed by the composite measure, are shown in Table 4.5.

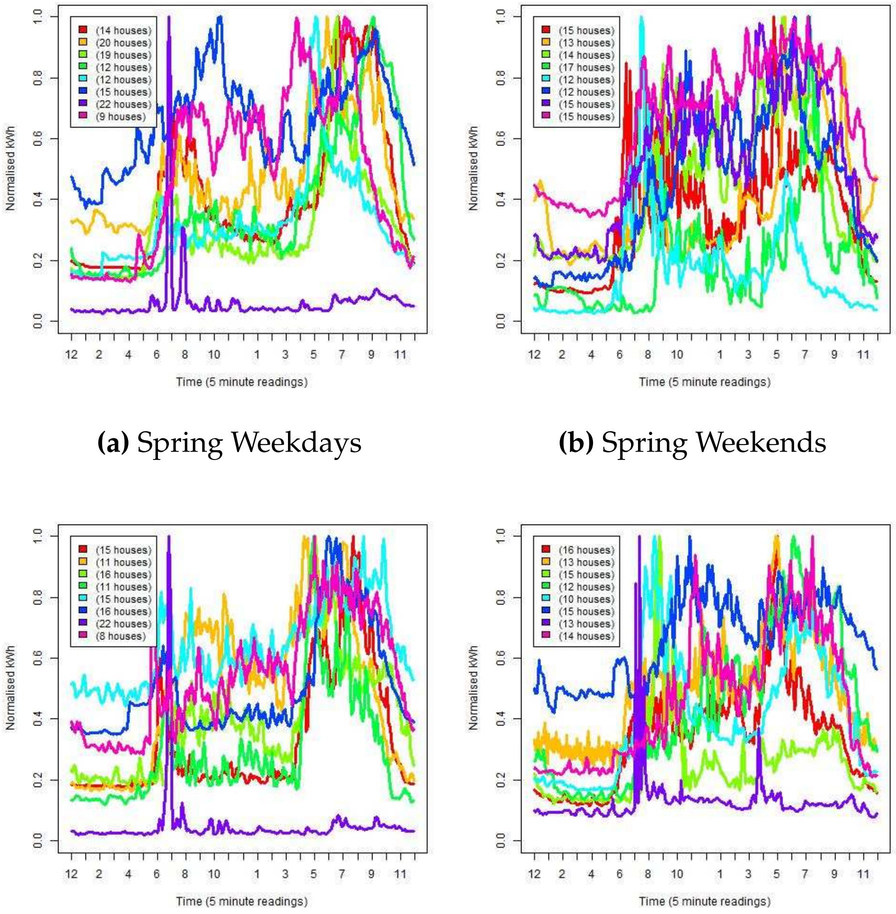
Across three of the datasets the top performing clustering algorithm, as assessed using the composite measure, is the random forest algorithm which is also second when using the Summer Weekdays dataset. The archetypes arising from the random forest algorithm for each of the datasets are shown at Figure 4.9.

The members of each archetype are shown on Figure 4.10 which can be compared with Figure 4.5 showing the same results for the Spring weekday period.

Comparing the results from the Spring weekends and Summer week-

Table 4.5: Clustering algorithms by measure across data periods

	Spring Weekdays	Spring Weekends	Summer Weekdays	Summer Weekends
RF	1	1	2	1
Gaussian	2	2	1	2
Heir - Ward	3	4	3	4
kmeans	4	5	5	5
SOM	5	3	4	3
Fuzzy	6	7	6	6
Heir - average	7	6	7	7



(a) Spring Weekdays

(b) Spring Weekends

(c) Summer Weekdays

(d) Summer Weekends

Figure 4.9: Archetypes using Random Forest algorithm for each dataset

days and weekends with the Spring weekday data detailed in the remainder of this chapter, the composite measure results show that the

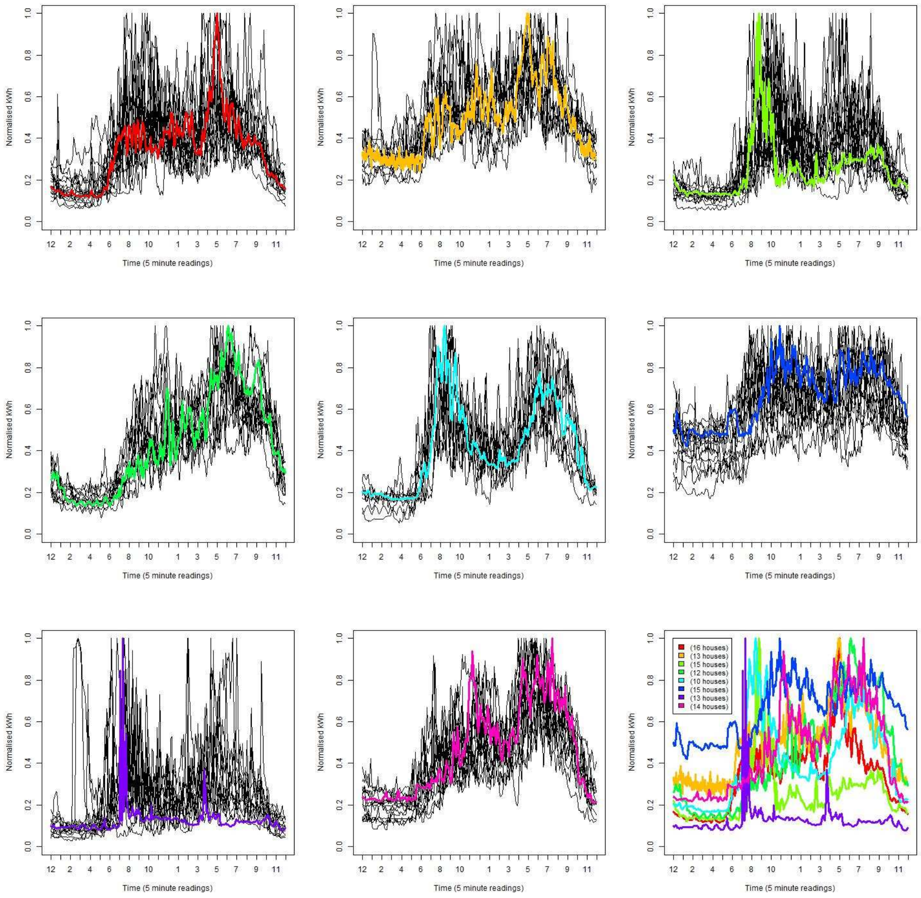


Figure 4.10: Load Profiles - Random Forest, Summer Weekend data (k=8)

clustering algorithms perform very similarly. The ranking of the clustering algorithms, using the composite measure, are similar across all the datasets.

Figure 4.9 shows an obvious difference between the weekday archetypes and the weekend archetypes for both the Spring and Summer seasons. Both types of days are similar in the early morning and over the breakfast peak but, whilst the weekday archetype tend to reduce in usage after breakfast and throughout the day, the weekend archetypes show much more variety in the household clusters. Some archetypes show a high electricity usage throughout the day whilst others do have a reduction after breakfast but less than on weekdays. The weekday early evening peak is still apparent at the weekends but is less pronounced.

The weekend usage profiles are, in general, harder to distinguish into easily understood archetypes and targeting a DSM intervention may be harder at the weekend than in the week.

Comparing between seasons, the weekday archetypes are very similar for Spring and Summer. At the weekend, the Summer archetypes show a midday peak which isn't apparent in the Spring archetypes.

The load profile for the households in each cluster are shown in Figure 4.10 and show a looser fit between the archetype for the cluster and the households using the Summer weekend dataset as compared with the Spring weekday dataset.

4.5 Conclusions

The results show that it is possible to produce a composite measure that can be used to assess the different approaches to clustering households using their electricity meter data. The composite measure has been defined by considering criteria that would be useful to professionals defining DSM interventions and, hence, provides an evaluation of the clustering results against the characteristics that would be most useful. The study does not test the results by implementing a series of trials in order to objectively assess which clustering results give the most useful impact in terms of changes in household behaviour.

The approach of using the extended measures is to focus on identifying the solutions that would be of most value to the individuals making use of the results to develop and implement interventions to change the behaviour of the population sampled. The criteria that mean a solution is useful to these individuals, who are likely to not be data miners, may not be the same as those that match the mathematical evaluations of "best" clustering.

Other chapters in this thesis consider different approaches to using the meter data to gain useful understanding of the households. The work in this chapter provides a baseline that can be compared with the further work. Using specific data collected from an area of the UK, that will also be used in the other chapters, it creates a set of archetypical profiles using clustering algorithms that have been applied to datasets from other geographical areas.

Arguments have been made that marketeers should concentrate on the "long tail" of a distribution of types of consumers with the most effective deployment of marketing resources often being towards the small number of consumers at the extremes of buying behaviour. As each intervention with consumers to change behaviour is likely to be expensive (in time and money), each of the created interventions need to be applied efficiently. At the moment, applying the DSM interventions to very small groupings of households in the long tail is unlikely to be a cost effective deployment of resources and is unlikely to lead to significant overall changes in behaviour. In the future, very low cost, personalised interventions could be deployed in a more targeted way to small, specific groups such as proposed in Dent et al. [85].

Based on the results obtained from the five minute data and, subject to the limitations of the study of:

- Using equal weights for the components of the composite measure
- Using data from a geographic region of the UK
- The clustering algorithms included in the study
- The parameter settings for the algorithms used

it can be concluded that the ordering of clustering techniques (by the composite measure) for electricity load profiling is as shown in the first column of Table 4.6. The other columns show the ranking resulting from using the CDI and MIA measures used in most other electricity load profiling work.

The analysis work has been repeated using different subsets of data drawn from the same database covering the different periods of Summer weekends and weekdays and Spring weekends and the results found are similar to those presented for the Spring weekday period.

It can be seen that the rankings of the differing algorithms vary between the evaluation measures used and the correlation (calculated using Spearman's rank correlation coefficient - rho) between the differing rankings

Table 4.6: Clustering algorithms by measure

	Composite	CDI	MIA
RF	1	1	7
Gaussian	2	4	3
Heir - Ward	3	6	4
kmeans	4	3	1
SOM	5	5	2
Fuzzy	6	7	6
Heir - average	7	2	5

is shown in Table 4.7. While there is some correlation between the results using the composite measure and those using the CDI, there is very little correlation between the composite measure and the MIA results. It is concluded that the composite measure gives an alternative approach for assessing differing clustering algorithms for electricity load profile clustering to the previous work using the traditional CVIs. This composite measure is potentially useful for electricity industry professionals working on demand side management interventions.

Table 4.7: Correlation of rankings of algorithms for each measure

	Composite	CDI	MIA
Composite	1.00	0.29	-0.07
CDI	0.29	1.00	-0.14
MIA	-0.07	-0.14	1.00

The consistency sub-measures for both the five minute and hourly data are both low (except for the Gaussian mixture model and self-organised maps) and suggest that other clustering algorithms should be considered for producing more consistent results. The period to use for comparison for the stability measure is influenced by the use to which the resulting cluster solution is to be applied. In this analysis an equivalent period from the following year (i.e., Spring 2012 working days) is used for comparison.

The analysis makes use of comparisons to a randomly generated set of data which is then clustered and the measures calculated. A different set of random observations can lead to a different (maybe better) random cluster solution and thus the value of the measures can change. For example, the clustering of the five minute and hourly data use different numbers of attributes. Each is compared to a random clustering solution using the appropriate number of attributes and thus each is compared to a different random solution. This possible problem is addressed by repeating the calculation of the compactness measure based on the random data a number of times.

Note that the composite measure is calculated in this analysis by using equal weights for each of the components. If information were available on the specific purpose for which the results would be used, it would be possible to assign differing importance to each of the sub-measures, and an appropriate weighting applied, which may lead to a different ranking of the clustering algorithms. However, the Gaussian mixture model shows a good score for each of the sub-measures and will likely be close to the best clustering algorithm whatever weighting method is chosen.

The question of frequency of usage monitoring to provide for effective load profile clustering is an important one and this work compares sampling at five minute and hourly intervals. The composite measure allows comparison between datasets containing differing numbers of attributes and hence between the two datasets. As the composite measure is higher for the five minute data for each of the clustering algorithms, it is concluded that the more frequent monitoring produces more useful partitions. One implication of this result is that the electricity industry should consider monitoring a representative sample of households at a higher frequency than the currently agreed UK rate of 30 minutes. These households can then be used to provide more detailed analysis and archetype creation than would be provided by the general population.

The approach has been developed to cluster household electricity load profiles but would be applicable to many situations involving individuals and the implementation of activities aimed at the cluster archetypes with the intention of obtaining a specific behaviour (e.g. purchase of a product, a change of electricity usage behaviour, or take up of a particular political message).

4.6 Summary

This chapter investigates the question (laid out in Chapter 1) of whether clusters of households can be created using electricity meter data collected in the UK.

The questions addressed include:

- Whether it is possible to create a composite measure, based on sub-measures, where each reflect an aspect of the partitioning, and which will allow choices to be made between different clustering algorithms
- Whether the sampling frequency affects the quality of the archetypes found
- Which clustering algorithms provide the "best" results as measured by the composite measure

Based on guidelines from literature in the field of marketing, four submeasures have been defined which reflect criteria for a "good" evaluation of a segmentation result. These sub-measures have been combined into a composite measure which is then used to assess the effectiveness of the clustering results to inform a DSM programme intended to change household behaviour. Using the composite measure provides meaningful results and thus the question of whether a composite measure can be created is found to be true.

The use of the widely used CVIs focuses just on the compactness (members of the same cluster are close and clusters are well separated) whereas the composite measure incorporates other criteria that makes the results more valuable for use with DSM. The different clustering algorithms have been assessed using the traditional CVIs as well as the composite measure and then ranked depending on the CVI results. It is shown that a differing ordering of algorithms is obtained using the composite measure and thus the question of whether the composite measure provides different evaluation results is found to be true.

Previous load profile clustering work assessed using the CDI and MIA measures has generally concluded that hierarchical clustering is the most effective algorithm with SOM and kmeans also providing reasonable solutions. Assessing the algorithms using the composite measure shows that the Gaussian mixture model provides the most effective results with kmeans being assessed lower than with traditional CVIs. The results in Table 4.6 show how the algorithms vary when assessed using the different methods.

The composite measure defined in this chapter is used to assess the clustering of households using the variability of occurrence of motifs as detailed in Chapter 6. The results from this chapter also provide a baseline which can be compared with the results from the following chapters when assessing differing ways of using the electricity meter data to cluster households.

One of the research questions to be assessed is whether the frequency of sampling of meter data affects the quality of the archetypical groupings found in the data. This chapter examines this question using datasets collected at five minute and one hour frequencies. The results are assessed using both the novel composite measure and the CDI and MIA validity indexes. The quality of the clusters found using the hour data are inferior to those from using the five minute data. Thus it is concluded that the more frequent sampling rate is preferable and provides results that are more effective in finding suitable clusters for DSM interventions.

This chapter demonstrates a novel approach to defining a composite measure that complements existing cluster validity indexes.

The results in this chapter can be used by electricity industry professionals to improve their analysis of load profiling clustering to develop effective clustering results to drive their implementation of demand side management techniques and hence lead to improvements in the electricity network efficiency. The chapter provides advice on the most effective clustering algorithms to use as well as providing a framework for the professional to use their opinion to set the weights of the various components of the composite measure.

The work laid out in this chapter provides the basis for investigating motifs within the data. The clustering using the base meter data (i.e. without considering motifs) provides a base line set of archetypes as have been produced in other work by other researchers using datasets from many countries. Repeating this work using the UK dataset, which will be used for exploring motifs, gives a starting point against which the motif work can be evaluated to assess whether or not it provides more useful information.

The definition of a composite measure will be used in assessing the clusters arising from the motif clustering work and, as the composite measure assesses the relative effectiveness of the clusters found, will allow for comparisons between the results from using the motif data and the base load profile data.

CHAPTER 5

Identification of Repeating Tasks within the Household

This chapter addresses the question as to whether a reasonable number of regular activities within the household can be identified solely using the electricity meter data. The intention is to identify activities that households may do regularly and not to identify the usage of particular appliances (e.g. toaster, kettle). The actual physical activity will not be determined from the electricity data but the research will determine whether repeating activities can be identified (without interpreting exactly what the activity is which would need input from the household members via questionnaires, diaries or similar).

The chapter adopts an approach of symbolising the electricity meter data and then finding repeating patterns within the symbolised data. Various ways of calculating the symbolised data and various parameter settings are explored to identify the most appropriate parameters for consistently finding repeating activities within the household. The goal is to, firstly, determine whether the data sampled at five minute frequency is sufficient to discover meaningful motifs, and, secondly, to determine the parameter settings that provide the most useful motifs.

If the research question can be answered positively then the motif finding method and parameter settings can be applied to the question of variability of the repeating activities which is addressed in Chapter 6.

This chapter includes the following sections:

- An explanation of what is meant by a motif.
- A description of the datasets used for the analysis.
- The general approach of symbolising the data and then matching the strings of symbols.
- Assessing the success or otherwise of the motif finding to answer the question of "What is a good result?".
- The effect of varying parameters within the symbolisation technique and selection of best parameters.

5.1 Motifs

A motif is defined as a previously unknown, frequently occurring pattern within a stream of data. It is distinguished from known patterns which can be found within a data set by a number of well researched methods [102].

In many cases the repeating patterns are not exactly the same and the power of a motif finding technique is in its ability to find similar patterns that, in real life, represent the same thing. For example, when considering electricity meter data, similar patterns may be discernible that represent the electricity usage during cooking.

The ability to use the meter data to discern particular activities can be useful to the electricity supplier in implementing demand side management techniques. If a particular pattern can be interpreted as a particular activity, then appropriate incentives or penalties can be offered/imposed on the particular household to gain a change in behaviour.

If a particular pattern is seen to recur then the characteristics of the recurrence can be used to drive demand management interventions. For example, if it is detected that a particular household undertakes the same activity at very different times of day, it can be assumed that there may be little or no requirement within that house to do the activity at a particular time and the household may be open to an incentive to change to a time that is more efficient for the overall network. However, a household that shows very regular behaviours may be less open to the incentive and the information can be used to help the utility company best target their interventions. This flexibility of behaviour will be explored further in Chapter 6.

As an example of what can be interpreted from the shapes of the electricity usage, Figures 5.1 and 5.2 show two days of meter readings for a single house (House number 5) from March 2011.

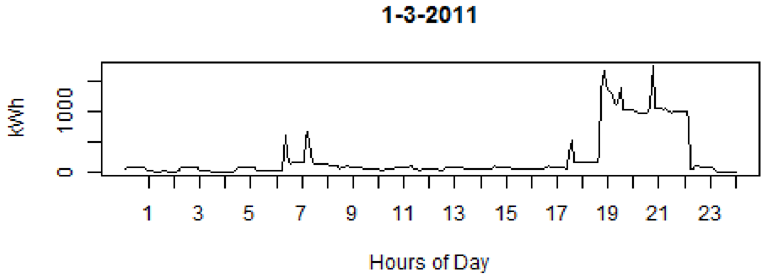


Figure 5.1: Example day of meter readings (household 5 on 1-Mar-2011)

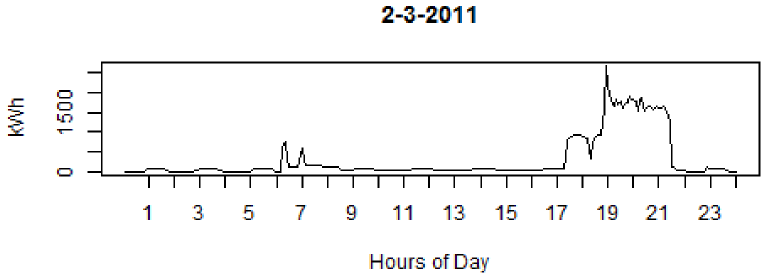


Figure 5.2: Example day of meter readings (household 5 on 2-Mar-2011)

Certain aspects of the household behaviour can be assumed as likely from inspection of the graphs:

- The household members seem to rise from bed at the same time each day
- It is likely that the household members are out of the house during the working day returning at about 5:00 to 5:30pm
- The members seem to retire to bed at about 10pm each day

- There is an underlying repeating pattern throughout the day and night which is likely to be a "always-on" device such as a fridge or freezer

Note that these assumptions are made from examining only two days of readings but are included to give an example of the types of conclusions that the detection of motifs can help people implementing demand response programs to reach.

The early morning (between 6am and 8am) pattern is repeated approximately on each day and may be a good motif that a larger scale analysis could use to detect the activity of "rising from bed / breakfast". Once the motif can be detected on a significant number of days, the variation in timing of the motif can be used to investigate variability of behaviour.

The challenge is for the motifs that the human eye can see as "similar" to be detected automatically by a suitable process. The early morning motif can be seen to be slightly different on each of the example days but an observer would use their common sense and knowledge of the types of things people may be doing in the morning (washing, making toast, making tea/coffee) to assume that the patterns show the same type of behaviour. The challenge to be addressed is to find a motif detection method that can automatically match the "common sense observer".

Some of the houses show a regular repeating pattern that suggests an "always-on" device in the house and which can swamp the interesting activity related patterns. For example, Figure 5.3 shows a zoom into the middle hours of a day for one house and it is apparent that there are a couple of appliances that regularly turn on and off. One appliance has a period of about ten minutes whereas the other has a period of about 70 minutes. This behaviour can mask the activities that would be interesting to detect for demand response reasons. One approach to addressing this issue is to set a minimum range for a motif to be considered. For example, setting a minimum range of 200W in this case would exclude the shown motifs as each has a range of about 150W.

Some houses have very high usage for short periods which can distort the shape of the whole day usage (e.g. see Figure 5.4 which suggests the use of a powered shower). This issue can be addressed by splitting the

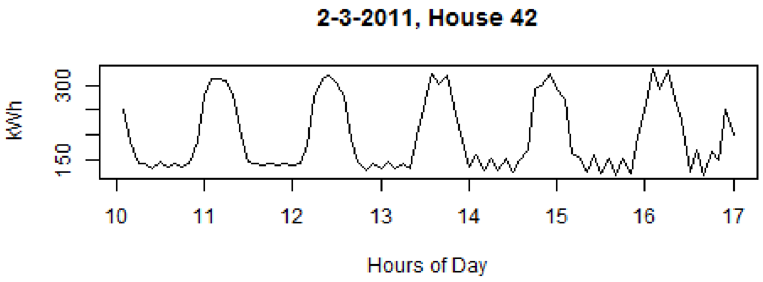


Figure 5.3: Example of very regular repeating patterns

motifs found within a household into a number of bands (e.g. low range, medium range, high range). If the definition of low and high ranges is taken from the range shown by the household (i.e., 8kW in this case) then only the shower motif will be classed as high range with any other motifs likely to be classed as low range. Instead of taking the definition of ranges from the household minimum and maximum usage, information on typical appliance energy requirements can be used to define low, medium and high ranges and this will allow some of the other motifs within the day (currently dwarfed by the shower) to be identified.

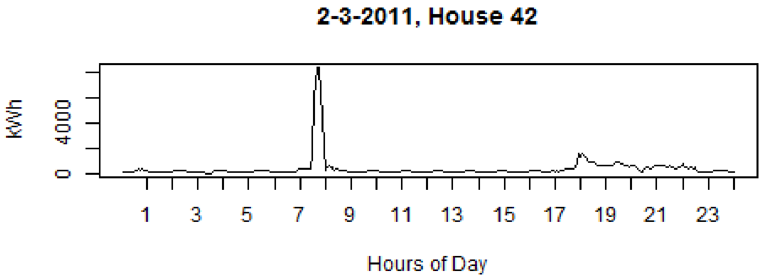


Figure 5.4: Example of large usage for a short period - shower?

Other work has considered the creation of derived datasets based on the presence (or absence) of particular patterns within the data. For example, Gay et al. [145] propose a method suitable for classification of data with binary valued attributes. Their work differs from the method proposed in this thesis which deals with real valued meter readings and includes a symbolisation step to handle approximate matching of patterns.

5.2 Motif Finding Method

Various motif detection algorithms are available and this work uses the SAX (Symbolic Aggregate approXimation) technique which provides for symbolic representation of time series data [106, 107]. Other motif finding algorithms could also be incorporated into the proposed approach (e.g. [146]).

The symbolisation step within the motif discovery process is important in allowing for approximate matching. As real valued meter readings are mapped to a few characters, as defined by the alphabet size used, various ranges of value are mapped to the same character and this process has the effect of compensating for some of the noise that is present in meter data collected from households. For example, doing the same activity on different days (e.g. making toast) would be expected to map to the same motif. However, other activities happening at the same time in the house (e.g. the fridge automatically running) or slight differences in the toaster settings (e.g. different level of brownness selected) may lead to slightly different meter readings. With the right symbolisation settings, it is more likely that the activity will map to the same symbolised representation.

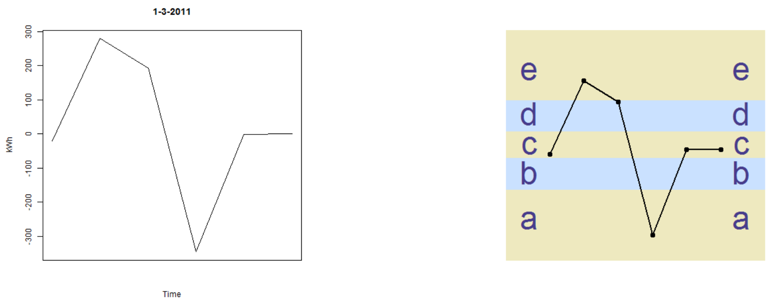
The first stage of searching for motifs is to index all of the patterns within the data. This is done by taking a moving window of a certain size (the motif length) and considering each of the meter readings within the window for each household. An example of this is shown at Figure 3.1.

Various different motifs are defined based on differing ways of considering the data and on how the data is normalised and converted to a symbolic representation. The different motifs created are:

- That created from considering the actual meter readings, normalised within the moving window being considered, and then converted to a symbolic representation. This process is graphically shown in Figure 3.1.

- The difference motif is created by taking the same data from the moving window and then calculating the differences between adjacent meter readings. This provides a series showing the relative increase or decrease in usage from the previous time point. This data is then treated as above (normalisation and symbolisation) to produce the difference motif. This is demonstrated in Figure 5.5.

- The overall normalised motif is created by considering all the data for each household and then normalising the data within the overall maximum range. The moving window is then taken across this normalised data and each window is then symbolised. Note that the normalisation within the window step is not executed.
- Compressed motifs are also created from each of the above approaches by removing any repeating letters within the symbolised motif. For example, a motif of "abcccb" will be stored in compressed form as "abcb".



(a) Example differenced data for 30 minutes **(b)** Translating the data into a symbolised string

Figure 5.5: Symbolisation process for differenced data

Values for the overall range of the meter readings within the motif window, and the range of the incremental change data (the differences between adjacent meter readings), is stored alongside the motifs.

The final result is that tables specific for each setting of the alphabet size and motif size are created providing for retrieval of the differing types of motifs together with details on the time and date of the motif, the household being considered, and the range information as described above.

The full list of data stored in the database tables containing the motifs is shown in Table 5.1.

The initial step in the motif finding process is to symbolise the data. This translates the real valued numbers obtained from the electricity meter

Table 5.1: Data stored for each motif

Field	Meaning
date key	Calendar date
time key	Time of day
household key	Household number
motif	Basic motif - e.g. accdee
diff motif	Motif calculated from differences
comp motif	Compressed version of basic motif - e.g. acde
comp diff motif	Compressed version of diff motif
norm diff motif	Normalised within all data rather than within the motif
comp norm diff motif	Normalised within all differences
spread	Range of readings across motif window
diff spread	Range of differences across motif window
mins in day	Minutes from start of day to motif starting time

into a restricted number of symbolic representations (generally represented as letters). This translation into a set number of symbols (e.g. the letters “a” to “e” for an alphabet size of five) necessarily incorporates some degree of approximation. As the alphabet size is restricted, the effect is to translate readings within a particular range into a given SAX letter and thus similar, although not identical, readings are translated into the same SAX letter. The resulting SAX string for two motifs may be identical whereas the original readings may only be approximately similar.

Alphabet sizes of 5, 7 and 9 are selected for analysis. An odd number is selected so that no change in meter readings between time points (or only slight changes due to background noise) will map to the “middle” character in the alphabet. An even alphabet size would provide two alternatives for the centre symbol and the selection will be driven by noise which is not the desired behaviour.

Too large an alphabet size will lead to very few repeating motifs being found whereas, similarly, too small an alphabet will allow very different activities to be mapped to the same symbolised representation. The selected sizes (between 5 and 9) are analysed further in the chapter to ensure that a "reasonable" number of motifs can be found.

The base motif size used for analysis is 6 which corresponds to a 30 minute (i.e. 6 x five minutes) period. This figure was selected as the UK electricity settlement market uses a 30 minute period and 30 minutes is also a reasonable estimate for a time period that will include most household activities such as breakfast, showering, etc..

To explore the setting for motif size that results in a reasonable number of motifs being found, additional motif sizes of 4 (20 minutes), 9 (45 minutes) and 12 (1 hour) are considered to find the most effective setting. Settings for the alphabet size and motif size are considered together to find a reasonable number of motifs.

The motifs found are compared on the basis of similar shape (represented by the same symbolic representation of the motif) without regard to absolute value of the data. A possible effect of this is to find motifs within what is the general noise associated with the meter readings. This is avoided by ignoring any motifs within a window which has a range of readings of less than a given size (100W for this analysis).

The flow of processing is shown graphically in Figure 5.6.

One issue arising is that normalisation means that the shape of the motif is matched without regard to the absolute values of the readings and this might not be valid. While small ranges of less than 100W are excluded, the actual size of the increase or decrease is arguably more useful when identifying particular activities (based on combinations of appliances) rather than the base shape. This can be addressed by normalising over the whole day of a household's readings rather then within the motif window and is the basis for including the normalised difference motif within the data collected.

While the motif matching approach based on shapes can, after normal-

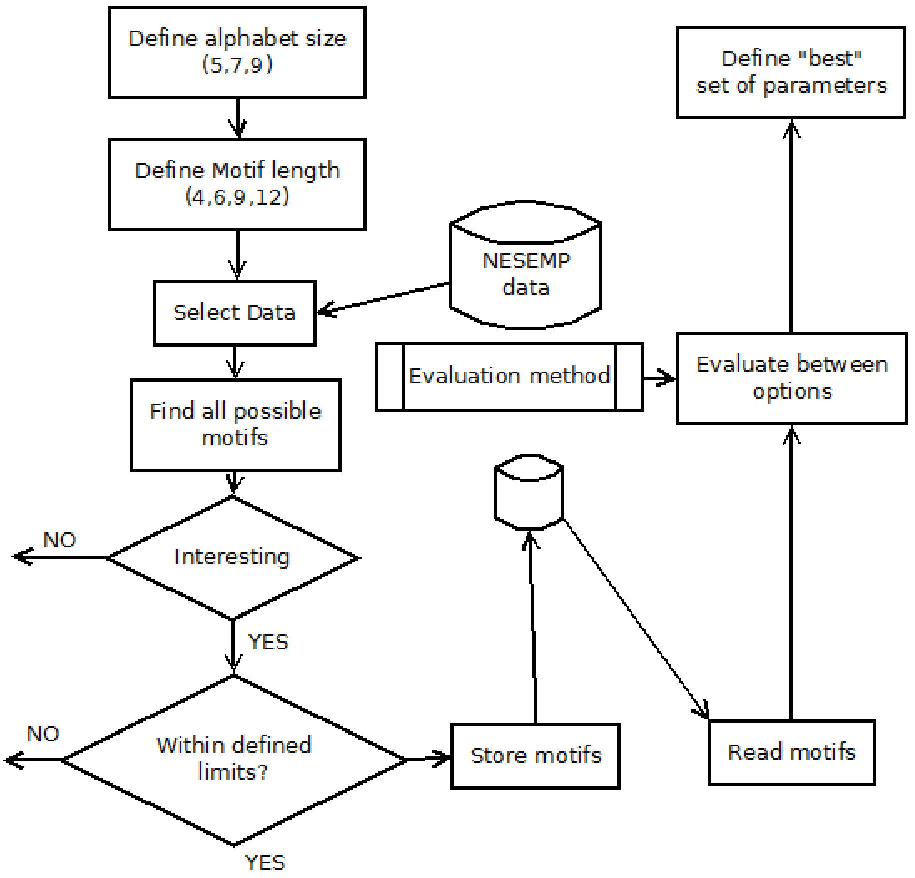


Figure 5.6: Flow of processing

isation within the analysis window, lead to two motifs being seen as identical when, in fact, one has a much larger range of power readings, this can be addressed by only considering motifs as similar when they fall into the same ranges of power usage.

As the analysis is to find activities during the day, these generally correspond to the use of certain individual appliances (or a combination of appliances) and thus the goal is to find motifs that map onto particular appliance combinations. The power usage characteristics of appliances in common use in the house are given by Kato et al. [12] (although for Japan rather than the UK).

The Centre for Sustainable Energy publishes a list of typical power usage figures for household appliances [1]. This table can be simplified into the following groups:

- High usage appliances such as power showers, over 5KW

[1] cse.org.uk/advice/advice-and-support/how-much-electricity-am-i-using

- Appliances generally used for heating - water or space - 3KW-5KW
- Laundry and larger cooking appliances 1KW-3KW
- Smaller and gardening appliances. Multiple lighting within rooms. 300W-1KW
- TV, lighting and audio equipment - under 300W

Note that fridge and freezers typically use up to 150W when running.

The ranges to be used can be set by selecting a reasonable number (e.g. five) and then splitting the range between the lowest and highest reading for a given house into the five ranges. Alternatively, range cut off points can be defined based on general industry standard values for power ranges of common appliances.

The suggested cut-off points in the list above can be used to group the motifs found into similar classes of appliance usage to avoid matching similar shaped patterns that actually cover much different ranges of usage. Five ranges are created, as per the wattage limits detailed above, and each motif falls into one of the ranges based on the overall spread of readings within the motif. For example, if the highest reading during the period of the motif is 4KW, and the lowest is 500W, then the range is 3.5KW and the motif is judged to fall into the range of appliances used for heating (3-5KW range).

5.3 Identifying "Interesting" Motifs

Many motifs are identified which may not be of interest to professionals making use of the results and these uninteresting motifs should be automatically excluded from the analysis. Decisions made on what is judged as uninteresting should be tested by reference to a team of electricity industry professionals.

Motifs arising from the small changes in the underlying electricity usage (the "noise") are excluded by setting a minimum range of 100W for a motif to be considered of interest. Any motif representing a behaviour which influences the electricity usage by less than 100W over the period of the motif is unlikely to be of interest to professionals designing DSM programmes.

As the motifs are created by shifting a moving window over the stream of data, overlapping periods are considered and a long period with no activity, except for one change in meter reading, will lead to a series of motifs that are similar. For example, when using an alphabet of five, a long period of no activity except for a decrease of 200W will lead to motifs such as ccccca, ccccac, cccacc, etc. As only one of these is interesting for further analysis, the others are excluded by omitting any motifs that start with two or more "c". With differently sized alphabets the same approach is adopted by excluding the motifs that start with multiple "middle" letters.

If a motif consists solely of increases in electricity usage then it is considered to be of no interest. The analysis is attempting to find motifs that reflect complete behaviours and a motif showing only increases in usage would only represent the beginning of the activity. A useful motif will consist of both starting (turning appliances on) and finishing (turning appliances off) an activity.

The check for motifs that consist of increasing readings only is further extended by checking the letters within the symbolised string as well as considering the absolute meter readings. Thus a motif of "abbccd" is considered to be uninteresting as it shows a motif that is continually increasing. In practice, the absolute meter readings may show a small difference between, say, the second and third readings which both map to the symbol "b".

Similarly, a motif which consists solely of decreasing meter readings is also designated as uninteresting.

Finally, as future analysis will be done by considering the timing of motifs within a day, any motifs that span midnight are excluded. While this can be criticised as a limitation, it is unlikely that electricity professionals will be considering programmes to change household behaviour around midnight and this limitation is unlikely to affect the usefulness of the results.

5.4 Evaluation of Different Techniques

The approach to finding motifs detailed above makes use of a number of parameters such as the type of motif to use (compressed, normalised, etc.), the size of alphabet to use, and the length of the motif. To choose between the various parameters a method is needed to select which set is most appropriate.

A useful behaviour to focus on for DSM would be one that repeats within a household on the majority of days under consideration and which is of a significant size in terms of power used. It should also be one that only occurs once or a few times per day. An example would be cooking which may only be done once during the conventional time period for the particular meal (e.g. 5pm to 8pm for the evening meal). However, with multiple members of the household, activities may repeat on the same day - e.g. when different groups of the household do their cooking separately. Another example would be that of using the shower which may be a regular behaviour for each member of the household and may thus repeat a number of times during the day as each person has one or more showers.

To detect a repeating activity a reasonable number of motifs need to be found. Some houses will have no repeating activity (their behaviour shows no regularity) and thus no useful motifs may be found. Other houses need to be evaluated in such a way as to give a reasonable number of motifs. The question as to "what is reasonable" needs to be addressed. Activity should happen on a sizeable proportion of the days under consideration to be seen as a regular activity.

One approach to evaluation is to consider the x most popular motifs within a household (i.e. the most popular is the one occurring most often, the second the one occurring second most often, etc.) and then see how the number of occurrences vary between the x motifs. For example, considering the 10 most common motifs each can be assessed by a useful measure (e.g. number of occurrences) and a graph similar to that of Figure 5.7 can be produced. The points marked as X, Y, and Z need to be defined in terms of reasonable numbers and then the results from the two sets of parameters displayed on the graph can be assessed against each other.

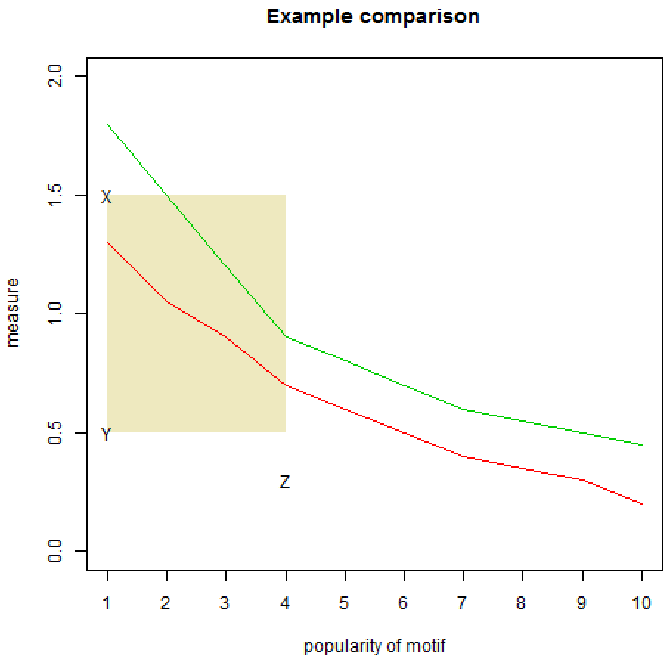


Figure 5.7: Example of motif evaluation

The point Z sets the number of popular motifs that are to be considered. While the graph shows the results for the top ten motifs, it may be considered that concentrating on the top Z motifs is sufficient and any results to the right of Z can be ignored.

The point X sets the value at which the measure (e.g. number of occurrences per day) is judged to be too high. Finding motifs that occur many times a day (say, ten times) is unlikely to be useful for finding behaviours that occur occasionally and can be influenced.

The point Y gives a value for the low point of the measure below which motifs are too infrequent to be useful.

These three points define the shaded area which is of "interest" and where the useful motif results will lie within. Whichever of the lines on the graphs falls more completely within the shaded area is considered to be the better set of parameters to use for future analysis. In the example, the red line would be judged to give better results as although the green line shows higher values, part of the line falls outside the preferred shaded area and can be considered to give too high results (e.g. too many occurrences per day).

There are many different measures that could be used to distinguish between the motif finding techniques. This analysis makes use of the following:

- The mean number of times that the motif occurs, per day, within the data period. This should not be too high (point X in the example) nor too low (point Y).
- The number of days that include the motif. There is no maximum value (point X) for this measure as a motif falling on every day would be useful. The value for a minimum number of days (point Y) needs to be selected.
- The percentage of days that include the motif. This differs from the second measure as it incorporates the number of days of readings for the household into the calculation. Not all households have a set of readings for every day within the period of analysis.

The relevant values for X, Y and Z need to be set for each of the measures being considered. The values chosen are detailed below. The analysis set out in the remainder of the thesis can easily be adjusted for differing values of X, Y and Z and could be explored in future work. For this analysis, arbitrary values are selected (with justification below) and other values could sensibly be selected.

For each measure, the top three motifs only are considered (i.e. point Z). The value of three is selected as a value between concentrating on too few motifs (e.g. if set as one then only a single motif for each household would be considered), and allowing too much influence to motifs that only occur occasionally for a household (e.g. if set as four or more).

When considering the average occurrences per day for each motif, a value for X need to be selected that exclude motifs that happen too frequently during the day and which may reflect regular behaviours which the household members may have little opportunity to influence. This could include regular automatic activities (e.g. fridge turning on and off) or behaviours which are necessary for efficient household functioning (e.g. turning lights on and off). A value for X of 2.0 is selected.

In a similar way, motifs that occur infrequently would be of little interest as they do not reflect a regular behaviour and hence Y should not be set too low. A value of Y of 0.3 is selected, reflecting that motifs occurring on average on 30% of the days is of interest.

Some motifs may occur many times on a few days but rarely on other days and the analysis of the number of unique days that contain the motif is used to identify this situation. The date of occurrence of each motif is found and the number of unique days for that motif is found and then averaged across all the households. In this situation, a motif that occurs on every day under analysis would be of interest and hence X should be set to a maximum value. The value of Y depends on the number of days in the period of analysis and, when considering a calendar quarter of working days (i.e. a maximum number of days of 65), a value for Y of 10 is selected.

Some households do not have readings for every day within the period of analysis and the percentage number of unique days with a given motif provides a measure that takes into account the number of days of analysis. In this case, a figure for X is selected of 90% which excludes the motifs that occur on nearly every day and may be considered to be too frequent and likely to reflect regular activities which the household has no control over (e.g. fridges). The value for Y is set at 20%.

The values assumed for each measure are summarised in Table 5.2.

Table 5.2: Values assumed for each measure

Measure	**X**	**Y**	**Z**
Motifs per day	2	0.3	3
Total days with motif	65	10	3
Percent of days with motif	90	20	3

5.5 Exploring the Different Ways of Treating the Motif Range

Three approaches to handling the range of the motif (the difference between highest and lowest meter readings during the period of the motif) are considered. These are:

- No consideration of range. Each motif with a similar shape, irrespective of actual values of the meter readings, is considered as the same motif. This approach matches similar shapes but has the possible downside of equating a motif with a relatively small range of readings (e.g. 200W) with the same shape over a relatively large range (e.g. 2KW). It is unlikely that these correspond to the same activity which would generally relate to the use of particular appliances and which would have particular electricity usage characteristics (e.g. Wattage ratings).
- Using the range of meter readings within each house under consideration. For each house, the maximum and minimum values for the range of each motif are found. 5 ranges are selected and these are set as being equally spread between the low and high motif range values. For example, for a house with motifs with a maximum range of 1100W and a minimum range of 100W, the ranges are set as 100-300W, 300-500W, 500-700W, 700-900W and 900W to 1100W. Motifs are only considered to be the same if, as well as having the same shape, the range of the motifs match.
- Based on typical appliance usage characteristics, the range boundaries are set as 300, 1000, 3000, 5000, and 60000W and the motifs are treated as above such that motifs are only judged to match when their shape is similar and they fall within the same range.

To assess the different approaches to dealing with ranges, the motifs using each approach are found for various values of alphabet size (5,7,9) and motif length (4,6,9,12). The results for each of the motifs detected (using the base data, using differences between readings and the compressed version of each) are plotted with the results shown at Figures 5.8, 5.9, and 5.10 for the percentage of days where the motif is found. The other measures (motifs per day and number of unique days) are also considered and the results are similar (graphs not included for space reasons). The graphs show the results for each value of alphabet size (5,7,9) down the vertical and each value of motif size (4,6,9,12) across the horizontal.

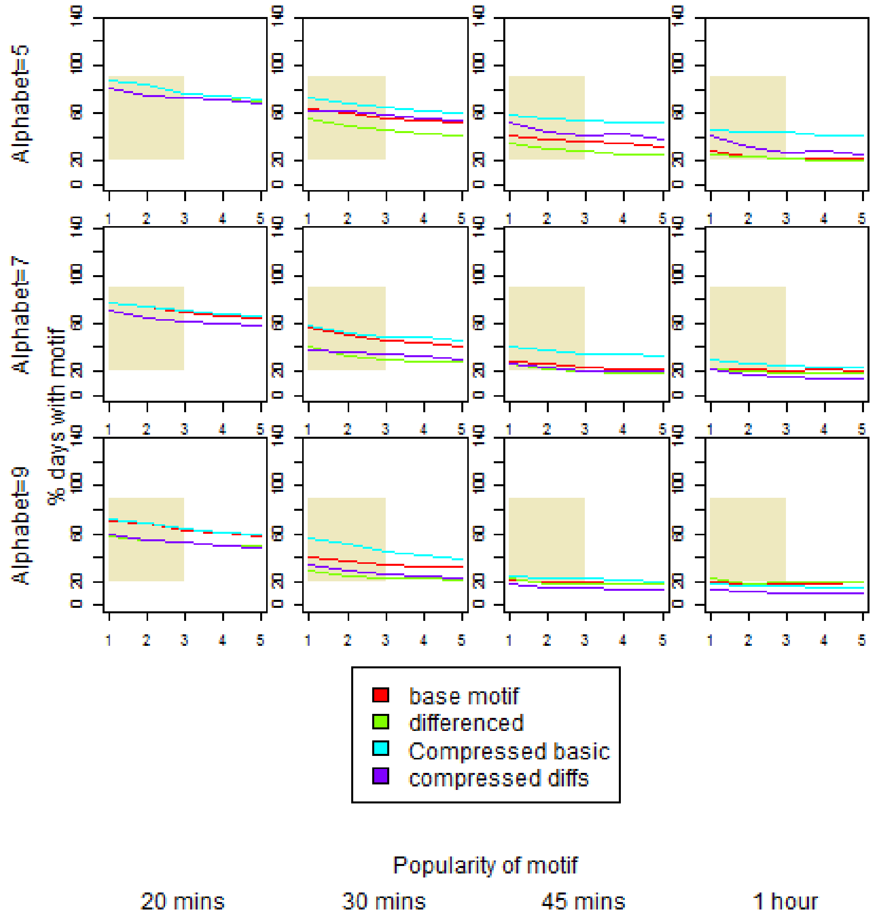


Figure 5.8: Percentage of days; Results with no range adjustment

Except for the largest values for motif size and alphabet size (where, intuitively, fewer motifs would be expected) all approaches to dealing with the range provide results within the acceptable and useful area (the shaded areas on the graphs). As the approach of making use of the ranges of possible appliances (the third approach) is intuitively more appealing, this option is taken for future analysis.

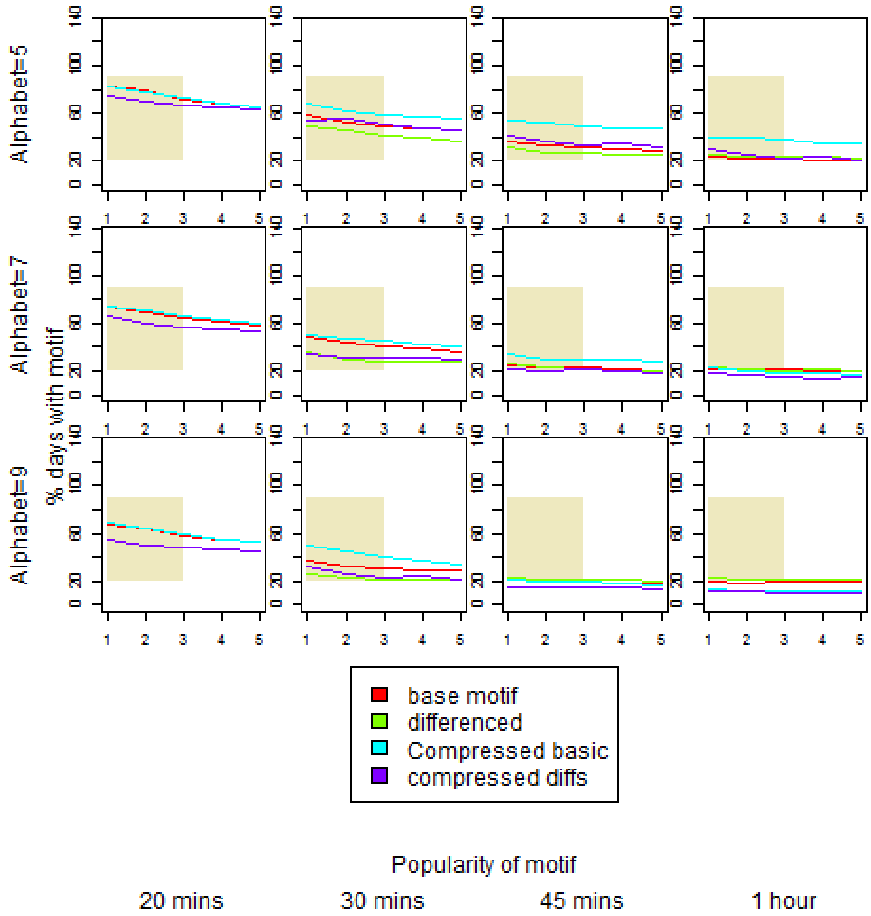


Figure 5.9: Percentage of days; Results with within house range adjustment

5.6 Exploring the Different Motif Representations

The motifs found within the data can be represented in various ways and the following are considered:

- The basic readings are translated directly into the symbolic representations.
- The differences between successive readings (i.e. the change in usage from time point to time point) are translated into symbolic representations.
- The basic readings are compressed such that adjacent letters that are the same in the symbolic representation are replaced with a single

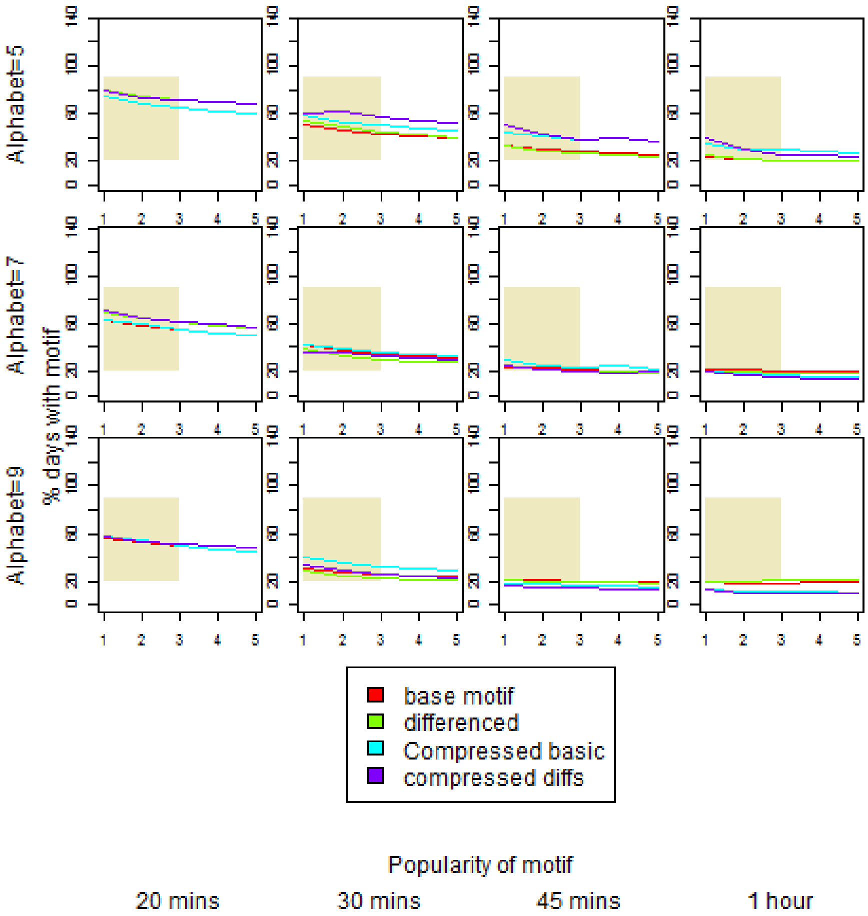


Figure 5.10: Percentage of days; Results with appliance range adjustment

letter - e.g. motif abccce becomes abce.

- The differenced readings are compressed as above.

To select the best way of representing the motifs the results for various motif sizes and alphabet sizes are plotted. Appliance based ranges (as described above in Section 5.5) are used. The results of the analysis are shown in Figures 5.11 and 5.12. The results for number of unique days is omitted for space reasons but gives similar results.

The compressed versions of the motifs (both basic and differenced) generally show higher values on the graphs. This would be expected as the compressed version of the motif is a more general representation of the motif than the original motif. For example, both motifs abccce and abbcce will, when compressed, be abce. In general, higher values on the

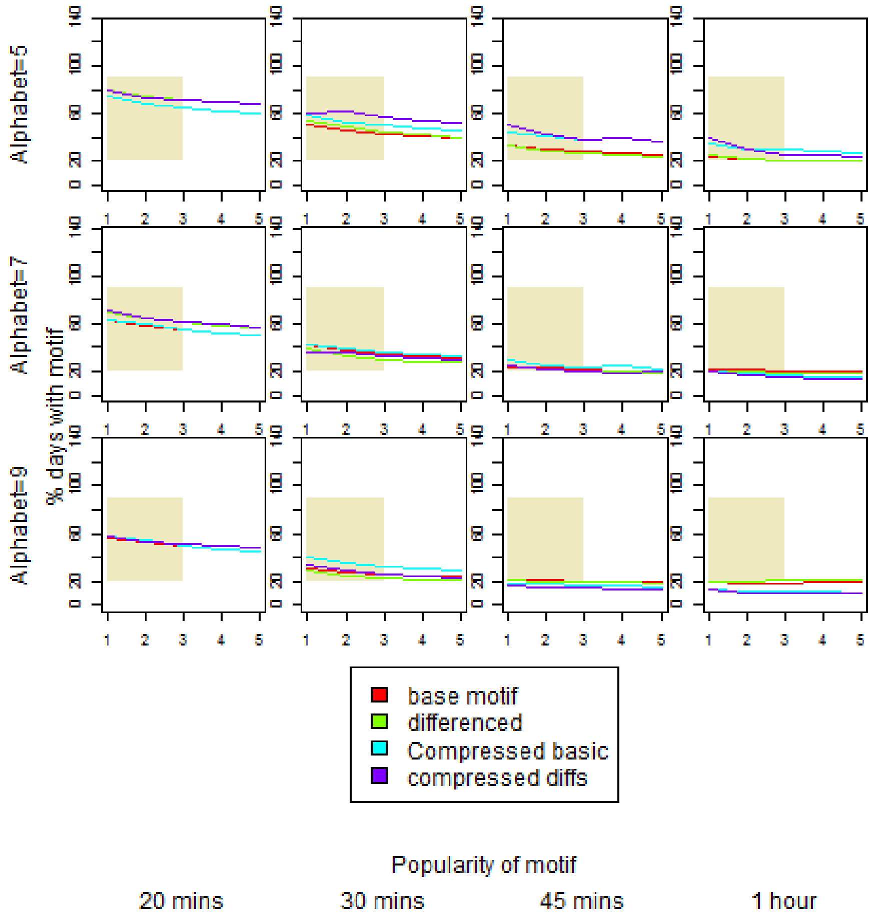


Figure 5.11: Percentage of days with different motif representations

graphs are preferred unless the values become too high and fall outside the shaded areas.

With the exception of the larger values for alphabet size and motif size, the compressed motifs fall within the useful shaded area on the graphs. The other exception is for the smallest motif size and alphabet size where the compressed basic motif shows too many motifs a day to be optimal.

There is little to choose between the results for the two compressed motif representations. As the differenced data is more closely related to changes in electricity usage during an activity (i.e. turning an appliance on or off) this is selected for further analysis.

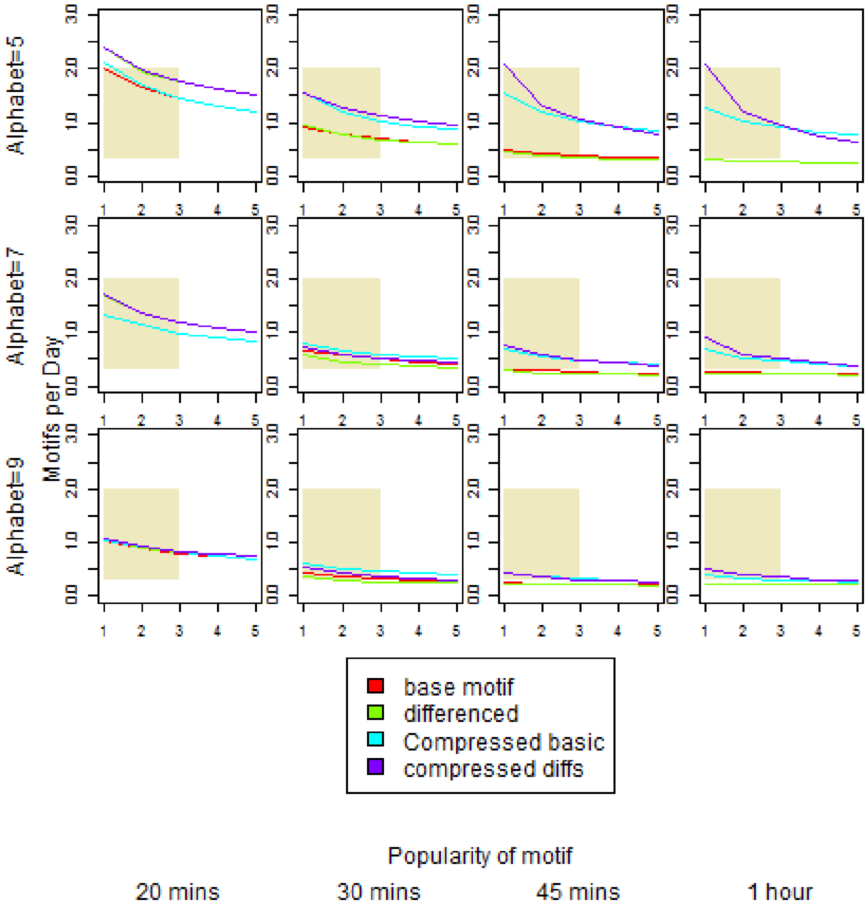


Figure 5.12: Motifs per day for different motif representations

5.7 Evaluating the Normalisation Methods

Options considered for normalisation of the data prior to conversion to symbolised representation are:

- Each motif is considered and the values within the motif are normalised before conversion to a symbolised string.
- Each household is considered and all the readings for that household are normalised prior to consideration of motifs.

The results from considering the two approaches to dealing with normalisation are shown in Figures 5.13 and 5.14. Again the results for the number of unique days is omitted for space reasons.

The motifs per day results show that the normalisation over the whole period for each household produces a lot of motifs per day and, in fact,

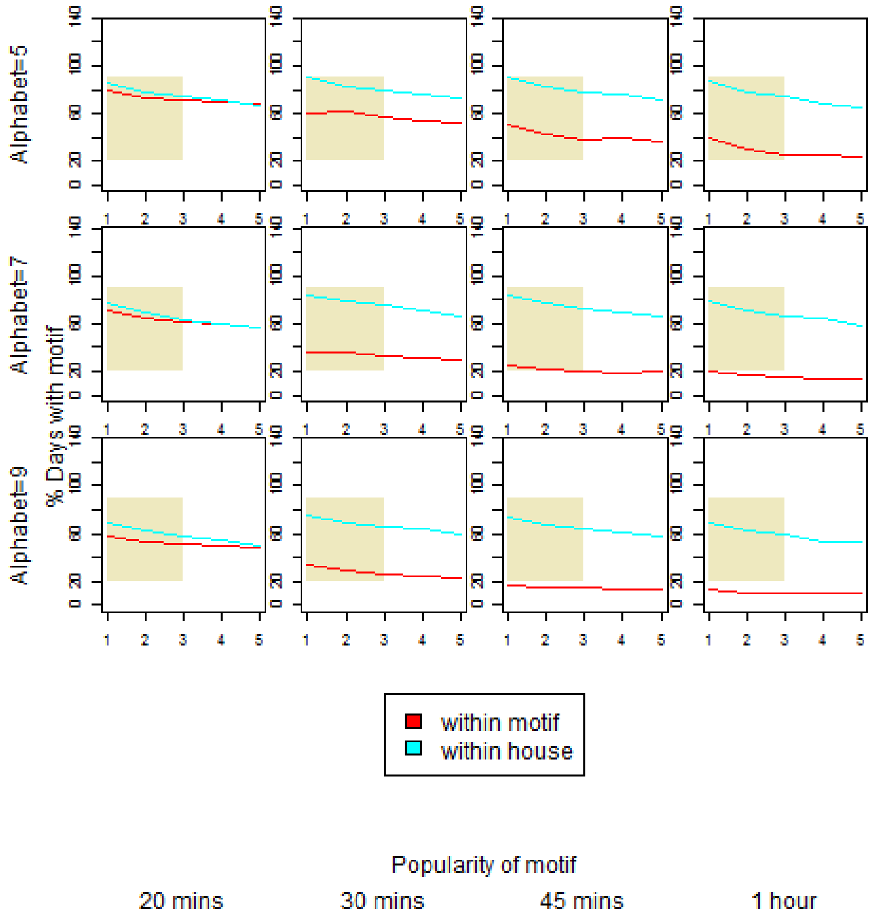


Figure 5.13: Unique days (%) for different normalisations

more than the maximum 2 per day that is considered useful as representing activities that occur occasionally and are likely to be able to be influenced. For that reason, the normalisation within a motif is considered the better approach and is selected for further analysis.

5.8 Selecting the Motif Size

As the motif size increases, the number of possible combinations of letters within the motifs increases and the number of motifs found reduces. The requirement for a useful size of motif is to provide useful numbers of motifs (but not too many). To analyse possible settings, results for motifs of size 4, 6, 9, and 12 (corresponding to 20 minutes, 30 minutes, 45 minutes and an hour) are each used with alphabet sizes of 5, 7 and 9. The results are shown in Figure 5.15.

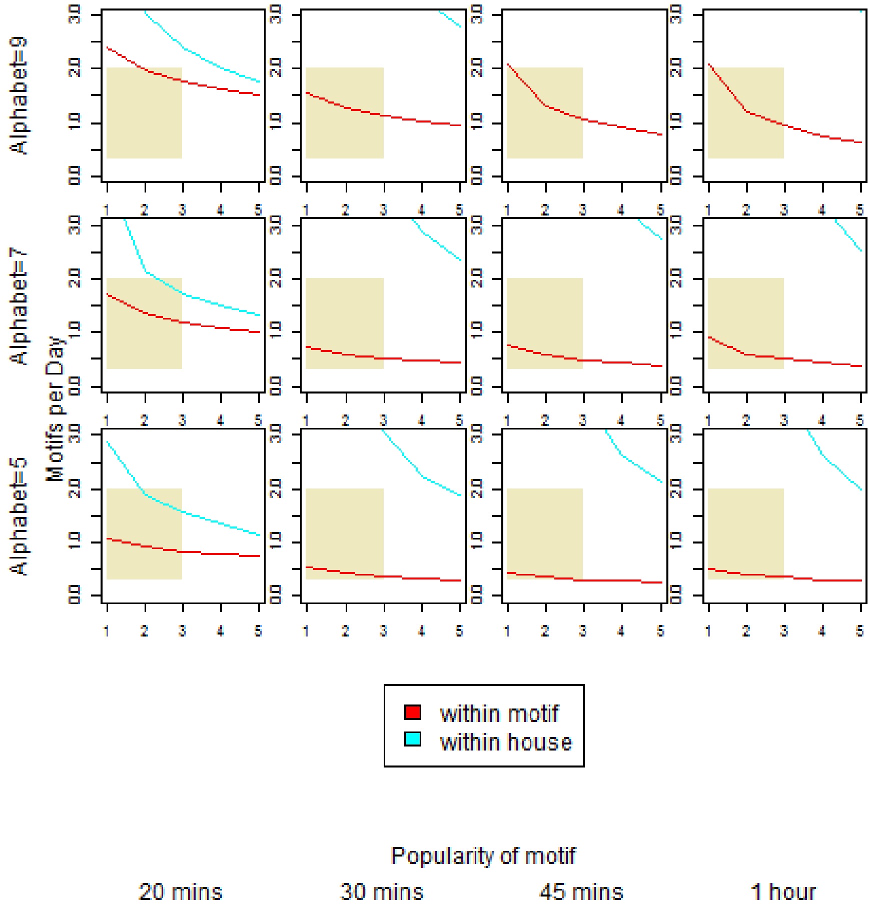


Figure 5.14: Motifs per day using different normalisations

For smaller values of alphabet size, the 30 minute motifs give more useful results. However, for larger alphabet sizes the 20 minute motifs are better. To investigate this further, both the 20 minute and 30 minute motifs will be considered when evaluating the alphabet size.

5.9 Selecting the Alphabet Size

As the alphabet size increases, the number of possible combinations of letters increases and the number of times motifs repeat will be expected to decrease. To evaluate the best settings for alphabet size, each of 5, 7, and 9 letters are considered with the motif sizes of both 4 and 6 (20 minutes and 30 minutes). The results are shown in Figure 5.16.

The results show that two combinations of settings are likely to provide

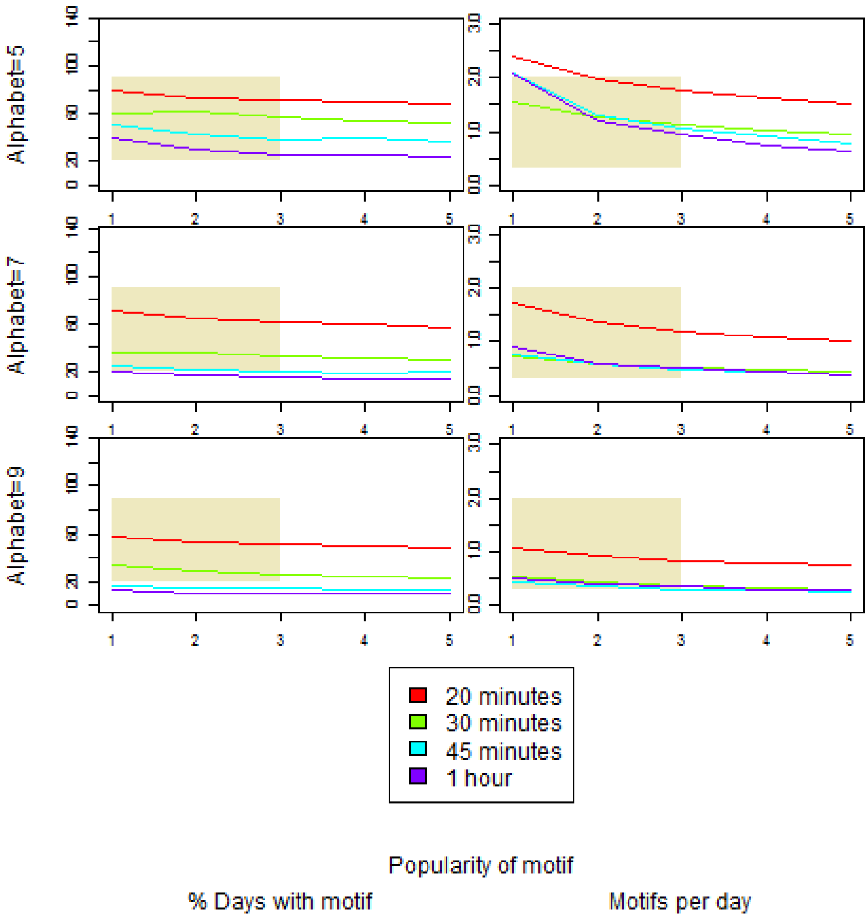


Figure 5.15: Different motif sizes

useful motifs for further analysis.

When using a small number of letters (5) in the alphabet when symbolising motifs, the best motif size to use is 6 equating to 30 minutes. However, an alphabet size of 7 is best used in combination with a motif size of 4 (20 minutes). Each of these combinations will be considered in further analysis.

5.10 Motifs Found

The 30 minute and 20 minute motifs found have been investigated to explore common motifs across multiple households. The charts at Figures 5.17 and 5.19 show the number of households that have the same commonest motif using the Spring weekday dataset.

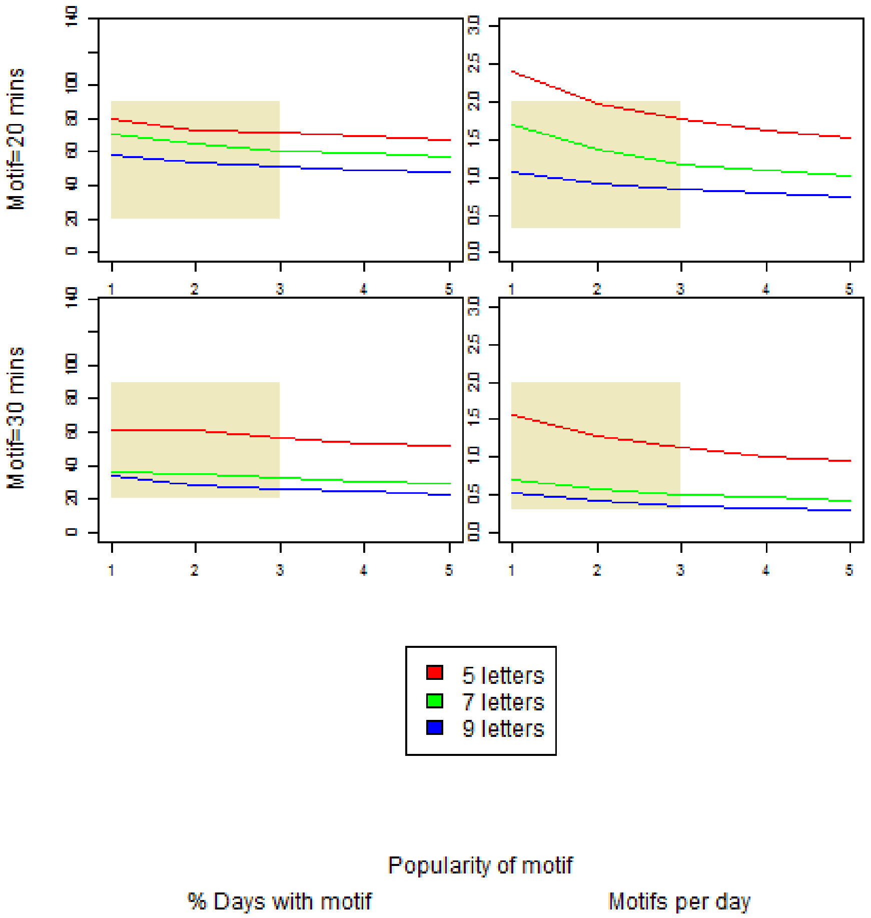


Figure 5.16: Different alphabet sizes

For the 30 minute motifs, it is clear that the top motif across a number of households is "ceac" when using an alphabet size of 5. This is a compressed motif and includes all the motifs that have one or more repeating letters such a "ceeac", "ceaaac", etc.. Some examples of the "ceac" motifs are shown in Figures 5.18 which, for ease of visualisation, shows a maximum of 10 days of readings for the selected houses.

Note that for the second example (household 286) there is an obvious repetition of the same activity at around 11am which hasn't been identified as the same motif as that at 7am and may demonstrate a shortcoming in the motif finding method.

Examples of households showing the 20 minute motif of "fb" are shown at Figure 5.20 which also demonstrates the effect of stratifying the motifs into similar ranges. For example, the results for household 561 show a number of peaks in the data where the range across the motif is fairly

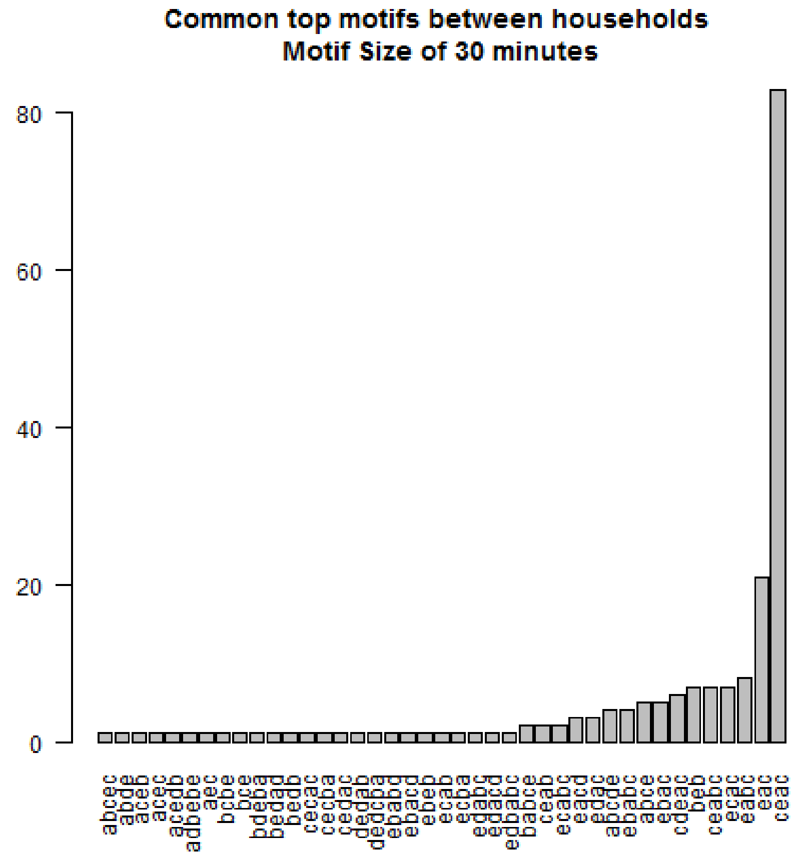
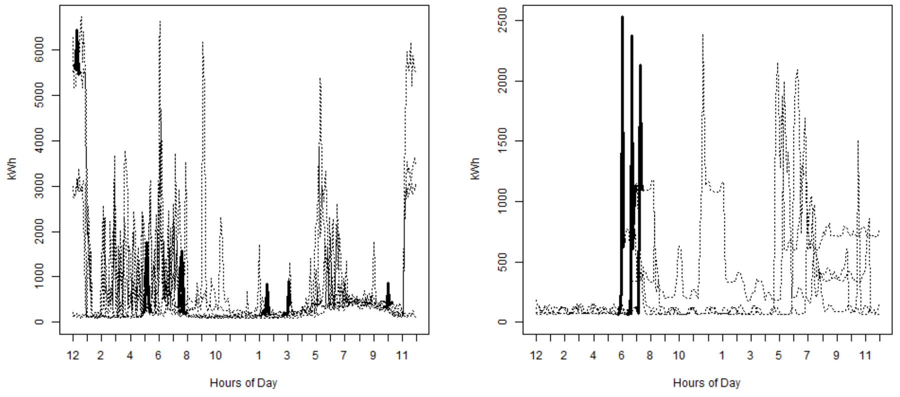


Figure 5.17: Common 30 minute motifs



(a) Household 517

(b) Household 286

Figure 5.18: Examples of households with "ceac" as most frequent 30 minute motif

small (less than 500W) whilst the similar shaped peaks elsewhere on the figure (but with larger ranges) are not detected as the same motif.

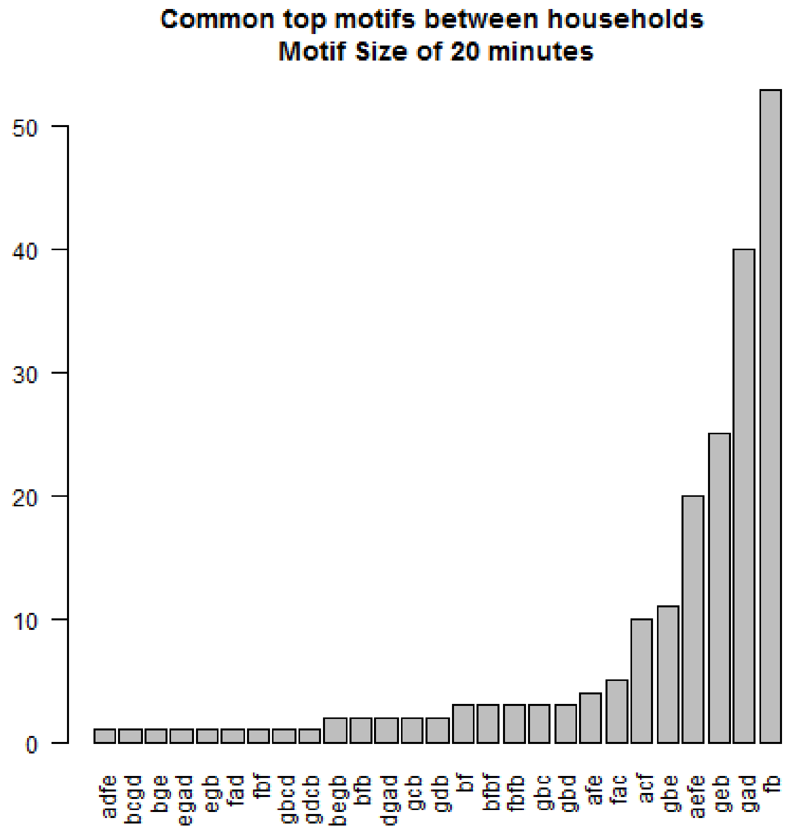
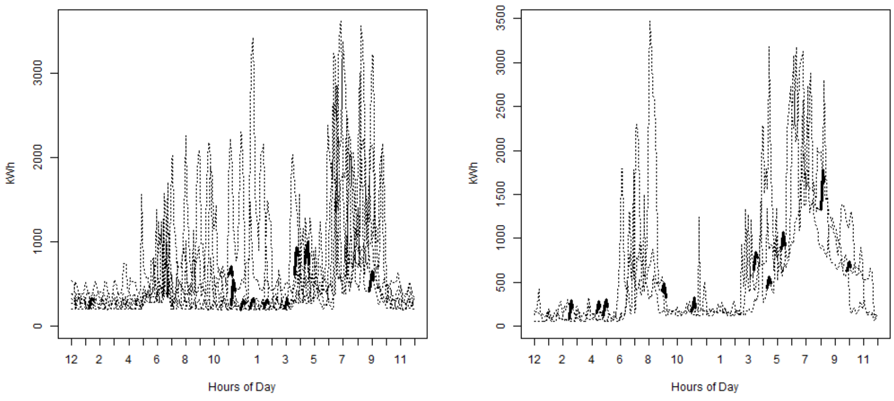


Figure 5.19: Common 20 minute motifs



(a) Household 571

(b) Household 561

Figure 5.20: Examples of households with "fb" as most frequent 20 minute motif

The analysis of the most common motifs across households is repeated using the Summer weekend dataset with the results shown at Figures

5.21 and 5.22.

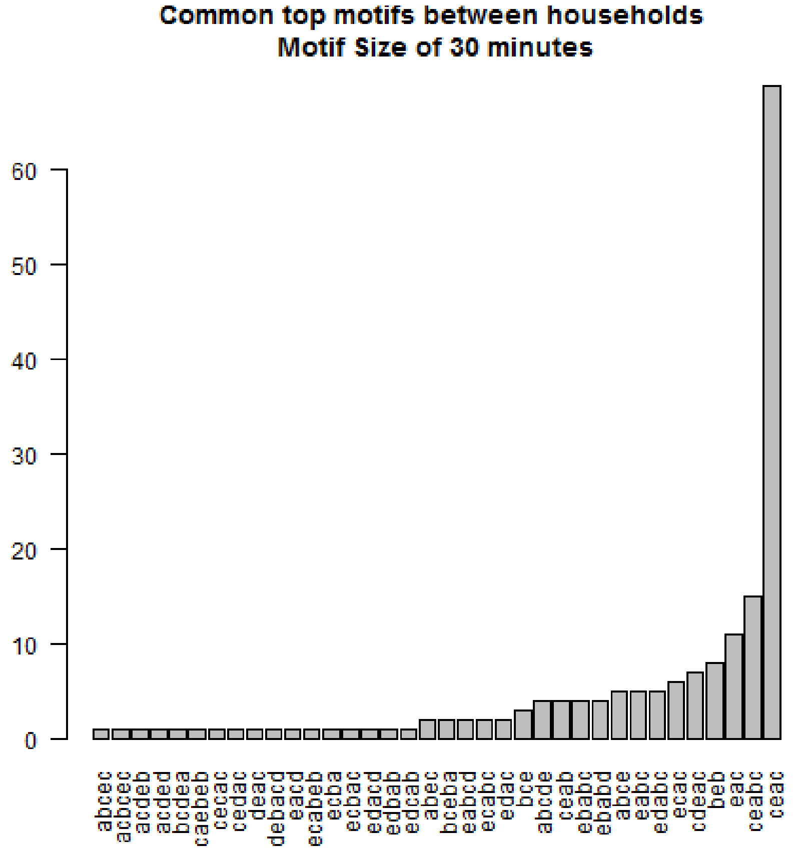


Figure 5.21: Common 30 minute motifs (Summer Weekends)

The most common motifs across the households show a high degree of similarity between the Spring weekday and Summer weekend datasets. Amongst the 30 minute motifs, the most common top motif across households is “ceac” for both datasets and most of the other top motifs feature in the results for both the Summer weekends and Spring weekdays.

For the 20 minute motifs, the 4 commonest motifs across both the Summer and Spring datasets are the same motifs.

With the short size of motif (20 minutes) only 4 characters are included in the motif (one for each 5 minute period) and, with removal of repeated letters, the resulting motifs are simple with the top motif for both datasets being “gf” representing just a single peak in electricity usage. More interesting analysis of motifs is likely to use longer motifs to detect more complicated behaviours.

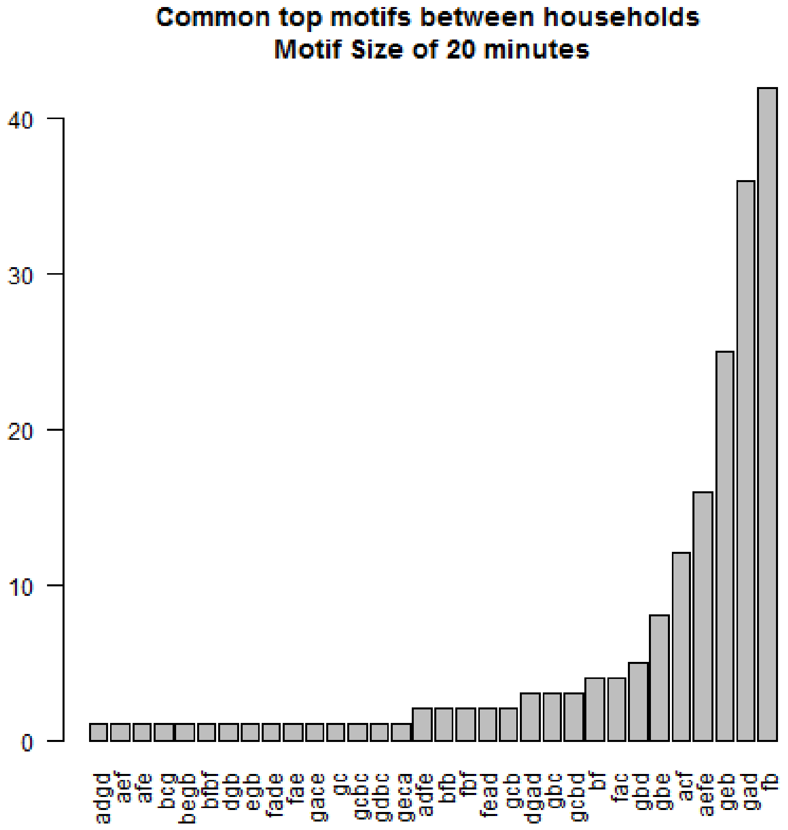


Figure 5.22: Common 20 minute motifs (Summer Weekends)

The results for the different datasets from Spring weekdays and Summer weekends show that the motifs found are generally similar and that the motif finding approach is equally valid for both datasets and is not dependent on peculiarities of the subset of data for a particular season or type of day (weekday or weekend).

The high number of households with the same motif occurring most frequently suggests that exploring the occurrence of the same motif at the same, or different, times across multiple households would be an interesting area for further study. This is listed as an area for future work in Chapter 7.

5.11 Conclusions

Based on the above analysis, a set of parameters is selected as the ones that are most likely to provide motifs that will potentially be useful for further analysis. The parameter settings are shown in Table 5.3.

Table 5.3: Parameter Settings selected for future analysis

Parameter	Setting	Meaning
Range	Using appliance characteristics	The typical ranges of appliances are used to distinguish between similar shaped motifs.
Data	Differences between consecutive readings	The differences between readings are used as representing changes in electricity usage.
Compression	Compressed data used	Repeating letters within the motif are removed.
Normalisation	Within motif	Values within the motif are normalised before being symbolised.
Motif size	4 (20 minutes) or 6 (30 minutes)	The length of each motif.
Alphabet size	5 or 7 letters (depending on motif size)	The number of letters used for the motif symbolisation step.

The combination of parameters has been shown to produce a usable number of motifs. The conclusion is drawn that using a sampling frequency of five minutes is sufficient for repeating patterns to be detected within the data streams and that sufficient repeating patterns can be found for the motif analysis to potentially be useful.

5.12 Summary

This chapter has addressed the question as to whether regular activities within the household can be identified solely using the electricity meter data captured. The analysis has shown that data sampled at a five minute interval is sufficient to provide a usable number of motifs which can be assumed to represent repeating activities within the household.

The symbolisation process is a suitable method for converting the real valued meter readings into a restricted set of characters with each representing a certain range of values. This provides for matching of data samples which, while not identical, map to the same symbols and can be assumed to be approximately the same.

A series of different parameter settings have been explored and a set of parameters selected that will be used in further analysis of the motifs in the following chapter.

The method of representing the motif in an alphabetised form has been explored with possible approaches of using the base meter data or differenced data. In addition, compression of repeating characters is considered.

Motifs may match but actually represent similar patterns within very differently sized data and different approaches to splitting the motifs into various bands of usage over the range of the motif have been explored with a solution based on a split into various appliance ranges (five ranges from small to large usage) selected as the best approach.

Various motifs that show uninteresting behaviour can be found (e.g. no change in usage over the time of the motif such as when the house is unoccupied) and an automatic method of rejecting some of the uninteresting motifs has been developed.

The next chapter will examine the point in time at which motif occurs and will then explore whether the variability in the timing of the motifs provides useful information on the variability of behaviour in a household.

An effective method of identifying meaningful motifs opens up a number of opportunities for electricity industry professionals to gain greater understanding of activities within given households. The chapter has not tested the implementation of initiatives to change behaviour but has only investigated whether reasonable numbers of motifs can be identified and established the most appropriate parameter settings. Now that these settings have been established, future work could explore the detection of behaviours and then test incentives to change these behaviours. This would be a very valuable step along the road of effectively deploying effective DSM initiatives.

CHAPTER 6

Using Variability to Define Archetypical Households

6.1 Background

This chapter finds the patterns in the stream of meter data that repeat and then examines how the times of these repeating patterns vary from day to day within a household.

The chapter describes the possible methods of assessing variability of behaviour and considers possible measures that do not use motifs. These are assessed and then compared with the approach of finding variability in the timing of motifs from day to day and the frequency of motifs from day to day. The results are compared with those arising from using just the load profile data and are found to provide complementary information.

Combining the partitions arising from the motif and the base meter data into a single set of partitions is explored but not found to be useful as the resulting combined partitions score lower on the evaluation measures and the results are hard to interpret by the layman.

This chapter addresses the question of whether making use of the variability of behaviour (as shown by the electricity meter data) provides "better" groupings of households for the purpose of demand side management (DSM) than those provided by using daily load profiles.

This chapter builds on the work in Chapter 4 which uses the traditional approach of using load profile shapes to create groupings of households and the analysis in Chapter 5 which has determined the most effective approach to searching for motifs within the meter data stream.

The approach taken in this chapter is to analyse meter data to find repeating patterns (motifs) which represent regular activities. The concept of "variability of behaviour" is used as the basis for clustering similar households and it is shown that this novel approach can complement the existing method of clustering using average load profiles. Differing clustering algorithms are considered with the quality of the clusters assessed. The hypothesis is that the added insight from the variability analysis allows for more efficient deployment of behaviour change interventions.

After finding clusters using the variability of timing of motifs within the period under analysis, the results are objectively assessed using both traditional cluster validity indexes and also the novel composite measure introduced in Chapter 4.

The data collected from the electricity meters of a number of households is used to develop measures to describe the flexibility of each household. These measures of flexibility are then input to a number of clustering algorithms. Once the most effective approach is determined, final clusters of households are calculated and these final clusters are validated against demographic and attitudinal data collected from each household. This demographic data is excluded from the meter data analysis and is used solely for validation purposes.

6.2 Methodology

6.2.1 Data Selection

A subset of the data is created for working days from Spring (March, April and May) 2011 where working days are weekdays excluding public holidays in Scotland. Not all households have a full set of meter readings. This data selection matches that used in the analysis in Chapter 4. The dataset represents approximately 2,653,000 individual meter readings spread between 122 households. The dataset differs by one household from that used in the analysis in Chapter 4 as House 528 is excluded due to only having a single day of valid data within the Spring period of analysis. Exploring how motif timings vary between days makes no sense when considering only a single day of data.

To allow comparison between the load profile and motif variability clustering, the results in Chapter 4 are recalculated in this chapter, using the modified dataset, alongside the motif variability results.

The motif analysis undertaken in Chapter 5 provides guidelines for the most effective parameters to use to find reasonable numbers of useful motifs within the datasets. These parameter settings (see Table 5.3) are used in this chapter.

6.2.2 Non-Motif Variability Clustering

Various different measures of flexibility of behaviour within the household can be defined [147] including:

- Calculating the variability of the time of peak usage during periods of interest.
- Assessing the variability of periods of peak usage (e.g. 15 minutes) during times of interest.
- Variability in the time of minimum usage.
- Detecting regular repeating patterns and calculating the variability in the times when these patterns occur.
- Creating various ratios such as the variability between each hour of the peak time analysed and the overall average usage during the peak time.

The list of possible measures to use to represent variability could be nearly endless and analysis of some of the suggestions is considered as possible future work.

6.2.3 Using Motifs to Assess Variability

As described in Chapter 5 the motifs in the data stream are found using the parameters assessed as most effective. Two different combinations of motif size and alphabet size (20 minutes/7 letters and 30 minutes/5 letters) are considered.

The top motif (the one that occurs most often within a household) is further examined for the times when the motif occurs. The number of times it occurs, and the standard deviation of the time of day of occurrence are calculated for each household. In a similar way, the second and third most common motifs within a household are identified and the variability in timing of each are also calculated.

Other useful measures relating to the motifs found within a household are also calculated including the number of different motifs (occurring at least twice during the Spring period) and the number of different motifs occurring on at least 30% of the days sampled for the household. The 30% figure is selected arbitrarily as a reasonable number to ensure only regularly repeating patterns are examined.

On investigation, it is found that some households have motifs that group around more than one period in the day. This is shown in an example for household 286 at Figure 6.1 which shows two obvious times around which the motifs cluster. Taking a measure of the spread around the overall mean will give a value that shows the variability across the full day. However, more useful measures would be related to how the motif timings vary around the means for the morning and the afternoon.

In this case, calculating the mean and the standard deviation around the mean is likely to overestimate the amount of variability that the household shows. While the standard deviation measure shows the variability within the whole day, a more intuitive measure would provide an estimate of how the motifs vary around the local (e.g. early morning) times. To provide an estimate in this case, the times of motifs for each household are further split into morning (am) and afternoon (pm) periods and the means and standard deviations calculated for each half day period. This leads to six further attributes reflecting the variation of the first, second and third motifs within the two half day periods.

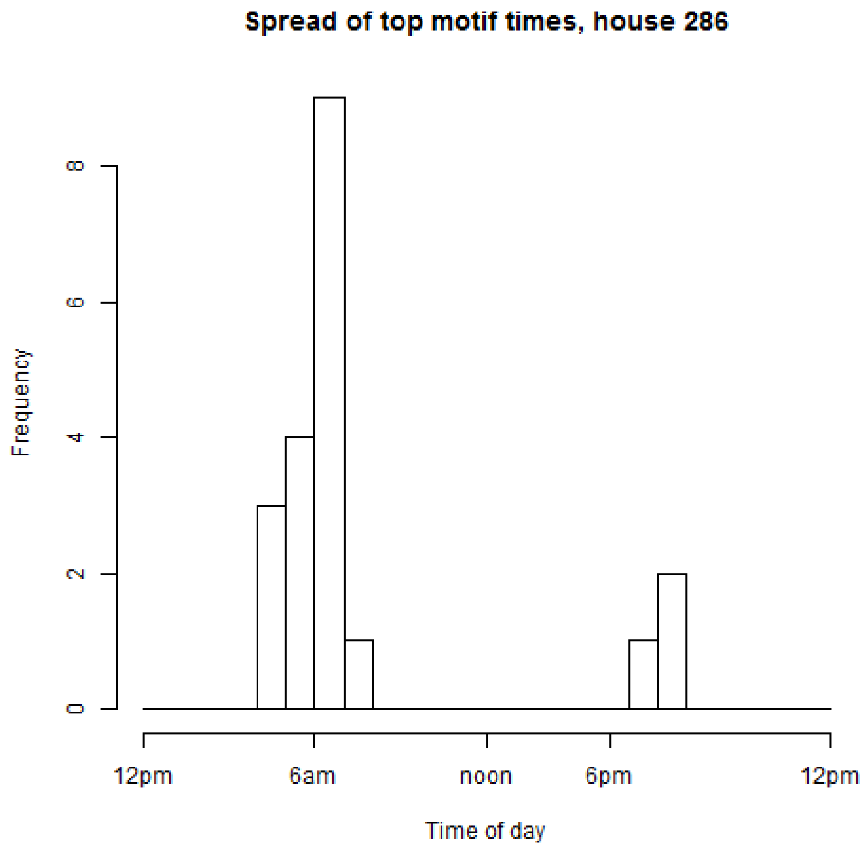


Figure 6.1: Example for 30 minute motifs

The attributes calculated for each household and used for the clustering are:

1. Number of occurrences of the motif occurring most frequently.
2. Number of occurrences of the second most frequent motif.
3. Number of occurrences of the third most frequent motif.
4. Total number of motifs found for the household that occur at least twice.
5. The variation in timing of the top motif during the morning period. This is calculated by examining the timing of the start of the motif and calculating the standard deviation of the timing (measured in minutes) around the mean start time over the morning.
6. The variation in timing of the top motif during the afternoon period.

7. The variation in timing of the second motif during the morning period.
8. The variation in timing of the second motif during the afternoon period.
9. The variation in timing of the third motif during the morning period.
10. The variation in timing of the third motif during the afternoon period.

The data generated from the exercise of searching for motifs, and calculating the variability in timing of the motifs, is used as the input to a clustering exercise to determine which clustering method produces the most useful clusters as assessed using the validity indexes.

6.2.4 Clustering using the Load Profile Data

Firstly, the work detailed in Chapter 4 is repeated using 6 clustering algorithms with the results assessed using the CDI and MIA cluster validity indexes and the novel composite measure. These results provide a baseline for assessing the motif variability approach detailed below.

The data is averaged to create a representative load profile for each household. For example, all the readings for 4pm for the household are averaged to create a representative (mean) reading for 4pm, similarly for 4:05pm, etc. The 122 representative profiles are then normalised within the 0 to 1 range.

The normalised profiles, each with 288 attributes, are then used as input to clustering exercises using a variety of clustering algorithms including kmeans, fuzzy cmeans, self-organised maps and hierarchical clustering (using Ward linkage) as these are the most commonly used as identified by Chicco [14].

A common issue with clustering exercises is the appropriate setting of a value for the number of clusters. To match common practice within the electricity industry, 8 clusters are selected for reasons discussed in Chapter 4.

6.2.5 Clustering using Timing of Motifs

While a number of statistics have been created that represent aspects of the motifs found within the data, only those statistics directly relating to the timing of occurrence of motifs are used in the clustering. Other statistics include the number of times a motif occurs within a household's data but, as this does not directly relate to variability of behaviour this is excluded. Hence, the attributes for the variability in timing of the first, second and third most popular motifs are used as the basis for the clustering.

The data generated from the exercise of searching for motifs, and calculating the variability in timing of the motifs, is used as the input to a clustering exercise to determine which clustering method produces the most useful clusters as assessed using the validity indexes.

The motifs found are compared on the basis of similar shape without regard to absolute value of the data except where the motifs are banded by range. For example, all motifs with a large range are considered together while those with a small range are considered together. Motifs with a similar shape but a differently sized range of values are considered to be different. To avoid the problem of finding motifs within what is the general noise associated with the meter readings any motifs within a window which has a range of readings of less than 100W are ignored.

Distance Measures

Extensive tuning of the clustering algorithms is not one of the goals of the thesis as the focus is on evaluating the difference in the results from using the load profile information as opposed to the motif variability data. However, a brief investigation to explore the use of different distance measures has been considered.

Not all the clustering algorithms are suitable for use with different distance measures. In particular, kmeans relies on calculating the mean of the various instances allocated to each cluster and is designed to operate using euclidean distance.

Hierarchical clustering allows for the use of different similarity matrices when forming clusters and various different methods of measuring similarity can be included. For the purposes of an initial investigation, the results arising from using Mahalanobis distance measures are compared with those from euclidean distance measures. Hierarchical clustering with the Ward agglomeration method is used to form three clusters using the variability of motif timing data. The composite CVI values are calculated for each method and the results are shown in Table 6.1.

Table 6.1: Comparing distance measures

	Cluster Sizes	Overall	Compact	Parsimonious	Substantial	Consistent
Euclidean	23,47,52	0.730	0.456	1	0.855	0.608
Mahalanobis	16,51,55	0.640	0.237	0.805	0.934	0.582

The use of the Mahalanobis distance measure produces poorer results than using euclidean (as measured by the composite CVI). However, it is clear that using the different distance measures produces significantly different results and, when tuning the clustering algorithms for optimum performance, is an area that will require attention.

6.2.6 Datasets to Use

The attributes that are available for use during clustering fall into three separate groups as follows:

- The meter readings for each of the five minute times during the period of analysis. This consists of 288 readings for the full 24 hour period. This provides a load profile dataset.
- The motif variability data generated by considering the variability in timing of the 1st, 2nd and 3rd most common motifs within a household. This variability in timing is extended to consider subsets of motifs just within the morning or the afternoon period leading to a total of six values per household. This provides the motif variability dataset.
- The motif frequency data that is obtained by counting the numbers of motifs per household. By considering the total amount and the

number of each of the three most popular, a total of four variables are available. This provides the motif count dataset.

Each of the three datasets is considered and the clustering, using the six algorithms previously defined, is undertaken with the results being assessed using the composite measure.

The result of this approach is that each household is allocated to an archetypical group within each of the three datasets. Thus a household is within a group that has a similar load profile of usage across the day, also within a group that has a similar degree of variability, and also within a group that has a similar number of repeating behaviours. From the membership of these groups, analysts can define a group of interest for a particular demand side management intervention (e.g. the households with a peak at early evening and that show a high degree of variability).

The load profile clustering uses a cluster number of eight (as discussed in section 4.3.1). However, the clustering based on variability of behaviour and the degree by which behaviour repeats use a cluster number of three. Three is selected for the number of clusters for the motif related clustering as the partitions can be easily understood. Households will fall into one of the three archetypes of high, medium or low variability and thus appropriate incentives can be designed to address a particular archetype.

6.2.7 Evaluation

To assess the quality of the clusters derived from one clustering exercise compared to those derived from a different exercise, the widely used CDI and MIA measures (as defined by Chicco [14]) are used. Lower values for the CDI and MIA measures denote "better" solutions.

In addition, the composite measure described in Chapter 4 is used to provide a measure of the overall usability of the results.

The results using different clustering algorithms for each set of data are combined using ensemble methods to provide an overall assessment of the usability of each of the three sets of data considered. The composite measure for the combined partitions is calculated with a higher value providing results that are preferred.

Combining all the results is also evaluated (i.e. all clustering algorithms across all datasets) to provide a single set of partitions which again is evaluated using the composite measure. However, the results from this are hard to interpret in real life terms as attributes relating to absolute meter readings are combined with motif timings and frequencies and the resulting archetypes do not provide easily understood example households.

6.3 Results

Table 6.2 shows, for each of the clustering algorithms used and for each set of data, the sizes of the partitions in the solution and the values for the MIA and CDI cluster validity indexes (lower is better). The MIA and CDI validity index calculations for the motif clustering are not comparable with the load profile clustering due to the differing set of attributes used. The sizes of the clusters have been sorted into ascending order.

It can be seen from the MIA and CDI values that the kmeans, SOM, hierarchical and mixed model techniques produce similar quality solutions using the load profile dataset with the Fuzzy Cmeans algorithm being significantly poorer. The random forest and pam combination provides a good result for the MIA measure (compactness within the cluster) but scores poorly on the CDI measure which includes both compactness and separation between the clusters.

When using the motif variability data, the kmeans, SOM and mixed model algorithms produce similar quality results with the Fuzzy Cmeans algorithm again producing poorer results. The hierarchical algorithm gives results that are worse than the kmeans and SOM algorithms, and the random forest/pam combination produces poor results. The ranking of algorithms using the CDI for the motif variability clustering is consistent across the two sizes of motifs (20 minute and 30 minute).

Table 6.3 shows how the results for each set of data compare across differing algorithms with the higher the value of the Rand index, the closer the match of partitions produced by the algorithm.

The results for the Rand index show that the values are consistently closer

Table 6.2: Clustering Results and Validity indexes

	Cluster sizes	MIA	CDI
Load Profiles			
Kmeans	4,8,13,13,16,17,20,31	1	0.445
Fuzzy Cmeans	1,2,12,20,26,30,31	1.706	0.517
SOM	4,9,14,16,18,19,20,22	1.103	0.447
Hierarchical	3,9,10,11,15,21,24,29	1.062	0.451
Model	3,7,9,9,20,20,26,28	1.058	0.451
RF	12,12,14,14,14,16,19,21	0.858	0.643
20 minute Motifs - variability			
Kmeans	7,45,70	2.448	0.878
Fuzzy Cmeans	26,36,60	4.204	0.914
SOM	7,44,71	2.416	0.88
Hierarchical	34,38,50	3.955	0.918
Model	7,48,67	2.393	0.913
RF	20,36,66	2.312	1.099
30 minute Motifs - variability			
Kmeans	24,48,50	3.244	0.932
Fuzzy Cmeans	30,43,49	3.951	0.953
SOM	18,50,54	3.045	0.936
Hierarchical	23,47,52	3.3	0.949
Model	35,42,45	3.811	0.97
RF	31,34,57	3.183	1.122
20 minute Motifs - occurence counts			
Kmeans	7,34,81	1.571	0.621
Fuzzy Cmeans	9,49,64	1.665	0.623
SOM	7,34,81	1.601	0.622
Hierarchical	31,34,57	2.684	0.7
Model	26,33,63	3.336	0.796
RF	29,37,56	3.115	0.763
30 minute Motifs - occurence counts			
Kmeans	4,34,84	1.138	0.606
Fuzzy Cmeans	28,33,61	3.746	0.801
SOM	4,34,84	1.129	0.606
Hierarchical	4,25,93	1.145	0.617
Model	4,47,71	1.378	0.737
RF	24,42,56	3.743	1.043

to 1 in the case of the clusters built using the load profile information rather than the motif variation information. The mean value for the Rand index (after omission of the values on the diagonal) are 0.441 for the load profiles and 0.379 (30 minutes motif) and 0.397 (20 minutes motif) for the motif variability approach. This shows a less consistent set of partitions (across the different clustering algorithms) are created when making use of the variation information than the partitions created using the load profile information.

Table 6.3: Modified Rand index using differing clustering algorithms

	Kmeans	Cmeans	SOM	Hierarchical	Mixture	RF
Load Profiles						
Kmeans	1	0.372	0.66	0.458	0.497	0.373
Fuzzy Cmeans	0.372	1	0.296	0.279	0.316	0.281
SOM	0.66	0.296	1	0.528	0.572	0.501
Hierarchical	0.458	0.279	0.528	1	0.798	0.301
mixture	0.497	0.316	0.572	0.798	1	0.376
RF	0.373	0.281	0.501	0.301	0.376	1
20 minute Motifs - variability						
Kmeans	1	0.589	0.969	0.435	0.442	0.377
Fuzzy Cmeans	0.589	1	0.575	0.57	0.43	0.549
SOM	0.969	0.575	1	0.431	0.423	0.398
Hierarchical	0.435	0.57	0.431	1	0.552	0.38
mixture	0.442	0.43	0.423	0.552	1	0.364
RF	0.377	0.549	0.398	0.38	0.364	1
30 minute Motifs - variability						
Kmeans	1	0.851	0.829	0.751	0.416	0.291
Fuzzy Cmeans	0.851	1	0.762	0.659	0.468	0.32
SOM	0.829	0.762	1	0.668	0.368	0.321
Hierarchical	0.751	0.659	0.668	1	0.418	0.28
mixture	0.416	0.468	0.368	0.418	1	0.425
RF	0.291	0.32	0.321	0.28	0.425	1
20 minute Motifs - counts						
Kmeans	1	0.539	1	0.336	0.171	0.283
Fuzzy Cmeans	0.539	1	0.539	0.263	0.155	0.266
SOM	1	0.539	1	0.336	0.171	0.283
Hierarchical	0.336	0.263	0.336	1	0.472	0.731
mixture	0.171	0.155	0.171	0.472	1	0.459
RF	0.283	0.266	0.283	0.731	0.459	1
30 minute Motifs - counts						
Kmeans	1	0.518	1	0.728	0.19	0.144
Fuzzy Cmeans	0.518	1	0.518	0.431	0.257	0.271
SOM	1	0.518	1	0.728	0.19	0.144
Hierarchical	0.728	0.431	0.728	1	0.109	0.071
mixture	0.19	0.257	0.19	0.109	1	0.211
RF	0.144	0.271	0.144	0.071	0.211	1

Applying the composite measure of usefulness introduced in Chapter 4 provides the results shown at Figure 6.2 for each of the load profile clusters, the variability of motif data sets (20 minute and 30 minute motifs) and the motif count datasets (both sizes of motif). The composite measures for the two differently sized motif datasets are shown in Tables A.1 (for the variability data) and A.2 (for the motif count data).

The relative performance of the clustering algorithms using the variability data show little difference in the results (evaluated using the composite measure) with Random Forest giving the best value for the 20 minutes motifs and kmeans giving the highest value for the 30 minute motifs. However, the small differences between the composite measure results shows that each algorithm demonstrates a similar level of performance.

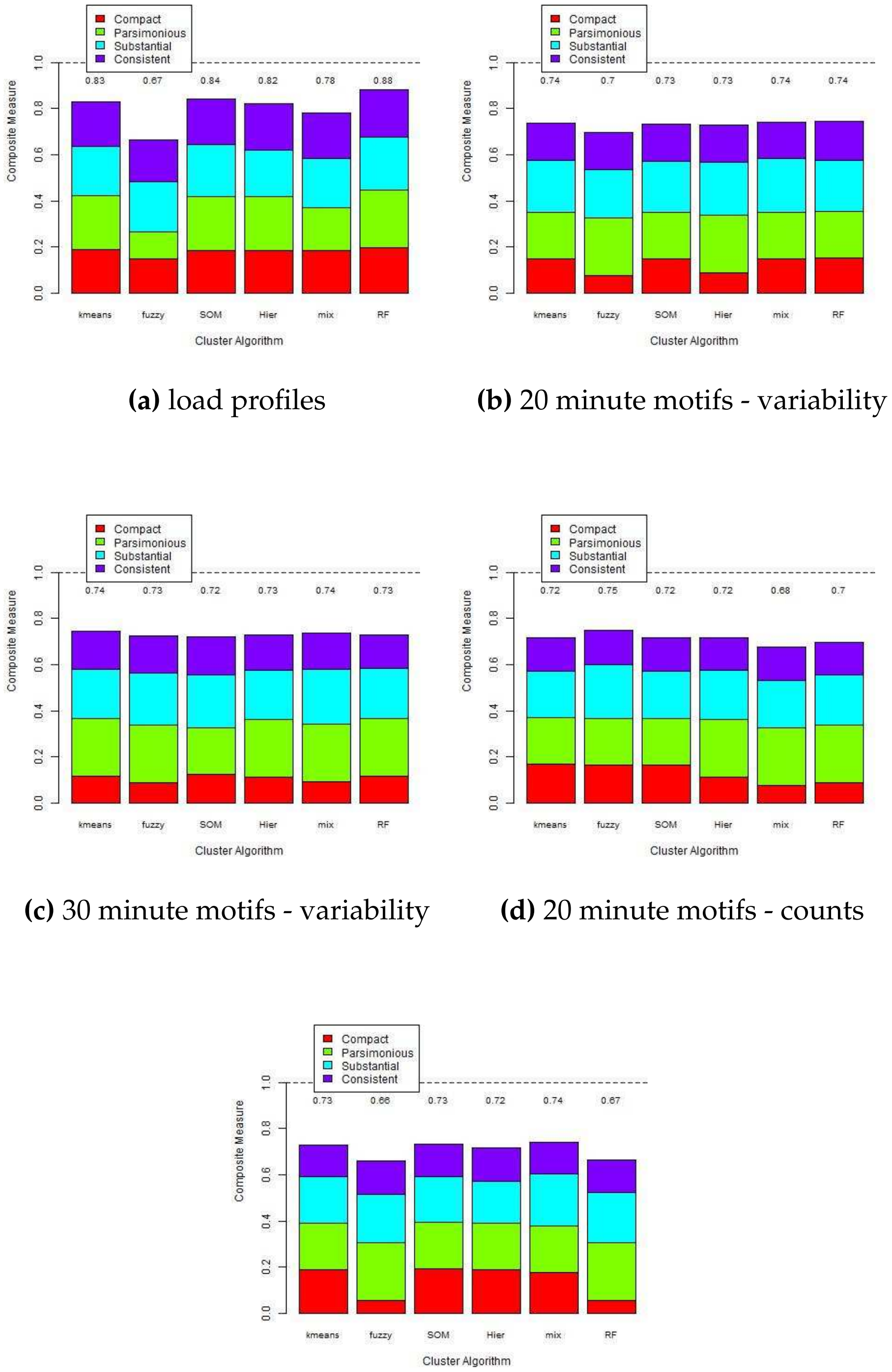
The cluster centres and the households allocated to each cluster using the random forest algorithm are shown at Figures 6.3 and 6.4. For comparison purposes the kmeans results are also shown at Figures A.1 and A.2.

The results from the kmeans algorithm using the variability data can be seen at Figures 6.5 and A.3. These graphs show the same information as in Figures A.1 and A.2.

From the 20 minute motif variability results using kmeans (Figure 6.5) it can be seen that the blue cluster (7 houses) shows relatively little variability in the timing of their regular activities and can be assumed to be "creature of habit" households which may not respond well to an incentive to change behaviour. The green (70 houses) cluster show a high variability in the timing of the motifs found and may be best to target for interventions.

Examples of households, their daily load profiles and the most frequent motif found are shown at Figure 6.6 for sample households from the blue cluster and at Figure 6.7 for samples from the green cluster. It is clear from the low variability examples that the timing of the motifs is grouped around a mean time in the morning and then again around a mean time in the afternoon. In comparison, the high variability examples show motifs occurring throughout the day with very little apparent arrangement around an average morning or afternoon time.

Considering the 30 minute motif results (Figure A.3), the blue (24 houses) cluster shows relatively low variability in timing of the motifs and example households from this cluster are shown at Figure A.4. The red



(a) load profiles

(b) 20 minute motifs - variability

(c) 30 minute motifs - variability

(d) 20 minute motifs - counts

(e) 30 minute motifs - counts

Figure 6.2: Composite measure for each clustering algorithm

cluster (48 houses) shows relatively higher variability and example households from this cluster are shown at Figure A.5.

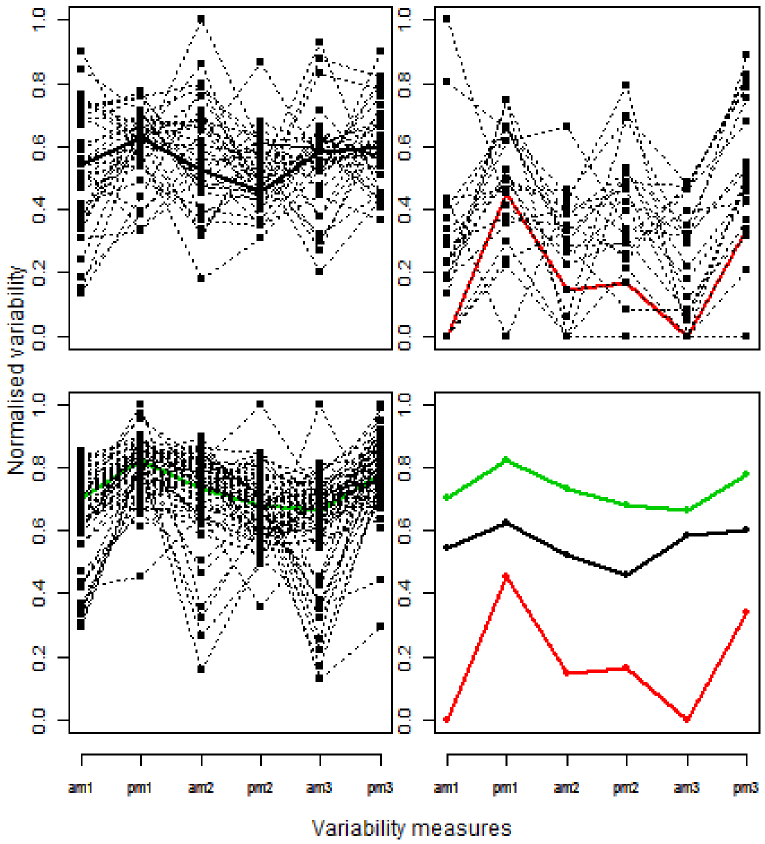


Figure 6.3: Random Forest clusters and members of each cluster using 20 minute motif

The results shown in Figure 6.2 demonstrate that there is very little difference in the composite measures for the 20 minute and 30 minute motif size datasets. Using either produces similar quality results across both the variability dataset and the motif count dataset.

The results show three separate cluster partitions for the same set of households arising from using three different sets of data about the households (the load profile data, the motif variability data and the motif count data). Therefore, each household is assigned to an archetype within each of the three datasets and, by examining the appropriate archetype, conclusions can be reached about the household such as, e.g.:

- The load profile information shows that the household uses most electricity during the early morning period.
- The motif variability information shows the household has high

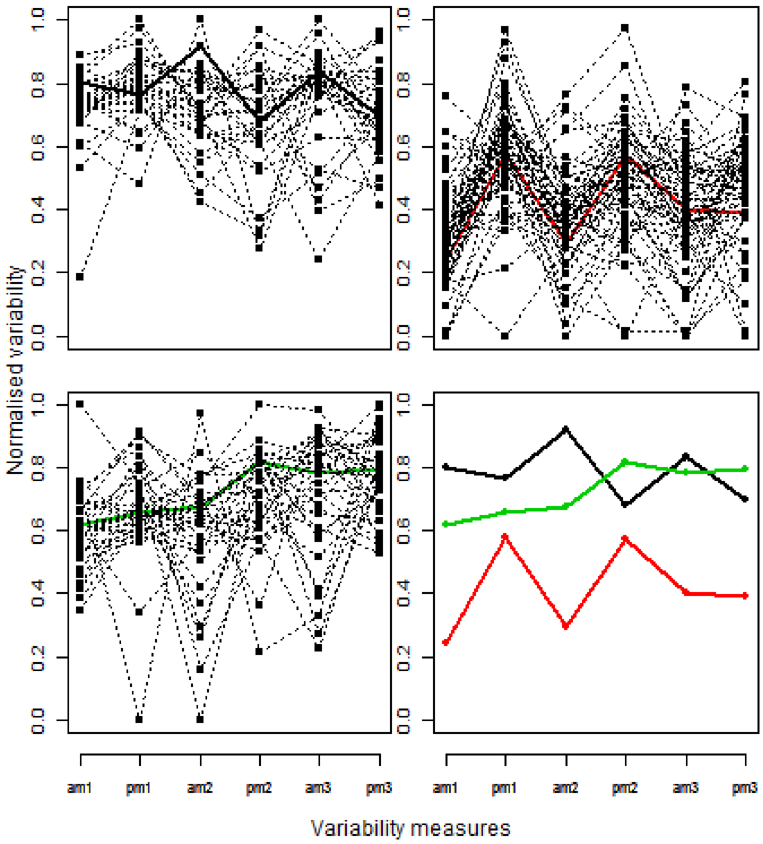


Figure 6.4: Random Forest clusters and members of each cluster using 30 minute motif

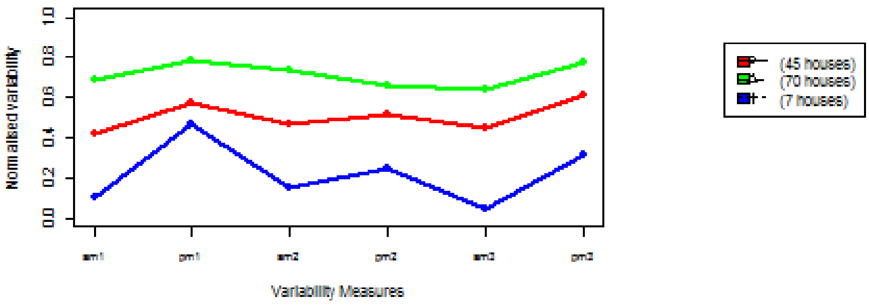
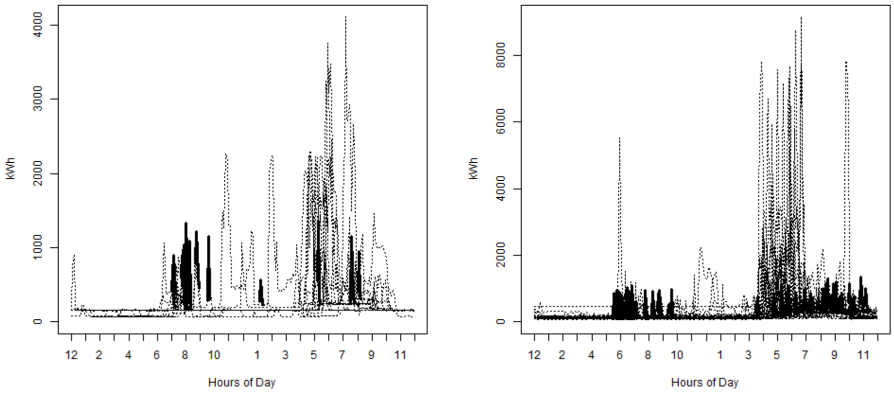


Figure 6.5: Kmeans clusters using motif variability with 20 minute motif

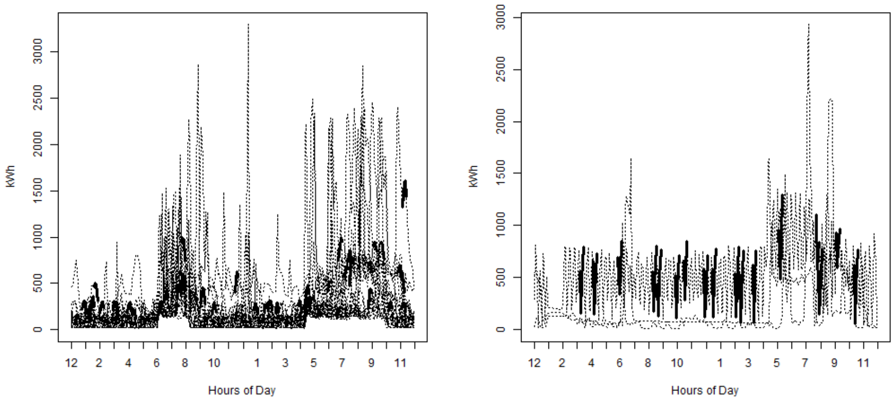
variability in their behaviour.

- The motif count information shows the household has a relatively large number of motifs which may suggest they repeat their behaviour frequently.



(a) Household 217 **(b)** Household 22

Figure 6.6: Examples of households from the low variability cluster (20 minute motif)



(a) Household 305 **(b)** Household 7

Figure 6.7: Examples of households from the high variability cluster (20 minute motif)

Based on this information an appropriate intervention targeting the required type of households can be designed.

The three sets of results are easily understood by people designing and implementing DSM interventions who can target households falling in one, two or three of the archetypes detailed above depending on the type of intervention and the behaviour change desired. For example, a focus on the shape of the usage pattern (e.g. to move usage from the evening peak period) may target those households in the cluster represented by an archetype with a large evening usage. Within this set of households, a subset can be targeted based on their degree of variability in behaviour with highly variable households targeted with a particular incentive to move their activities to a different, non-peak time.

As well as considering the three aspects of the household (as detailed above), combining the results from the clustering using the three different datasets will allow a single set of archetypes to be defined. One method to combine the results is to use ensemble clustering methods to take the cluster partitions from the application of a clustering algorithm to each dataset and use these partitions as input to a further clustering exercise. This further clustering groups households that are commonly grouped in the same partitions in the underlying results into the same partition in the combined results.

The clustering results are combined across the datasets for each of the six clustering algorithms - i.e. the results for each dataset using the kmeans algorithm are combined into an overall set of partitions. Similarly, the results for the other algorithms are also combined. This leads to the composite cluster validity results shown in Figures 6.8 and 6.9 with the data shown in Table 6.4.

Table 6.4: Ensemble clustering across datasets

20 minute Motifs						
	Cluster Sizes	Overall	Compact	Pars'ous	Subst'al	Consistent
Kmeans	1,1,2,4,8,47,59	0.544	0.643	0.006	0.934	0.591
Fuzzy Cmeans	1,3,5,24,39,50	0.543	0.62	0.058	0.887	0.607
SOM	1,2,4,11,47,57	0.474	0.491	0.058	0.786	0.56
Hierarchical	1,1,2,3,4,31,35,45	0.494	0.402	0.058	0.919	0.598
Model	1,1,1,2,3,18,40,56	0.506	0.514	0.058	0.842	0.61
RF	1,1,1,3,4,26,42,44	0.526	0.502	0.058	0.921	0.623
30 minute Motifs						
	Cluster Sizes	Overall	Compact	Pars'ous	Subst'al	Consistent
Kmeans	1,1,2,3,5,10,38,62	0.45	0.414	0.058	0.776	0.552
Fuzzy Cmeans	1,2,2,35,35,47	0.53	0.54	0.058	0.928	0.593
SOM	1,1,4,5,11,41,59	0.48	0.493	0.058	0.802	0.567
Hierarchical	1,1,3,4,14,36,63	0.461	0.463	0.058	0.78	0.543
Model	1,1,2,5,17,44,52	0.51	0.548	0.058	0.843	0.592
RF	2,2,3,4,14,15,35,47	0.506	0.415	0.213	0.783	0.615

It is clear that combining households grouped into three clusters (using the motif variability and count datasets) with households grouped into

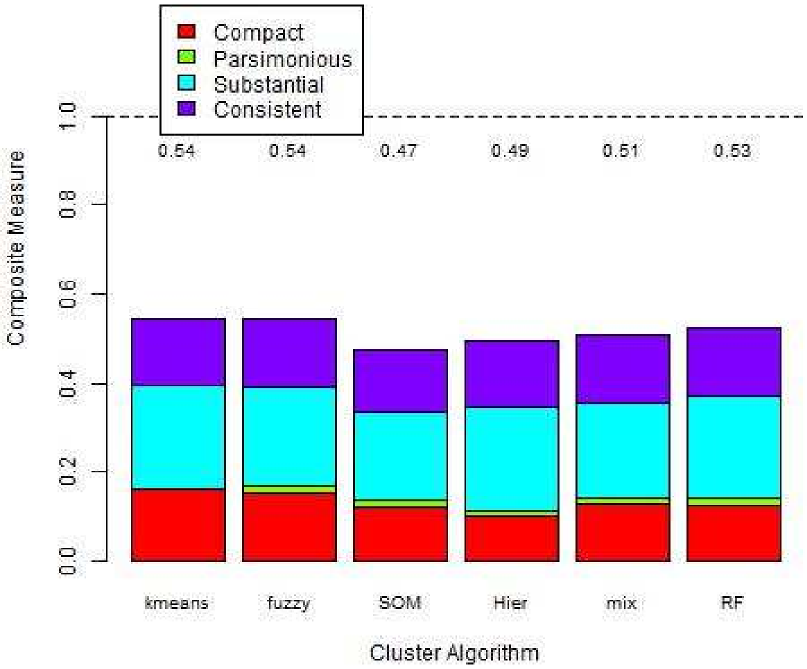


Figure 6.8: Ensemble results across datasets using 20 minute motif data

eight clusters (using the profile dataset) produces results which have few adequately sized clusters. Calculating the composite measures produces very low values for the parsimonious measure as the number of reasonably sized clusters is far away from the optimum of 8. In fact, for each algorithm, only three reasonably sized (four for random forests using 30 minute motifs) clusters are found. Thus the composite measure is very sensitive to an appropriate setting for the optimum parameter. As discussed below, the combination of the results for each dataset is hard to interpret and, for that reason, and for the problems in finding sufficient reasonably sized clusters, is not the preferred method of using ensemble clustering.

For the 20 minute motif, the best performing algorithm is the fuzzy cmeans algorithm. This is surprising as the fuzzy cmeans algorithm performs poorly when clustering the individual datasets but well when using an ensemble approach to combine the individual clustering results. The individual results for the fuzzy cmeans algorithm using the profile

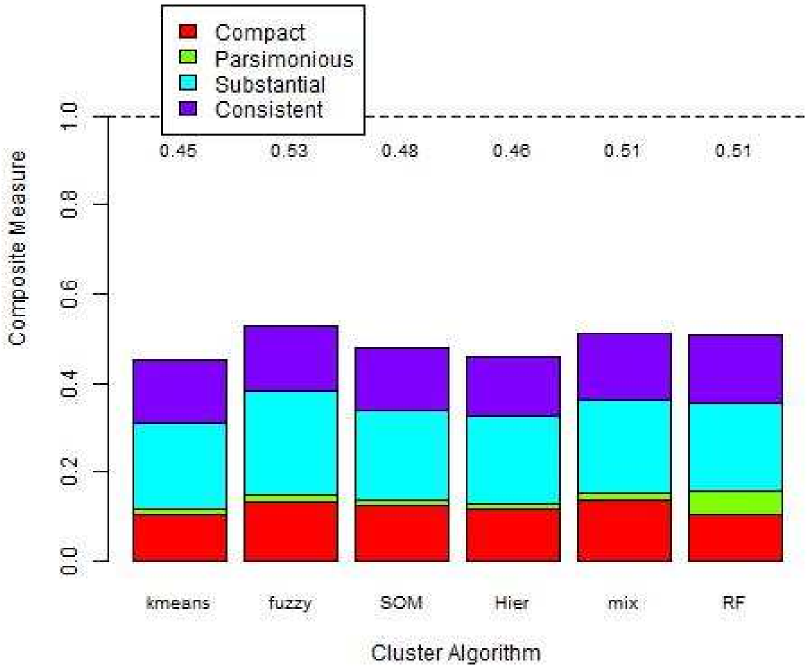


Figure 6.9: Ensemble results across datasets using 30 minute motif data

data has less final clusters than for the other algorithms and this is a possible reason for the good results when applying ensemble techniques.

The results from the ensemble clustering across the different datasets are hard to interpret for a professional developing a DSM intervention. While the individual datasets can be understood (i.e. a particular shape of load profile, low, medium or high degrees of variability, etc.), the combined results merge the easily understood characteristics to produce one set of clusters.

A more meaningful application of ensemble clustering is to combine the results for each dataset across each clustering algorithm. This maintains the easily understood interpretation of the results but provides a set of partitions that make use of the results from each clustering algorithm.

All the results (for each algorithm and for each dataset) can be combined into a single set of partitions although this, again, is hard to interpret in a meaningful way. The results for the overall comparison of all algorithms and all datasets are shown alongside those for each dataset in Figures 6.10 and 6.11 and in Table 6.5.

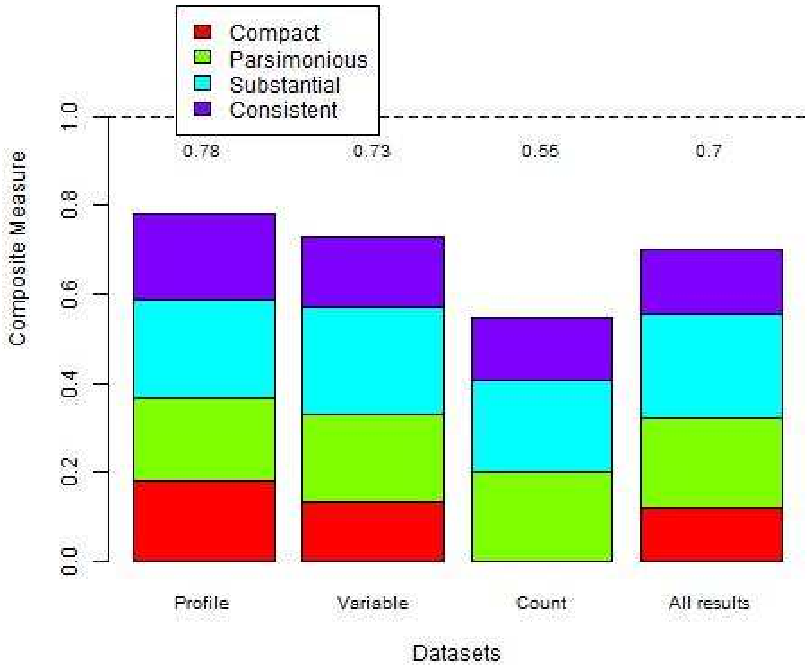


Figure 6.10: Ensemble results across datasets using 20 minute motif data

Table 6.5: Ensemble clustering across algorithms and overall

20 minute Motifs						
	Cluster Sizes	Overall	Compact	Pars'ous	Subst'al	Consistent
Profile	4,6,9,12,21,21,22,27	0.782	0.73	0.74	0.88	0.778
Variable	12,49,61	0.729	0.525	0.805	0.951	0.633
Count	20,30,72	0.548	0	0.805	0.828	0.558
All results	1,14,49	0.657	0.312	0.805	0.939	0.571
30 minute Motifs						
	Cluster Sizes	Overall	Compact	Pars'ous	Subs'al	Consistent
Profile	4,6,9,12,21,21,22,27	0.782	0.73	0.74	0.88	0.778
Variable	27,47,48	0.744	0.44	1	0.888	0.65
Count	4,34,84	0.729	0.765	0.805	0.795	0.551
All results	15,39,68	0.615	0.201	0.805	0.881	0.572

If the results for the ensemble clustering are compared with the results

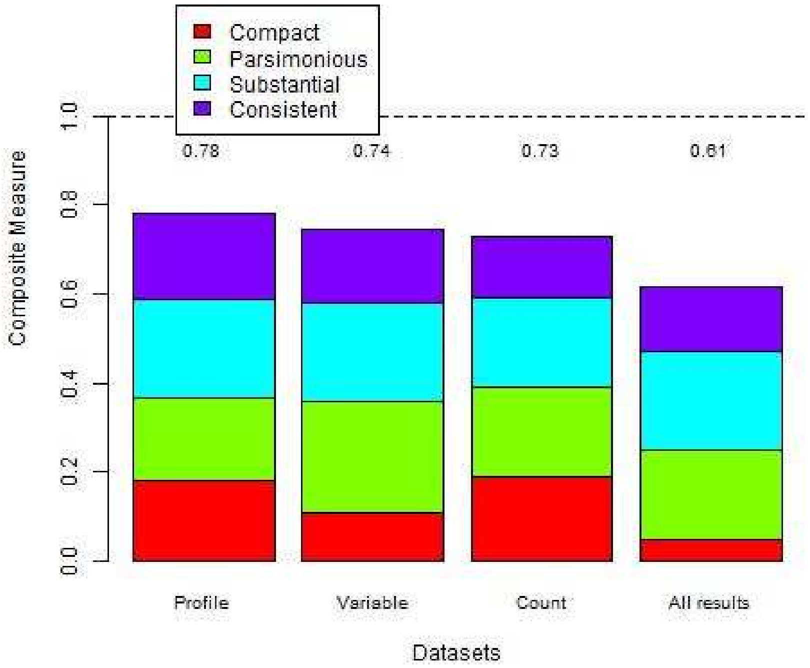


Figure 6.11: Ensemble results across datasets using 30 minute motif data

for each of the clustering algorithms (Figure 6.2) it can be seen that the composite measure for the motif variability dataset is similar to that for the individual clustering algorithms (around 0.73) for both the 20 minute and 30 minute motif data.

Combining the results for each algorithm can provide some degree of "averaging" and may provide more useful results by reducing the impact of peculiarities of a particular algorithm. The cluster centres and the members of each cluster for each of the three datasets after applying the ensemble clustering are shown in Figures 6.12, 6.13 and 6.14.

As a reminder of what the graphs show, Figure 6.12 shows the average meter reading data for each household arranged by cluster to which the household has been assigned. For each household the data is averaged across all the days of readings for that household and is then normalised for the readings to fit within the 0-1 range. Similar shaped average household profiles are then grouped together. Note that one of the implications of this process is that similar shaped usage patterns across the days are grouped together irrespective of the total usage. For instance a household that averages a daily usage of 10kWh will be grouped with another household that averages a daily usage of 1kWh if the shape of that usage is similar across the day.

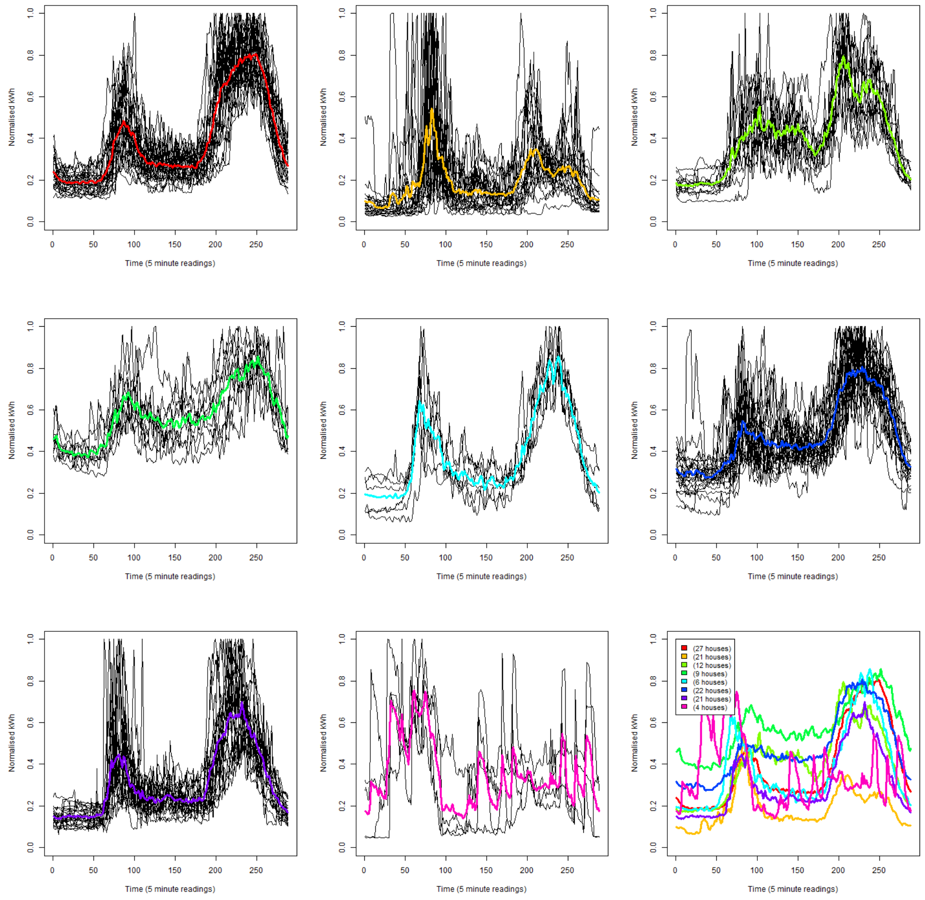


Figure 6.12: Ensemble clusters for all algorithms using load profile data

Figure 6.13 shows the allocation of households to each cluster based on motif variability. Each motif found within the data has been examined for its time of occurrence. The times are then split into morning (am) and afternoon (pm). The three most frequently occurring motifs for each household are then examined and the variation in timing of each of the popular motifs is found (represented by the standard deviation of the times around the morning and afternoon means). Note that the most frequently occurring motif for one household may be a different motif from that of another household. The values of standard deviation for each motif are then normalised (linearly) to fall into a 0-1 range. The households are then grouped using the similarity of variability in timing of the motifs resulting in three partitions that can be seen as high, medium and low degrees of variability in behaviour.

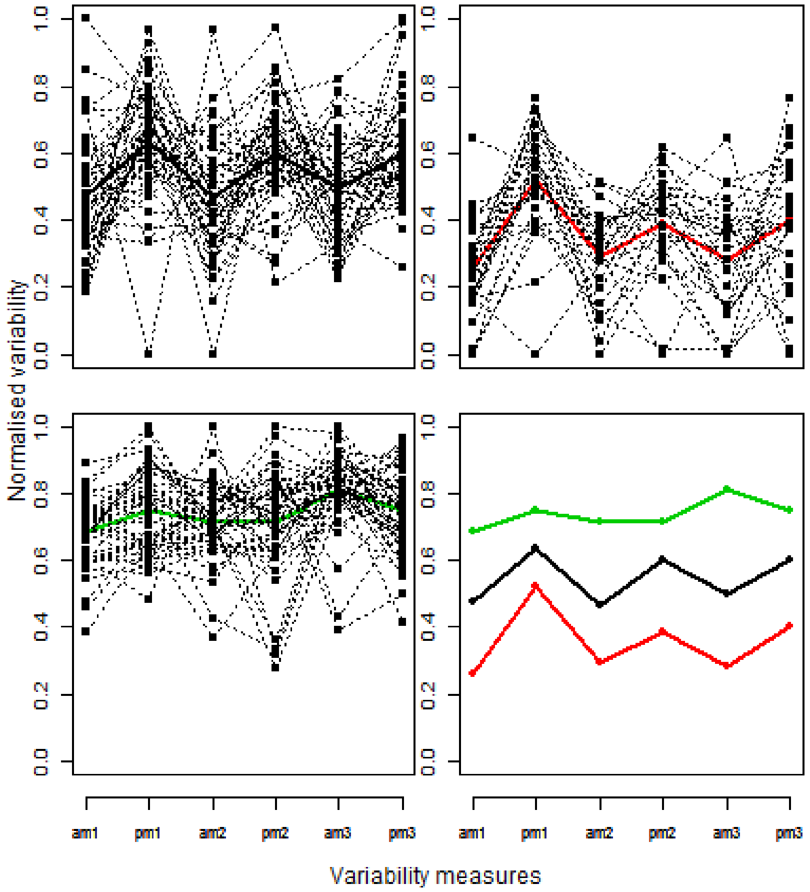


Figure 6.13: Clusters and members across all algorithms - 30 minute motifs, variability data

Figure 6.14 shows the allocation of households to each cluster based on the number of motifs occurring in each household's meter data. The three most frequently occurring motifs for each household are examined and the total number of occurrences of each motif within the household are calculated. In addition, the total number of different motifs that occur at least twice within the household meter data is calculated. The values of counts for each motif (and the total) are then normalised (linearly) to fall into a 0-1 range. The households are then grouped using the similarity of numbers of motifs resulting in three partitions that can be seen as high, medium and low occurrences of motifs.

A higher value for the count of the third most popular motif compared to the most popular (e.g. for the blue cluster) may be surprising but arises from the normalisation stage of the processing. The underlying count of the third motif will be less than that for the first but, when normalised into the 0-1 range, may result in a higher normalised value.

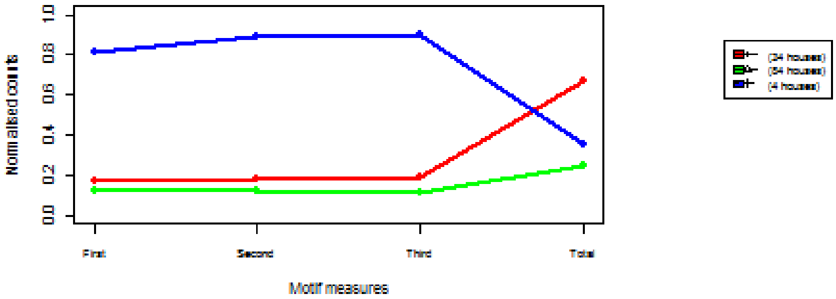


Figure 6.14: Clusters and cluster sizes for motif count data across all algorithms

The motif variability results show that the distribution of variability of timings around the mean tends towards lower values for the morning data when compared with the afternoon data. This matches expectations as a lot of households will be inactive during the night with a peak of activity around breakfast time before several of the household members leave for school or work. This will mean that a lot of the activities will be concentrated around the mean time for breakfast and hence the variability in timing will tend to be relatively low in the morning.

In comparison, activities can occur throughout the afternoon and evening and, while there will be some concentration around the evening meal time, it would be expected that regular activities may be spread a lot wider across the whole afternoon period. The data matches expectations with higher variability for the afternoon compared to the morning period.

6.4 Comparison with Simple Measures

The question may be posed as to whether the clusters resulting from the derived information (such as the timings of motifs) are better than taking the very simple approach of splitting households into a few groups based on simple measures.

For example, the simple measure of the variation in total electricity usage from day to day can be used to split households into 3 nearly equal groups (41, 40, and 41 households) to give three groups which can be labelled as households with "low", "medium" and "high" degrees of variability. Are the partitions arising from the motif variability analysis any better than this simple approach to splitting the households?

It can be argued that the approach of splitting households based on a coarse measure can certainly produce three partitions but gives little insight into the variation in behaviour within the total electricity usage of the household. A household that uses different amounts of electricity from day to day may just represent one in which the members of the household are sometimes out of the house all day (little electricity usage) and sometimes in the household (higher electricity usage) giving a relatively high degree of variability. However, the behaviour of the household members when they are in the house (i.e. the times when they do certain activities) may not vary from day to day. Hence the two approaches are providing different insights into the data.

Consider the examples shown in Figure 6.7 which shows households 7 and 305 which fall into the high variability cluster as calculated using the motif variability data. The normalised variability of the total usage falls into the 0-1 range with cut points for 41 households and 81 households falling at 0.058 and 0.136. The value for household 7 is 0.201 putting 7 into the "high" variability simple grouping. The value for household 305 is 0.060 placing the household into the "medium" grouping.

Comparing the groupings from the 20 minute motif variability clustering (using kmeans) with the grouping using the simple measure above, the partition memberships can be compared using the Rand measure. This produces a figure of 0.513 suggesting that while there is some overlap between the grouping using each approach, the amount of overlap isn't extensive.

The conclusion is that, while the simple measures of variability may give some insight into the households' behaviour, the insight doesn't match that arising from the motif variability analysis.

6.5 Comparison with Socio-demographics

The NESEMP project includes the collection of socio-demographic and attitudinal information from the participants in the trial through the use of questionnaires which have been analysed from a social psychology point of view. This information has been excluded from the analysis reported in this thesis but is available for comparison of the final results with the "real world" socio-demographic information collected as part of the study.

Note that there is no absolute information available on each household's degree of variability of behaviour and it is only possible to explore how the clusters arising from the analysis of the meter data match against questionnaire answers that may indicate attitudes and life styles that suggest variable behaviour. There is, therefore, no "ground truth" that can be used to validate the analysis results.

Details of the questionnaire questions and the possible responses are included in Appendix B.

For each question, the clustering results from the ensemble step that combines all the clustering algorithms (for each dataset) is compared with the overall split of answers to the question. The split of answers within the members of each cluster can be compared with the overall split to identify any discrepancies which may indicate a mapping between a partition and a particular response to the questionnaire.

The comparison uses the data from the 30 minute motif variability and count analysis for comparison with the questionnaire responses. The second cluster for the variability and motif count data is the low variability cluster and the third is the high variability cluster. For the load profile information, the clusters are numbered in the same order as on Figure 6.12.

6.5.1 Questionnaire responses

To explore the matching of the partitions arising from the clustering using the data with the questionnaire responses, each of the questions has been considered and the responses mapped onto the partitions. If a particular partition shows a large difference from the overall spread of responses then it suggests that further investigation of that cluster is merited. It may be possible to match a partition to a particular demographic or attitudinal archetype although the results are not conclusive.

Appendix B shows the results of the analysis. Some correlation between the clustering results and the Size of House, the Number of Bedrooms, and the Type of Household can be seen although the results are not compelling. Little matching between the answers to the questionnaire questions and the derived archetypes can be found for the attitudinal questions.

6.6 Discussion

Various measures that represent the variability of behaviour can be constructed and this chapter considers two based on the underlying meter readings (the variability in time of maximum usage and the variability in total usage). However, as each measure is intended to represent the same thing (i.e. the variability of behaviour), the fact that there is little correlation between the measures means that at least one provides a poor representation of the characteristic.

When clustering using just the load profile information, the Fuzzy cmeans algorithm shows the least useful results with higher values for the MIA and CDI as well as less useful sizes of clusters including a partition of one household.

The data used for the variability of motifs analysis are not ordered in that the differing attributes are not directly related to each other in the same way as within the load profile information. For example, the attribute for a particular time (e.g. 5pm) always falls between 4:55pm and 5:05pm and treating the attribute in isolation is not likely to be meaningful. However, a selection of the variability attributes can be included in the clustering exercise (e.g. the information for the third most popular motif could be omitted). This provides much more flexibility in the attributes to define, and to include, and allows for the application of tools such as Principal Component Analysis to determine the "most important" attributes affecting the quality of the clustering.

While the compactness of each cluster and the separation between differing clusters is generally used to define various cluster validity indexes to measure "quality" of the cluster, other criteria used within marketing campaigns (e.g. the size of the clusters, the intuitive understanding of the clusters to the layman) also influences the effectiveness of the solution found.

Using the motif timing variability data to create archetypes based on the timing data complements the archetypes created using the load profile data. Similarly, using the motif count data also provides additional information on the households.

Combining the data to create a single set of archetypes (that incorporates the meter data, the motif variability data and the motif count data) does not produce useful archetypes as evaluated using the cluster validity indexes or by considering the simple meaning of the archetypes. The combination of all the data generated into a single set of archetypes does not produce better results than using just the load profile information.

However, treating each dataset as providing different information about the household and then applying clustering algorithms to each dataset separately produces 3 collections of archetypes where each household is a member of one of the clusters in each collection. This provides additional information on the household in that different aspects of the household behaviour are provided by each archetype (e.g. a particular load shape, a degree of variability of behaviour, and a degree of amount of repeating behaviour). These combinations of archetypes can usefully be used to create DSM interventions. Hence, the use of motif information, and the variability in timing of the motifs, is a useful addition to knowledge about the households.

6.7 Summary

The symbolisation technique is shown to be effective in being able to detect repeating patterns (motifs) that are approximately the same shape. Depending on the type of intervention planned for a subset of the households (e.g. incentives to change overall electricity usage from day to night, or to influence short periods of usage during the peak period), differing sizes of motifs may be used.

The approach of clustering using the variability of motifs representing regular activity patterns within households extends and complements the widely used approach of clustering using average load profiles. It is found that creating archetypes based on the motif information complements the archetypes arising from load profile clustering and provides a more complete understanding of the household characteristics than that resulting from using just the load profile results.

Combining all the results from each of the three datasets into a single set of archetypes does not produce useful results (as measured using the cluster validity indexes) and is not easily interpretable. The conclusion is that, while simply adding data generated using the motif information to the household meter readings does not allow the creation of more useful archetypes, treating the motif data separately to produce 3 sets of archetypes does provide a more complete understanding of the households.

The ability to cost effectively partition domestic households into a few meaningful archetypes, based on the household electricity usage, is an important problem for the electricity industry. Identifying a few archetypical representations of households is essential to cost effective implementation of demand side management techniques which itself is necessary to allow the electricity industry to meet the upcoming challenges. Producing more consistent and more descriptive archetypes than currently possible will allow the deployment of more effective behaviour modification interventions.

CHAPTER 7

Conclusions

This thesis has investigated the information about household behaviour that can be derived from using household electricity meter data. The ability to derive knowledge of households' behaviour on a large scale using widely deployed technology (e.g. smart meters) is essential for effective deployment of DSM techniques to address the current and forthcoming challenges for the UK electricity industry.

The work has taken the approach of discovering a few archetypical households which can be used as representatives of the groups which each contain large numbers of similar households.

Chapter 4 implements an approach that has been widely used in the academic literature. Using a database of electricity meter data collected in Scotland at a frequency of five minutes, households are clustered based on the shape of the average daily electricity usage. The meter readings from 123 households for working days from the Spring 2011 period are considered for the main analysis with a comparison to Spring weekend days and Summer working and weekend days also included. This work involves the application of existing techniques to data collected in a region of the UK. The UK has specific geographic and historical differences from other countries and thus results arising from work in other countries may not be directly applicable to the UK. The data used is from a area of northern Scotland and differing climatic conditions and daylight hours may impact on the applicability of the results to other areas of the UK.

To assess the results of the load profile clustering, an extended composite measure of the quality of the partitions resulting from the clustering analysis has been defined. Previous published clustering work on electricity meter data has used various cluster validity indexes but all focus on the compactness of the households within a cluster and the separation of clusters from each other. This thesis proposes that additional criteria used in the field of marketing segmentation are included in the composite measure of the quality of the results to give a more complete representation of the usability of the results.

Using this composite measure, commonly used clustering algorithms have been assessed and ranked for their applicability to the load profile clustering application domain. The rankings of the algorithms are found to be different from those that arise from using the traditional CVIs that focus purely on compactness and separation. This work provides guidance on the ranking, as assessed using traditional validity indexes alongside the proposed composite validity index, of clustering algorithms for researchers in the electricity application area to use for load profile clustering.

Chapter 5 investigates how a symbolisation approach may be used to find patterns within a household's stream of electricity meter data that are approximately similar and which may represent regular activities within the household. Motif detection within household electricity meter data has not previously been researched in detail.

Many different approaches to finding motifs are possible and an evaluation method is defined to determine which approaches give effective results. Defining DSM initiatives to influence household behaviour relies on finding an appropriate number of motifs that represent regular behaviour: not too many nor too few. The differing parameter settings for motif finding are evaluated and a set of parameters determined that produce a reasonable number of motifs for further analysis.

The options considered include the use of different lengths of motifs (20 minutes, 30 minutes, 45 minutes and 1 hour), differently sized symbolisation alphabets, using differences between data points or the absolute data, and using the full motif or compressed versions. An important investigation is for the method of banding the motifs such that they are matched on similar shape but only if the range of the motif (i.e. the difference between the highest and lowest meter readings during the duration of the motif) falls within a similar range. In this way, appliances that produce similar shaped motifs when in use can be differentiated if the electricity usage for each is different (i.e. the shape of usage is similar but the total amount of electricity usage is very different).

Automatic methods of discarding motifs that are judged to be of little interest are defined and used in the analysis. Applying these methods reduces the motifs to be considered and excludes periods of electricity usage which may be similar but of little use for DSM initiatives (e.g. periods of no electricity usage when the house is empty).

The analysis results in a set of parameters and methods that provide a reasonable number of useful motifs for understanding household behaviour. The guidelines found by the analysis in Chapter 5 are then applied in Chapter 6.

Chapter 6 introduces the concept of variability of behaviour and defines how it may be represented by variability from day to day in the timing of occurrence of a particular motif within a household.

The variability of timing of motifs and the number of motifs found within each household's meter data are used to create two different approaches to clustering the households that complement the load profile clustering described above. This results in three separate datasets (profile data, variability of motif data, and motif count data) which each represent aspects of the households.

Combining the partitions from the three datasets using ensemble clustering shows poorer performance than the clustering done using just the load profile data and also produces archetypes that are hard to interpret. The alternate approach of considering the clustering results from all three datasets separately is preferred and, for each household, provides membership of three archetypes which can be understood as relating to the shape of the usage throughout the day, the degree of variability in regular behaviours (low, medium or high) and the number of repeating activities per day (low, medium or high). The partitions resulting from clustering of the motif variability data are of comparable quality (as measured by the composite measure) as the partitions arising from the load profile data only. Viewing the households as members of three archetypes based on different criteria provides greater insight into the households than using just the load profile data.

The partitions from each of the six clustering algorithms used is combined using ensemble clustering techniques to produce three sets of results (one for each of the three datasets).

"Real world" data arising from questionnaires completed by the household members, as part of the NESEMP project, is used to determine whether any of the archetypes arising from the clustering can be aligned with a particular type of household or attitude. As there are no questionnaire results that give a direct measure of variability of behaviour in the household, it is not possible to validate the clustering results against a "ground truth". However, some relationship between the archetypes and information on household type (e.g. single, family) and number of people in the household can be seen. No direct relationship between the archetypes and the responses to attitudinal questions can be found.

7.1 Key Results

The research hypotheses for this thesis were laid out in Chapter 1 and are repeated here.

The overall thesis is "Can variability in behaviour within a household be identified by finding motifs within UK electricity meter data and can these motifs then be used for clustering households into a few archetypes?"

The results shown in Chapter 6 show that motifs can be detected within the electricity meter data and that these can be used as the basis for clustering households into a few archetypes that can be characterised as relatively "low", "medium" or "highly" variable households. The archetypes can be used alongside the archetypes arising from the load profile clustering (as shown in Chapter 4) to provide a more complete understanding of the household characteristics.

The thesis raises a number of questions that the research addresses.

7.1.1 Can clusters of households be found using UK electricity meter data?

This question leads to a number of sub-questions:

- Can a method of evaluating load profile clustering be defined that incorporates aspects of the effectiveness of the results for use in DSM?
- Does the sampling frequency affect the quality of the archetypes found?
- Which clustering algorithm provides the best clustering results, as defined by the evaluation method and using data collected in the UK?

The work documented in Chapter 4 describes the use of various clustering algorithms applied to household data to determine a small number of archetypical households. In this work, a target of eight archetypes is used and this allows for well distinguishable load profiles to be identified by each of the clustering algorithms used.

Defining a method of evaluating the clustering results

Assessing the results using CVIs used in other electricity load profiling work produces a ranking of how effective each of the clustering algorithms is in producing the archetypes. However, to provide a more complete evaluation of the results using criteria beyond compactness and separation, a composite measure is defined and also used to rank the clustering algorithms.

Based on guidelines from literature in the field of marketing, four sub-measures have been defined which reflect criteria for a "good" evaluation of a segmentation result. These sub-measures have been combined into a composite measure which is then used to assess the effectiveness of the clustering results to inform a DSM program intended to change household behaviour.

Using the composite measure, the clustering algorithms are ranked differently than when using the well used CVIs and the composite measure can be used to evaluate different clustering algorithms using the additional criteria included in the composite measure.

Chapter 4 describes the clustering evaluation method and its application is shown in Chapters 4 and 6. The composite measure provides a more complete assessment of the quality of a given partitioning solution and incorporates an assessment of cluster sizes, number of clusters and consistency of results between comparable time periods.

The work shows that it is possible to create a composite measure to evaluate the clustering results and that the composite measure provides results different from those arising from using the traditional CVIs.

Evaluating the data sampling frequency

Chapter 4 examines this question using datasets collected at five minute and one hour frequencies. The one hour dataset is derived from the five minute dataset by aggregating readings. The results, assessed using both the novel composite measure (see Figure 4.3) and the CDI and MIA validity indexes (see Tables 4.2 and 4.3), show that the quality of the clusters found from the hour data are inferior to those from the five minute data.

Thus it is concluded that the more frequent sampling rate is preferable as assessed using the composite cluster validity index.

Application of load profile clustering to Scottish data

Previous load profile clustering work assessed using the CDI and MIA measures has generally concluded that hierarchical clustering is the most effective algorithm with SOM and kmeans also providing reasonable solutions. Assessing the algorithms using the composite measure shows that random forest clustering provides the best results with kmeans being assessed much lower than with traditional CVIs. The results in Table 4.6 show how the algorithms vary when assessed using the different validity indexes.

The data used is collected from part of northern Scotland (Aberdeenshire) and household behaviours driven by UK wide social conventions (e.g. common television programming) are likely to be similar across all parts of the UK. However, climatic conditions and daylight timings which are different to parts of the UK that are geographically far away (e.g. southern England) may mean that some aspects of the results are not generally applicable across the whole of the UK.

The work is specific to the UK and, while the approach would be applicable to data collected in other countries, the results on the ranking of the clustering algorithms may be different due to differences in electricity usage practice between the UK and the countries providing the data.

The work detailed in Chapter 4 shows that load profile clustering can be applied to the data used in this study.

7.1.2 Is it possible to find a reasonable number of interesting motifs within the electricity meter data?

A sub question arising from this work is "Whether it is possible to define a set of parameters to use for the motif finding technique to provide a reasonable number of motifs". "Reasonable" is defined as being not too many for the motif to represent behaviour that occurs too frequently to be influenced, nor too few for influencing of the behaviour to be of little interest.

The work establishes a method of how to find motifs and evaluate between different parameters and is detailed in Chapter 5 which includes:

- The definition of a motif finding approach and consideration of the many parameters that can be varied.
- The identification of simple non-interesting motifs and their exclusion from the analysis.
- The definition of a qualitative evaluation method to choose between the differing sets of parameters. Parameter settings are particular

to the problem of motifs within electricity meter data although the approach would be valid for other application domains.

- The identification of the set of parameters to use for the motif finding technique to provide a reasonable number of motifs.

Chapter 5 addresses the question as to whether regular activities within the household can be identified solely using the electricity meter data captured. The analysis shows that data sampled at a five minute interval is sufficient to provide a usable number of motifs which can be assumed to represent repeating activities within the household.

Various motifs that show uninteresting behaviour can be found (e.g. no change in usage over the period of the motif such as when the house is unoccupied) and an automatic method of rejecting uninteresting motifs has been included in the approach.

The results show that motif finding using the symbolisation technique and the five minute sampling period is possible.

Definition of a set of parameters

Various approaches to representing the motif including using compression, different alphabet sizes, and different motif lengths have been explored and evaluated using the inspection method detailed in Chapter 5.

Motifs may match but actually represent similar patterns within very differently sized data. Different approaches to splitting the motifs into various bands of usage over the range of the motif have been explored and a solution based on a split into various appliance ranges (five bands from small to large usage) selected as the best approach.

Parameter settings have been identified and detailed in Section 5.11 thus showing that a set of parameters can be defined.

7.1.3 Does extending the household attributes to include measures of variability alter the archetypical clusters obtained?

This work is detailed in Chapter 6 and builds on the work in previous chapters.

The symbolisation technique is shown to be effective in being able to detect repeating patterns (motifs) that are approximately the same shape. Depending on the type of intervention planned for a subset of the households (e.g. incentives to change overall electricity usage from day to night, or to influence short periods of usage during the peak period), differing sizes of motifs may be used.

The work in Chapter 6 shows that archetypical households can be identified using the motif variability information as well as the motif frequency count information. However, combining these results into a single set of archetypes that combine information on the meter readings, the variability of timing of motifs and the frequency of motif occurrences, produces archetypes that are less useful than the archetypes based on just the load profile data. The combined archetypes cannot be easily interpreted in layman terms (e.g. low variability households) and, as measured using the composite measure, show a lower effectiveness. It is concluded that combining the motif data with the load profile data does not produce more effective archetypes.

However, using the archetypes produced from the variability of motifs data alongside the archetypes from the load profile data and the motif count data provides a deeper understanding of the households as each is described by a combination of three archetypes reflecting different aspects of the household. The archetypes can be well understood by the layman as representing a particular shape of daily usage, the degree of variability of behaviour (low, medium or high) and the number of repeating activities (low, medium or high).

It is concluded that using the motif variability information to produce a single set of archetypes is not shown to be effective. However, using the motif variability information to produce an additional set of archetypes to use alongside the load profile archetypes provides an enhanced understanding of the households.

7.2 Future work

During the work detailed in this thesis, various additional areas for investigation have been identified. Some of these arise from extensions to the work detailed here and others arise from ideas from alternative approaches to addressing some of the research areas explored. These areas for further work are described below.

7.2.1 Evaluation using Additional Datasets

The work in this thesis has made use of a large dataset arising from the NESEMP project and which contains data for 100s of households captured over a period exceeding a year at a five minute sampling rate.

The question of frequency of usage monitoring to provide for effective load profile clustering is an important one and this thesis compares sampling at five minute and hourly intervals (see Chapter 4). As the composite measure is higher for the five minute data for each of the clustering algorithms, it is concluded that the more frequent monitoring produces more useful partitions. Further levels of aggregation can be explored using the same base data.

Datasets from other sources could be used as the basis for the same analysis to explore whether the same results are obtained. These could be from other UK sites or from alternative international studies

The work in this thesis has compared two Spring seasons of data to calculate a measure of consistency over time. In addition, the work on creating load profile archetypes and finding motifs has been repeated using data from Spring weekends and Summer working days and weekends with the results compared to those from the Spring working days. No substantial differences have been found in the results between the different periods of data. As the NESEMP project is ongoing, further data could be loaded into the analysis database over time and further alternative seasons can be used for verification of the approach across different periods of the dataset.

Datasets collected from varying numbers of households and at various sampling frequencies are available for analysis and could be made use of in future work. These data sources include:

- Ireland Commission for Energy Regulation (CER) Electricity Customer Behaviour trial which comprises data from 5000 homes and businesses collected during 2009 and 2010 [148]. This data is collected at half hourly frequency.
- Data collected as part of the Electric20 trail within the Horizon Digital Economy Centre at the University of Nottingham. This consists of data collected for 18 households at a sampling frequency of six seconds [149].
- Data collected from 22 households in Loughborough at a sampling frequency of one minute [150].

7.2.2 Test Targeted Intervention

The approach detailed in this thesis has proposed and analysed a number of approaches to determining the most appropriate households to target for DSM interventions to improve the overall electricity network operation. Alternative and complementary approaches to defining a subset of households for particular DSM interventions have been presented.

It is believed that these approaches will allow for more effective interventions. However, the hypothesis that the interventions will be more effective due to the improved targeting has not been tested and a real world deployment of the results is needed to test the underlying hypothesis (i.e. that the defined clusters will allow for more effective DSM interventions). A useful next step would be to undertake a field trial with appropriate interventions and to test the success of deployment of the interventions.

To test this, a targeted subgroup should be subject to a specific DSM intervention that is intended to change household behaviour. By comparing the change in behaviour over time with a separate, randomly selected group of households, a measure of the effectiveness of the targeting can be obtained. This approach is also reliant on the quality of the intervention proposed so a suitable experimental design is needed that will either adjust for the quality of the intervention or use a well tested and known successful intervention.

Interventions could be of many possible types with previous studies exploring the impact of feedback to individual users [100], dynamic pricing to create financial incentives to change behaviour [96], gaming approaches to engagement of household members, and many other possible innovative activities.

Data needs to be collected for a reasonable period for sufficient data points to be available for analysis and multiple data collection periods need to be defined (e.g. to collect a base level, to collect data after a number of test interventions). These requirements mean that a proposed study would likely last for multiple years.

7.2.3 Enhancements to the Composite Measure

The composite measure used in this thesis addresses four of the desired marketing characteristics identified for a good segmentation (see Section 2.4.1). With additional information on the "ground truth" more of the desired criteria could be addressed and extending the composite measure could be addressed in future work.

The results from the compactness measure when using few attributes (as with the motif variability work in Chapter 6) show a problem in comparing clustering results with a random kmeans clustering and an alternative approach to assessing compactness should be investigated. This is particularly of use when dealing with low numbers of attributes and impacts on the ability to compare between two sets of data with very different numbers of attributes.

7.2.4 Considering Different Representations of Variability

The hypothesis explored in Chapter 6 is that producing archetypes based on the variability of regular behaviour would allow for more effective targeting of DSM interventions than those arising from clustering just on average load profiles. Further work on assessing alternative (or additional) measures of variability could be undertaken.

Approaches to be explored in future work include:

- Generating a larger number of possible measures of variability to be included as attributes for the households to be clustered. These attributes could include the statistics suggested in Section 6.2.2. A nearly endless set of possible representations can be defined and a structured method of assessing the effectiveness of each possible measure (or combination of measures) would provide useful information. This information can complement the results from the load profile clustering and from the motif variability work described in this thesis.

- Exploring motifs beyond the three most popular for each household.

- Restricting the households analysed to those with a certain minimum level of motifs within the period of analysis.

Some of the motif variability results show anomalous results with households included in the analysis which have unreasonably low numbers of readings (and hence a low or zero degree of variability as measured by the standard deviation of motif timing). Future work should explore the detection and exclusion of households which show very few motifs. While the household may be suitable for behaviour change interventions using other data (e.g. total usage or overall load profiles) the application of interventions based on the variability of behaviour is not possible without a higher number of motifs being identified.

A two stage process could be defined that firstly identifies the numbers of motifs (as has been done in Chapter 6 and then applying the motif variability processing only to those households that meet certain minimum criteria.

7.2.5 Exploring Additional Usage of the Motif Information

The work in this thesis has concentrated on finding motifs within a household and comparing between households by considering how the timing of the most common motifs within the household vary between households. However, no consideration has been taken of the possible occurrence of the same motif within more than one household and how that may demonstrate how the same activity occurs at different times in different households.

Effective investigation of motifs across households will require testing that the same activity in different households does generate motifs that are sufficiently close in shape as to be identifiable as the same activity.

Other motif finding algorithms can also be incorporated into the proposed approach to better identify the flexibility of behaviour (e.g. Mueen et al. [146]).

The same activity on different occasions within a household may take differing elapsed time periods (e.g. showering) and the motif detection process could be enhanced by being able to detect these differing length motifs and recognising that they represent the same activity. The SAX approach includes some capabilities for time warping and these could be explored in future work.

7.3 Summary

The investigation of household electricity load profiles is an important area of research given the centrality of such patterns in directly addressing the needs of the electricity industry, both now and in the future. This work extends existing load profile work by taking electricity meter data streams and developing new ways of representing the household that can be used as the basis for clustering exercises. The identification of repeating motifs and the investigation of how the timing of these motifs varies from day to day, as a key behavioural trait of the household, is a novel area of research. An improvement in creating useful archetypes that can be effectively addressed by DSM initiatives can have major financial and environmental benefits.

A novel approach to defining a composite measure that complements existing cluster validity indexes is included in the thesis. While all of the sub-measures comprising the composite measure make use of simple or well known approaches, the combination into a single measure, comparable between different electricity load profile clustering algorithms, and across datasets, is novel. The results using the composite measure can be used by electricity industry professionals to improve their analysis of load profiling clustering to develop effective clustering results to drive their implementation of DSM techniques and hence lead to improvements in the electricity network efficiency.

Advice is provided on the most effective clustering algorithms to use for load profile clustering, based on the Scottish dataset and the evaluation measures.

Symbolisation techniques for motif detection have been used in various application areas but rarely within the domain of electricity meter data. This thesis has applied the symbolisation techniques and includes novel work to allow selection between many different combinations of parameters and methods to find the most effective for finding reasonable numbers of motifs. A set of guidelines for other researchers exploring motifs within electricity meter data are produced.

The concept of variability of behaviour as a characteristic of a household that can be used to cluster similar households together has not previously been considered. An novel approach to incorporating this concept, based on the occurrence of motifs, is described and tested.

This thesis has introduced the issues affecting the UK electricity industry, as well as the opportunities to make use of the newly available stream of data from smart meters. Domestic usage of electricity in the UK is 30% of the total usage and is thus an important area of research. Industry commentators are agreed that the implementation of incentives that change domestic behaviour patterns is a key part of the solution needed to meet the challenges. To achieve better targeting of the incentives a better knowledge of households and their behaviour is needed. This thesis provides approaches, results and guidelines that allow this improved targeting to be achieved.

7.4 Dissemination

In this section a list of refereed publications, produced as part of the work on this thesis, are presented.

Table 7.1: Refereed Publication List - Conference and Journal Papers

Authors	Title	Conference	Year
Ian Dent, Tony Craig, Uwe Aickelin, Tom Rodden	Variability of Behaviour in Electricity Load Profile Clustering; Who Does Things at the Same Time Each Day? [151]	Industrial Conference on Data Mining (ICDM)	2014
Tony Craig, Gary Polhill, Ian Dent, Carlos Galan-Diaz, Simon Heslop	The North East Scotland Energy Monitoring Project: Exploring relationships between household occupants and energy usage [15]	Energy and Buildings journal	2014
Ian Dent, Tony Craig, Uwe Aickelin, Tom Rodden	Finding the creatures of habit; Clustering households based on their flexibility in using electricity [147]	Digital Futures 2012, Aberdeen, UK	2012

Continued on next page

Table 7.1 – continued from previous page

Authors	Title	Conference	Year
Ian Dent, Tony Craig, Uwe Aickelin, Tom Rodden	An Approach for Assessing Clustering of Households by Electricity Usage [152]	UKCI 2012, the 12th Annual Workshop on Computational Intelligence, Heriot-Watt University	2012
Aristides Kiprakis, Ian Dent, Sasa Djokic, Stephen McLaughlin	Multi-scale Dynamic Modeling to Maximize Demand Side Management [153]	IEEE Power and Energy Society Innovative Smart Grid Technologies Europe 2011, Manchester, UK	2011
Ian Dent, Christian Wagner, Uwe Aickelin, Tom Rodden	Creating Personalised Energy Plans. From Groups to Individuals using Fuzzy C Means Clustering [85]	Digital Engagement 11, Newcastle, November 2011	2011
Ian Dent, Uwe Aickelin, Tom Rodden	The Application of a Data Mining Framework to Energy Usage Profiling in Domestic Residences using UK data [154]	Buildings Do Not Use Energy, People Do Research Student Conference, Bath, UK	2011
Ian Dent, Uwe Aickelin, Tom Rodden	Application of a clustering framework to UK domestic electricity data [155]	UKCI 2011, the 11th Annual Workshop on Computational Intelligence, Manchester	2011

The papers listed in Table 7.1 contribute to sections of the thesis as follows.

Dent et al. [151] was presented at the Industrial Conference on Data Mining (ICDM) in July 2014 in St. Petersburg and published in the Springer journal “Advances in Data Mining”. The paper covers the approach to using variability in the timing of motifs as the basis for clustering households as detailed in Chapter 6. The paper was awarded "Best Paper" at the conference.

Craig et al. [15] describes the NESEMP project which has contributed the household data used in the studies in thesis. Amongst other topics, it describes the data collection process detailed in Chapter 3.

Dent et al. [155] and Dent et al. [154] apply some of the clustering of load profile techniques described in Chapter 4 to old data collected in Milton Keynes in 1990 to build a few archetypical households. Dent et al. [85] describes a particular benefit of applying the Fuzzy Cmeans algorithm to the same data, to not only produce archetypical users but, to also make use of the fuzziness output to allow personalised offers to be made to customers based on their degree of membership to each archetype.

Kiprakis et al. [153] describes the Desimax project of which the work in this thesis forms part.

Dent et al. [152] and Dent et al. [147] use data as described in Chapter 3 and clusters the households using a simple measure of variability of timing of peak usage within a day with the results then assessed using various cluster validity indexes. This is an early version of the work described in Chapter 6.

Appendices

Appendix A

Additional Results from Variability of Motifs analysis

This appendix contains additional results from Chapter 6.

Table A.1: Cluster sizes and components of composite measure for motif variability data

	Cluster Sizes	Overall	Compact	Pars'ous	Subst'al	Consistent
20 minute Motifs						
Kmeans	7,45,70	0.736	0.596	0.805	0.898	0.644
Fuzzy Cmeans	26,36,60	0.695	0.307	1	0.842	0.634
SOM	7,44,71	0.733	0.601	0.805	0.889	0.638
Hierarchical	34,38,50	0.728	0.348	1	0.923	0.641
Model	7,48,67	0.743	0.605	0.805	0.922	0.638
RF	20,36,66	0.744	0.619	0.805	0.877	0.675
30 minute Motifs						
Kmeans	24,48,50	0.744	0.465	1	0.863	0.648
Fuzzy Cmeans	30,43,49	0.727	0.348	1	0.913	0.645
SOM	18,50,54	0.721	0.498	0.805	0.926	0.657
Hierarchical	23,47,52	0.73	0.456	1	0.855	0.608
Model	35,42,45	0.738	0.371	1	0.954	0.627
RF	31,34,57	0.728	0.475	1	0.866	0.572

Table A.2: Cluster sizes and components of composite measure for motif count data

	Cluster Sizes	Overall	Compact	Pars'ous	Subst'al	Consistent
20 minute Motifs						
Kmeans	7,34,81	0.719	0.676	0.805	0.807	0.586
Fuzzy Cmeans	9,49,64	0.749	0.656	0.805	0.939	0.595
SOM	7,34,81	0.716	0.67	0.805	0.807	0.582
Hierarchical	31,34,57	0.715	0.446	1	0.866	0.549
Model	26,33,63	0.678	0.312	1	0.817	0.585
RF	29,37,56	0.699	0.357	1	0.874	0.563
30 minute Motifs						
Kmeans	4,34,84	0.729	0.765	0.805	0.795	0.552
Fuzzy Cmeans	28,33,61	0.66	0.227	1	0.833	0.579
SOM	4,34,84	0.732	0.767	0.805	0.795	0.562
Hierarchical	4,25,93	0.717	0.764	0.805	0.721	0.577
Model	4,47,71	0.742	0.716	0.805	0.902	0.545
RF	24,42,56	0.665	0.228	1	0.863	0.57

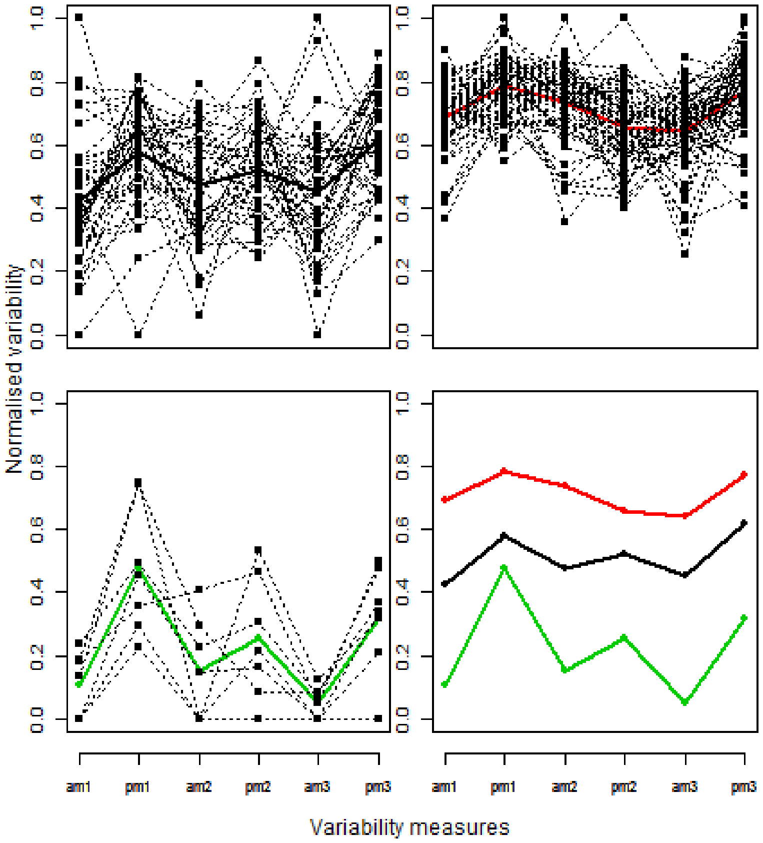


Figure A.1: Kmeans clusters and members of each cluster using 20 minute motif

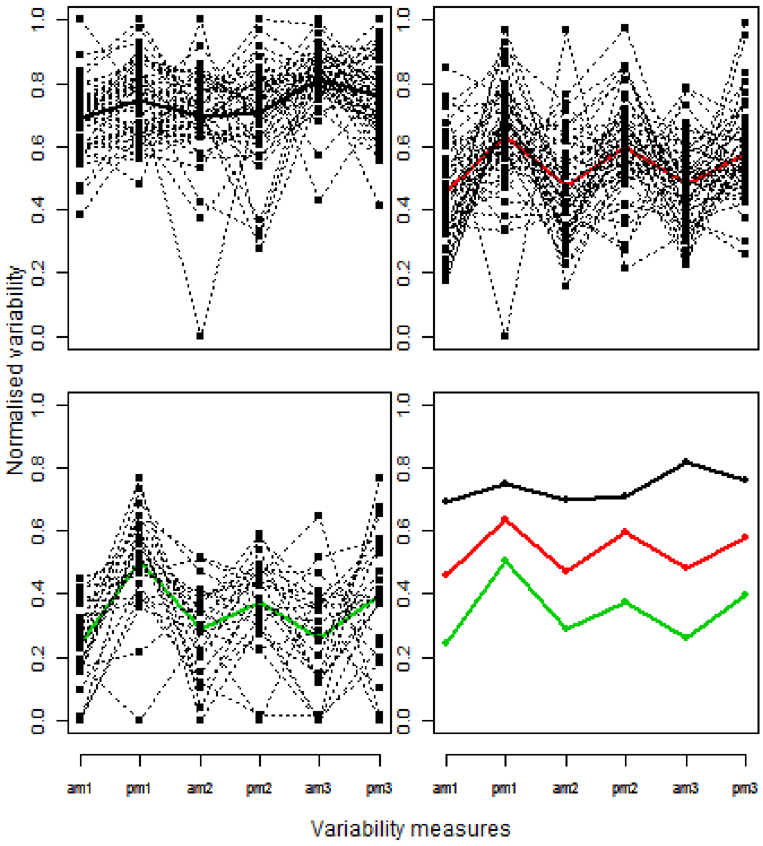


Figure A.2: Kmeans clusters and members of each cluster using 30 minute motif

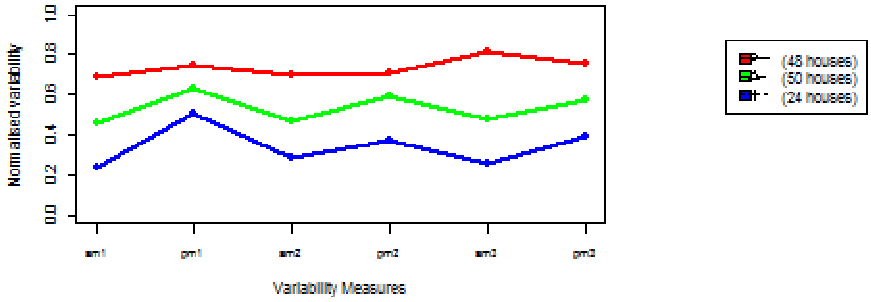
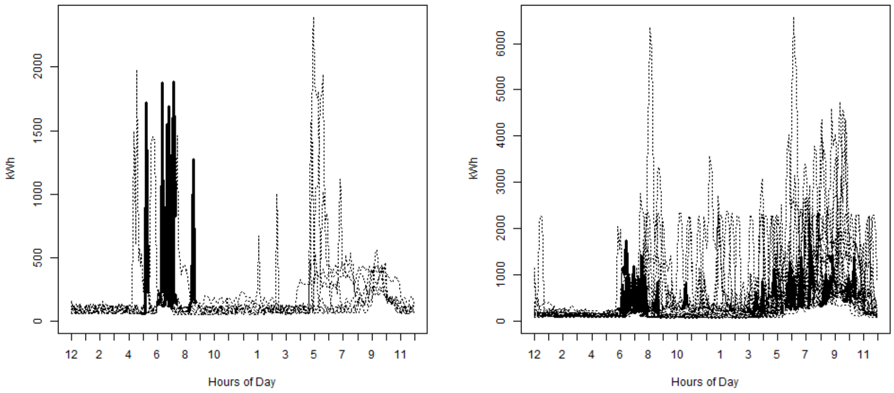
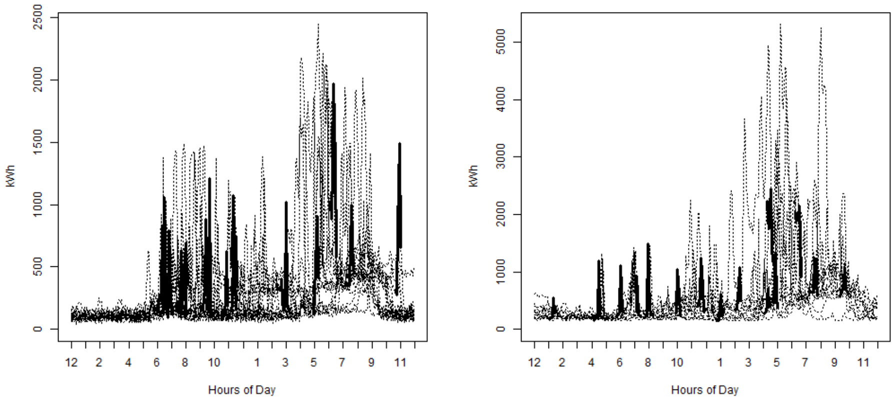


Figure A.3: Kmeans clusters using motif variability with 30 minute motif



(a) Household 397 **(b)** Household 250

Figure A.4: Examples of households from the low variability cluster (30 minute motif)



(a) Household 360 **(b)** Household 562

Figure A.5: Examples of households from the high variability cluster (30 minute motif)

APPENDIX B

Comparisons with Socio-Attitudinal data

The information collected by the questionnaires includes answers to the following questions.

1. Number of people regularly living in the household (ranges from 1 to 5)
2. Type of house (with the values as listed in Table B.1.
3. Number of bedrooms (1,2,3,4,5 or more than 5). This includes spare rooms and studies.
4. "As the last person to leave the room, I switch off the lights"
5. "I leave electrically powered appliances (e.g. TV, DVD player, stereo, printer) on standby".
6. "I wait until I have a full load before doing the laundry"
7. "In winter I turn down the heat whenever I leave the house for more than 4 hours"
8. "I commute to work by car"
9. "I think I can contribute to tackling climate change by saving energy"
10. "It is pointless to save energy to tackle climate change"

11. What is your household type? (Single Household, Living together with partner, Living together with partner and children, Single Parent, Other)

The fourth to eighth questions are part of a series asking "how often do you do this?" and giving possible responses of Never, Seldom, Occasionally, Often, or Always. Questions 9 and 10 request responses from the range of Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, and Strongly Disagree.

These questions have been selected for exploration as questions 1-3 and 11 give background demographic information on the household, questions 4-8 gives some kind of indication of thinking about when to do activities in the house, and questions 9 and 10 give some kind of indication of people already actively thinking about their behaviour.

Table B.1: Types of house collected in questionnaire

Type Code	Type of House
1	Bungalow (detached)
2	Bungalow (semi-detached)
3	Flat / maisonette
4	House (detached)
5	House (semi-detached)
6	House (mid-terrace)
7	House (end-terrace)
8	Other

Analysis of important attributes

The results from this exercise are shown in Table B.2 which shows, for each of the questionnaire questions, the most important attribute for predicting the questionnaire answer, the number of different answers to the question, and the error rate for predicting the answer using the meter data.

The most important attribute is found by using a random forest approach.

Some of the questionnaire questions have a poor spread of responses (e.g. "Do you switch off the light?") and the error rate reflects the relatively easier ability to predict a response with a narrow spread of responses.

Overall the number of motifs appears to be more important than the variability in timing of motifs when using the motif data to predict the questionnaire responses.

Table B.2: Important attribute for each question

Question	Most important	Class values	Error rate
People in household	third_motif_time_pm	5	0.615
House Type	num_top_motif	8	0.689
Number of bedrooms	num_of_motifs	6	0.697
Switch off light	num_second_motif	5	0.5
Leave on standby	num_second_motif	5	0.615
Full washing load	num_second_motif	5	0.525
Turn down heating	num_third_motif	5	0.672
Car commute	num_second_motif	5	0.328
Can save energy	num_of_motifs	5	0.467
Energy saving pointless	third_motif_time_am	4	0.41
Household Type	num_top_motif	5	0.541

Questionnaire responses matched with partitions arising from using the Variability of motifs data set are shown below.

Number of people in the household

The number of people in the household tends to be correlated with the size of the house and the number of bedrooms.

The information compared with the load profile clusters is shown in Figure B.1.

It can be seen that the load profile for cluster 3 (12 houses) shows a high usage throughout the day (in comparison to other clusters which tend to have a morning and evening peak) and corresponds to households with larger numbers of members.

Comparison with the motif variability clusters is shown in Figure B.2.

The low variability cluster (the second) shows a comparatively large number of single person households.

Comparison with the motif count clusters is shown in Figure B.3.

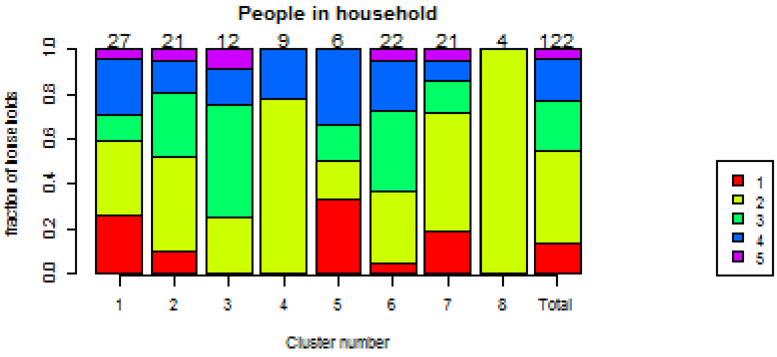


Figure B.1: Load Profile Ensemble Clusters related to numbers in the household

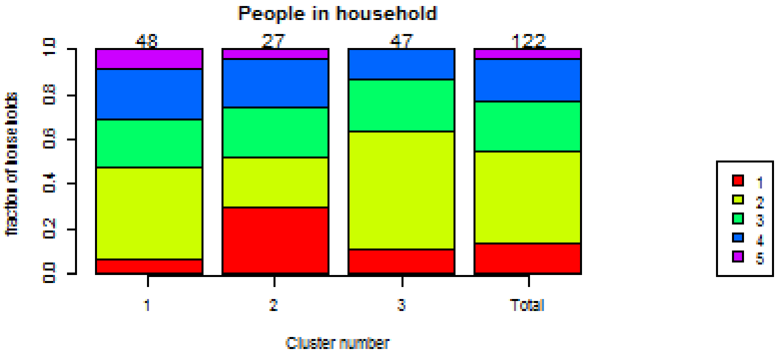


Figure B.2: Motif variability Ensemble Clusters related to numbers in the household

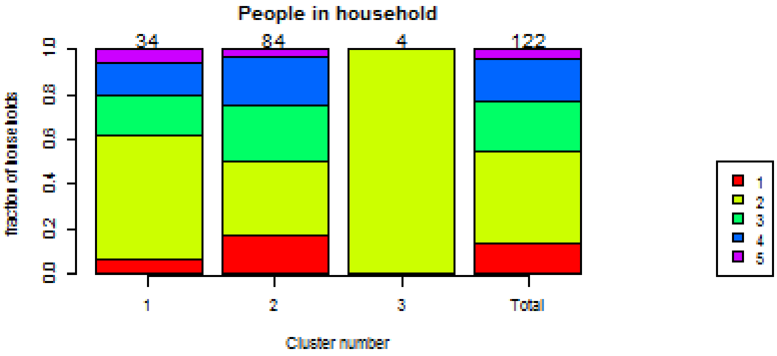


Figure B.3: Motif count Ensemble Clusters related to numbers in the household

Type of house

The information compared with the load profile clusters is shown in Figure B.4. Comparison with the motif variability clusters is shown in

Figure B.5.

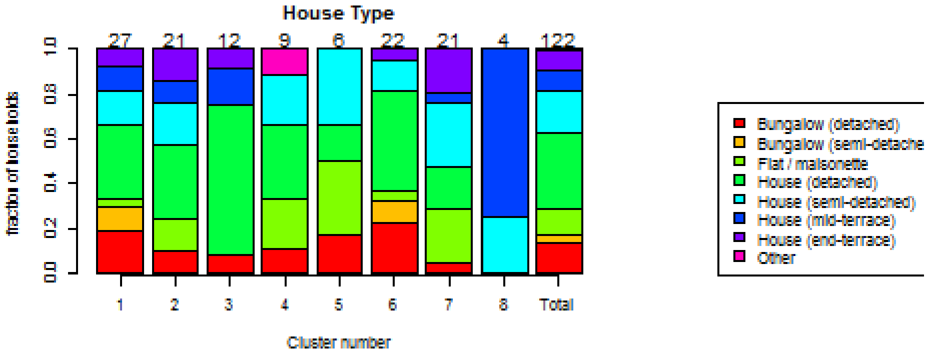


Figure B.4: Load Profile Ensemble Clusters related to house type

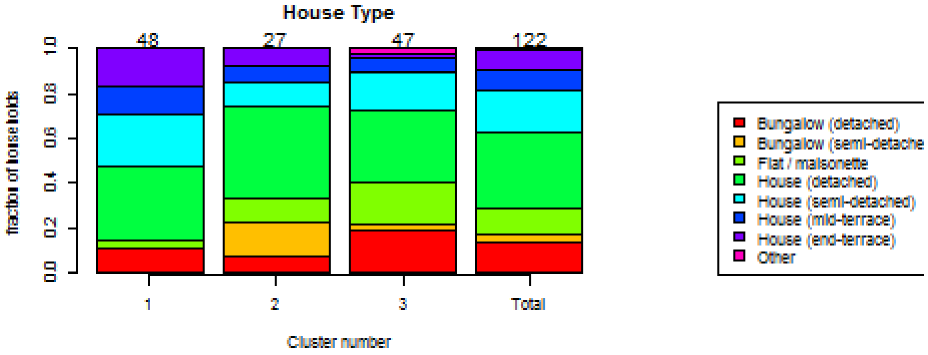


Figure B.5: Motif variability Ensemble Clusters related to house type

Number of bedrooms

The information compared with the load profile clusters is shown in Figure B.6. Comparison with the motif variability clusters is shown in Figure B.7 and with motif count clusters in Figure B.8.

The motif count cluster with low numbers of motifs (cluster 2) contains all the households with single bedrooms.

Household type

The information compared with the load profile clusters is shown in Figure B.9.

Comparison with the motif variability clusters is shown in Figure B.10.

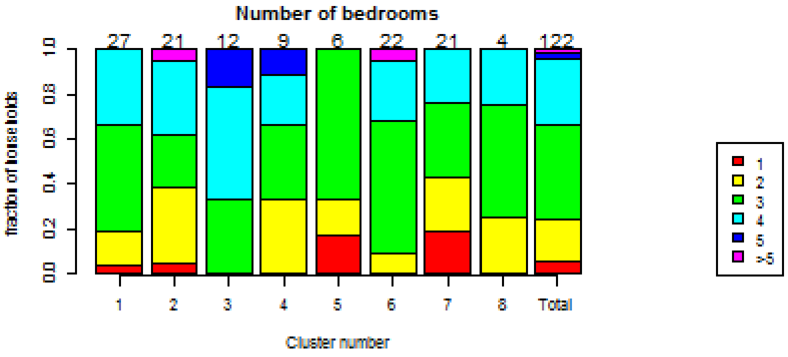


Figure B.6: Load Profile Ensemble Clusters related to number of bedrooms

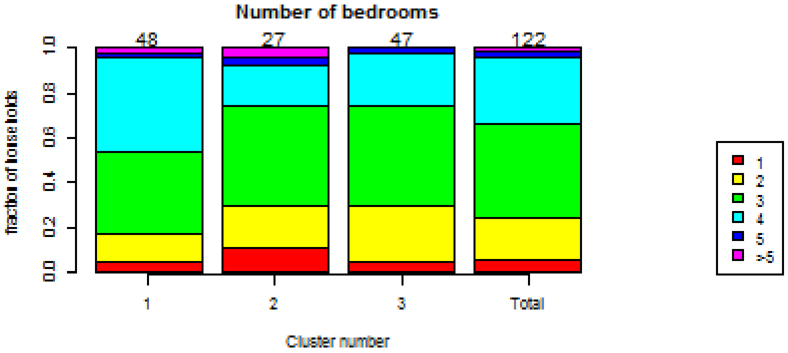


Figure B.7: Motif variability Ensemble Clusters related to number of bedrooms

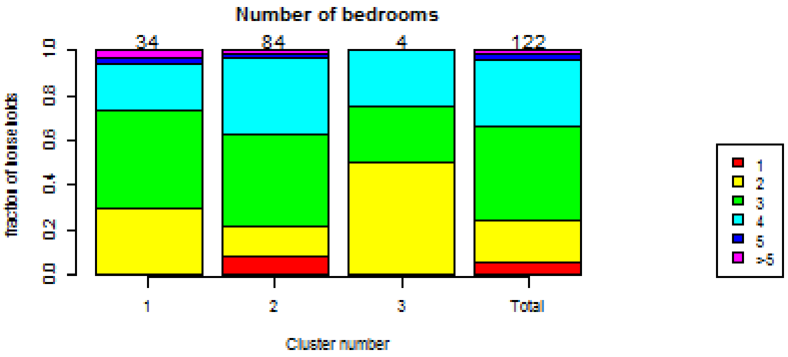


Figure B.8: Motif count Ensemble Clusters related to number of bedrooms

Comparison with the motif count clusters is shown in Figure B.11.

The single households seem to map onto the low variability cluster for both the motif variability and motif counts data. The profile data seems

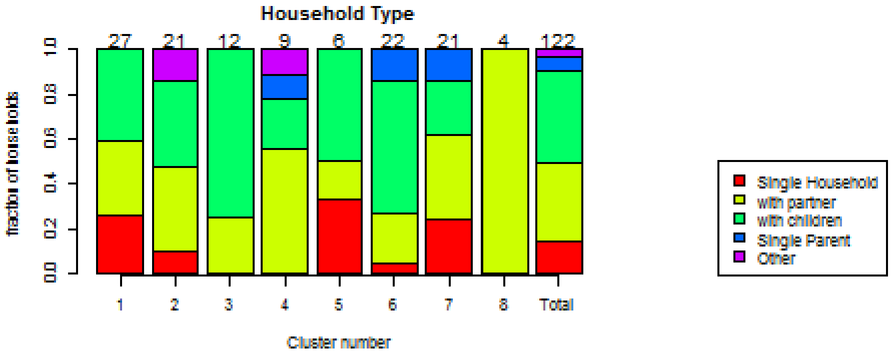


Figure B.9: Load Profile Ensemble Clusters related to household type

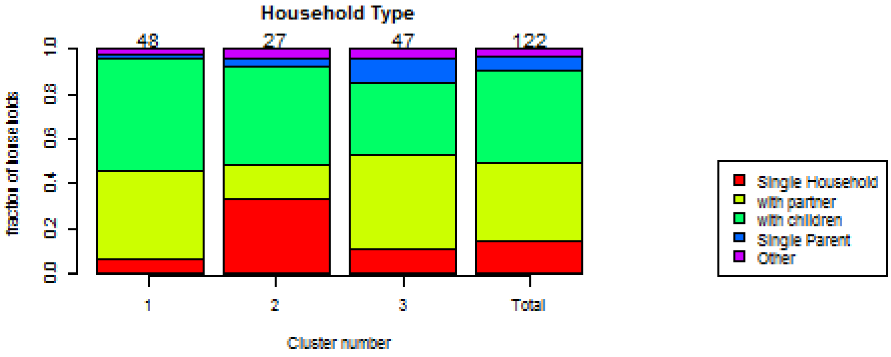


Figure B.10: Motif variability Ensemble Clusters related to household type

to identify the single parent households (in clusters 4, 6 and 7) although total numbers of this type of household are low.

Switching off lights attitude

The information compared with the motif variability clusters is shown in Figure B.12. The questionnaire responses tend to be mainly "often" and "always" with few people giving negative answers so it is hard for the clustering to distinguish between the different answers.

Leaving appliances on standby

The information compared with the motif variability clusters is shown in Figure B.13.

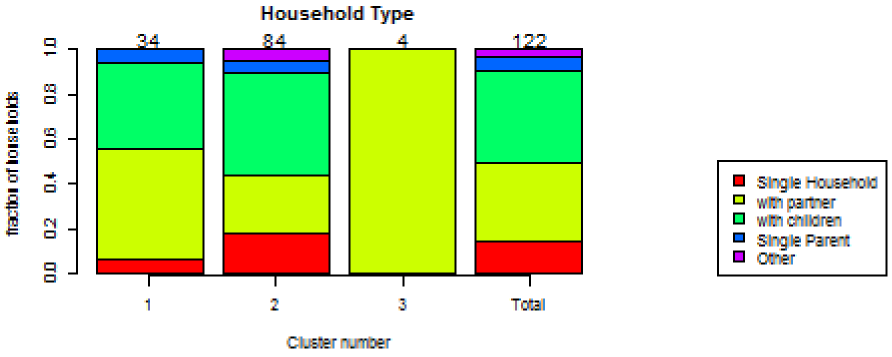


Figure B.11: Motif count Ensemble Clusters related to household type

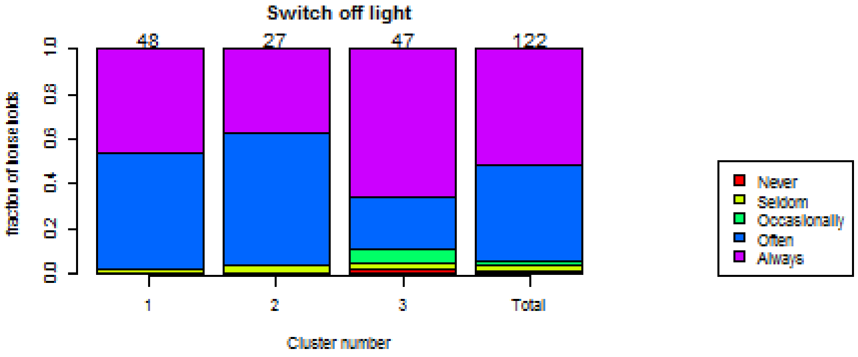


Figure B.12: Motif variability Ensemble Clusters related to attitude to switching off lights

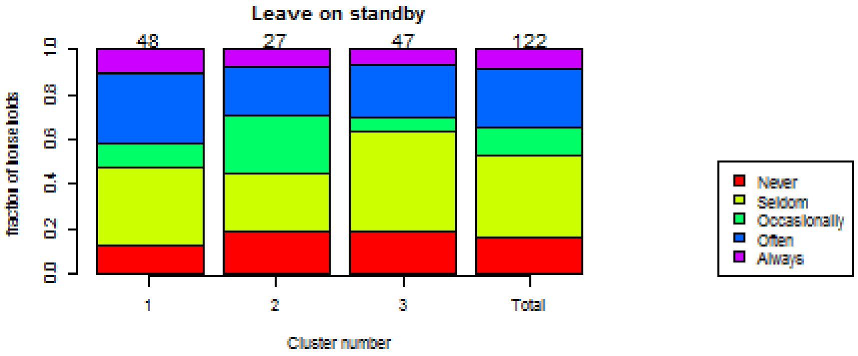


Figure B.13: Motif variability Ensemble Clusters related to attitude to using standby

Full washing load

The information compared with the motif variability clusters is shown in Figure B.14. There is a poor spread of responses to the question and the clustering analysis is not able to find an obvious differentiation between the responses.

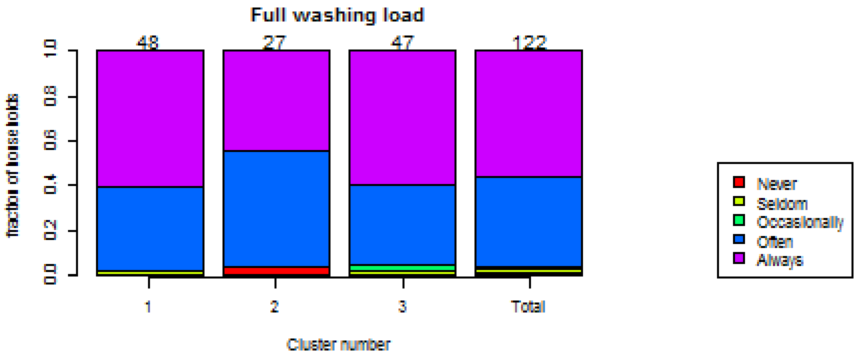


Figure B.14: Motif variability Ensemble Clusters related to using a full washing load

Turn down heating

The information compared with the motif variability clusters is shown in Figure B.15.

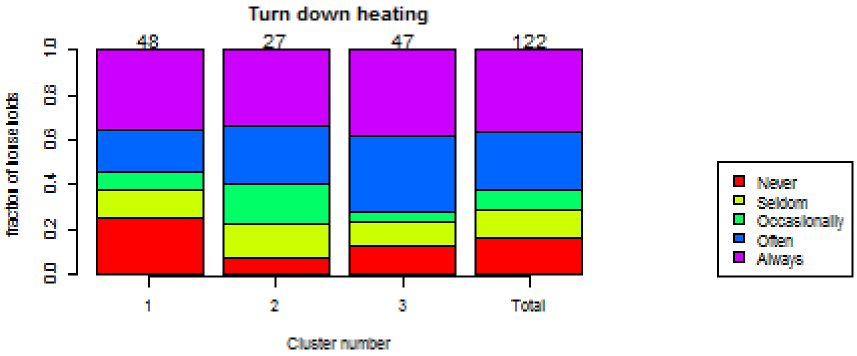


Figure B.15: Motif variability ensemble clusters related to heating adjustment

Commuting by car

The information compared with the motif variability clusters is shown in Figure B.16.

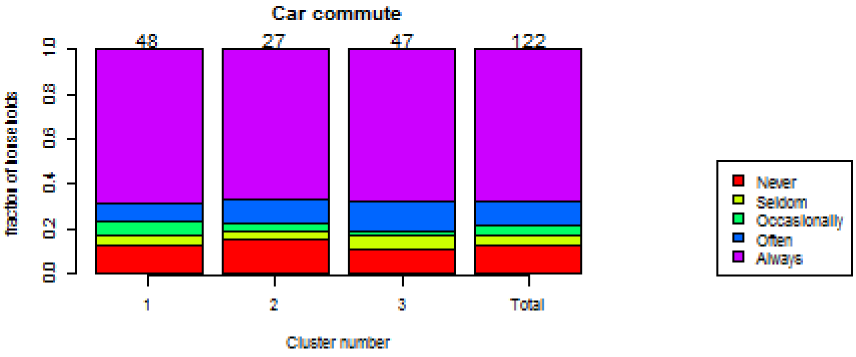


Figure B.16: Motif variability Ensemble Clusters related to commuting by car

Attitude to saving energy

Questions 9 and 10 from the questionnaire ask the same question but phrased differently so the results have been presented together. The information compared with the motif variability clusters is shown in Figures B.17 and B.18.

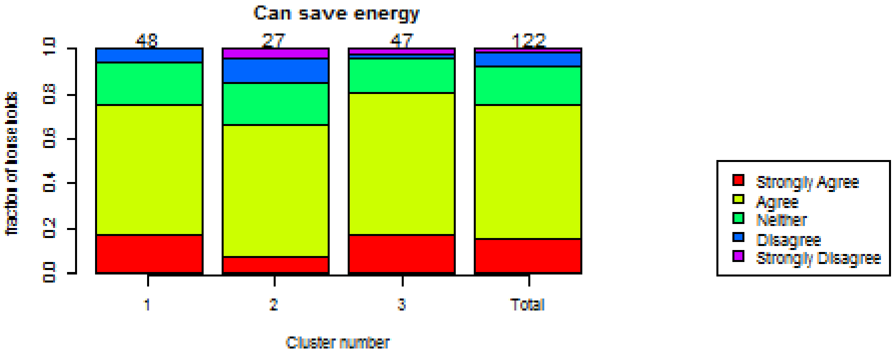


Figure B.17: Motif variability Ensemble Clusters related to belief that can save energy

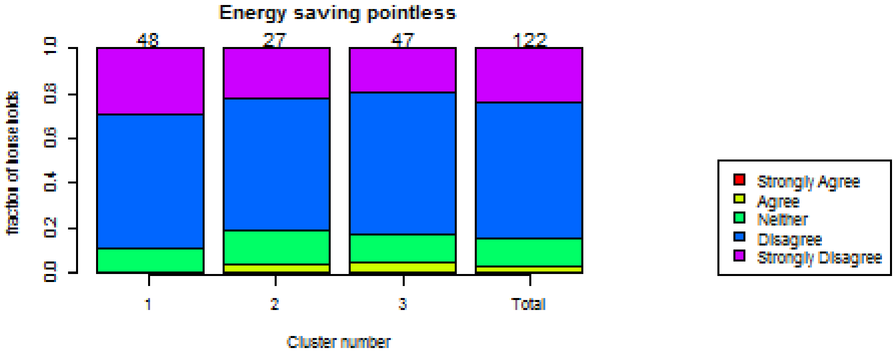


Figure B.18: Motif variability Ensemble Clusters related to view that saving energy is pointless

Bibliography

[1] Iain MacLeay, Kevin Harris, and Anwar Annut. Digest of United Kingdom Energy Statistics 2013. 2013.

[2] Aidan Rhodes. Smart Grids: Commercial Opportunities and Challenges for the UK. Technical report, Energy Generation and Supply Knowledge Transfer Network, 2010.

[3] DECC. Towards a Smarter Future, Government Response to the Consultation on Electricity and Gas Smart Metering. 2009.

[4] River. Primer on demand-side management with an emphasis on price-responsive programs. *prepared for The World Bank by Charles River Associates, Tech. Rep*, 2005.

[5] Serena Hesmondhalgh. GB Electricity Demand - 2010 and 2025. Initial Brattle Electricity Demand-Side Model - Scope for Demand Reduction and Flexible Response. Technical report, The Brattle Group, 2012.

[6] M. Sarstedt and E. Mooi. *A concise guide to market research: The process, data, and methods using IBM SPSS statistics*. Springer Verlag, 2011.

[7] Kajsa Ellegård and Jenny Palm. Visualizing energy consumption activities as a tool for making everyday life more sustainable. *Applied Energy*, 88(5):1920–1926, 2011.

[8] Richard Smith. UK Future Energy Scenarios. Technical report, National Grid, 2013.

[9] Department of Energy & Climate Change. Smart Meters Programme Plan. Technical report, Department of Energy & Climate Change, 2012.

[10] Xiaofan Jiang, Stephen Dawson-Haggerty, Prabal Dutta, and David Culler. Design and implementation of a high-fidelity ac metering network. In *Information Processing in Sensor Networks, 2009. IPSN 2009. International Conference on*, pages 253–264. IEEE, 2009.

[11] A.J. Collin, I. Hernando-Gil, J.L. Acosta, and S.Z. Djokic. An 11 kV steady state residential aggregate load model. Part 1: Aggregation methodology. In *PowerTech, 2011 IEEE Trondheim*, pages 1–8. IEEE, 2011.

[12] Takekazu Kato, Hyun Cho, Dongwook Lee, Tetsuo Toyomura, and Tatsuya Yamazaki. Appliance Recognition from Electric Current Signals for Information-Energy Integrated Network in Home Environments. In Mounir Mokhtari, Ismail Khalil, Jeremy Bauchet, Daqing Zhang, and Chris Nugent, editors, *Ambient Assistive Health and Wellness Management in the Heart of the City*, volume 5597 of *Lecture Notes in Computer Science*, pages 150–157. Springer Berlin / Heidelberg, 2009.

[13] CompSci. *Computer Science Undergraduate study brochure*. University of Nottingham, 2014. URL `http://www.nottingham.ac.uk/ugstudy/downloads/school-brochure-pdf/computer-science.pdf`.

[14] G. Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, Volume 42, Issue 1:68âĂŞ80, June 2012.

[15] Tony Craig, J Gary Polhill, Ian Dent, Carlos Galan-Diaz, and Simon Heslop. The North East Scotland Energy Monitoring Project: Exploring relationships between household occupants and energy usage. *Energy and Buildings*, 2014.

[16] Iain MacLeay, Kevin Harris, and Anwar Annut. *Digest of United Kingdom Energy Statistics 2012*. The Stationery Office/TSO, 2012.

[17] Ian Marlee. Project Discovery; Options for delivering secure and sustainable energy supplies. Technical report, Ofgem, 2010.

[18] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods: part I. *ACM Sigmod Record*, 31(2):40–45, 2002.

[19] M. Zeifman and K. Roth. Nonintrusive appliance load monitoring: Review and outlook. *Consumer Electronics, IEEE Transactions on*, 57 (1):76–84, 2011.

[20] Poyry. Options for Low-Carbon Power Sector Flexibility to 2050. 2010.

[21] DECC. Implementation Programme: Response to Prospectus Consultation: Overview Document. Technical report, Department of Energy and Climate Change, 2011.

[22] The Parliamentary Office of Science and Technology. Future Electricity Networks. Technical report, The Parliamentary Office of Science and Technology, 2011.

[23] UK Parliament. Climate Change Act 2008, 2008.

[24] Committee on Climate Change. The Fourth Carbon Budget Reducing emissions through the 2020s. 2010.

[25] European Parliament and the Council of the European Union. Directive 2009/28/EC of the European Parliament and of the Council. *Official Journal of the European Union*, 2009.

[26] Committee on Climate Change. Next steps on Electricity Market Reform - securing the benefits of low-carbon investment. Technical report, Committee on Climate Change, 2013.

[27] DOE. Grid 2030: A National Vision For Electricity's Second 100 Years. 2003.

[28] Cap Gemini. Smart electricity; threat and promise. Technical report, Cap Gemini, 2010.

[29] Cap Gemini. Smart Grid: Enabling Operational Efficiency and Distributed Generation. Technical report, Cap Gemini, 2010.

[30] Cap Gemini. Smart Home: The Human Side of the Smart Grid. Technical report, Cap Gemini, 2010.

[31] Cap Gemini. Smart Metering: The Foundation for Smart Grid. Technical report, Cap Gemini, 2010.

[32] B. Dupont, L. Meeus, and R. Belmans. Measuring the smartness of the electricity grid. In *7th International Conference on the European Energy Market (EEM)*, pages 1–6. IEEE, 2010.

[33] Microsoft Power and Utilities. Smart Energy Reference Architecture. Technical report, Microsoft Power and Utilities, 2009.

[34] Ofgem. RIIO: A new way to regulate electricity networks. Technical report, Ofgem, 2010.

[35] DECC. Quantitative Research into Public Awareness, Attitudes, and Experience of Smart Meters. Technical report, Department of Energy and Climate Change, August 2012.

[36] Colette Cuijpers and Bert-Jaap Koops. Smart metering and privacy in Europe: Lessons from the Dutch case. In *European data protection: coming of age*, pages 269–293. Springer, 2013.

[37] Gary Raw and David Ross. Energy Demand Research Project: Final Analysis. Technical report, Ofgem, 2011.

[38] DECC. Smart Metering Equipment Technical Specifications: version 2. Technical report, Department of Energy and Climate Change, 2013.

[39] M. Newborough and P. Augood. Demand-side management opportunities for the UK domestic sector. In *Generation, Transmission and Distribution, IEE Proceedings-*, volume 146, pages 283–293. IET, 1999.

[40] Ahmed Faruqui; John Chamberlin. Principles and Practice of Demand Side Management. Technical report, Barakat & Chamberlin Inc, 1993.

[41] V. Hamidi, F. Li, and F. Robinson. Demand response in the UK's domestic sector. *Electric Power Systems Research*, 79(12):1722–1726, 2009.

[42] S Zachary, CJ Dent, and DJ Brayshaw. Challenges in quantifying wind generation's contribution to securing peak demand. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1–8. IEEE, 2011.

[43] S.J. Darby and E. McKenna. Social implications of residential demand response in cool temperate climates. *Energy Policy*, 2012.

[44] Jacopo Torriti, Mohamed G Hassan, and Matthew Leach. Demand response experience in Europe: Policies, programmes and implementation. *Energy*, 35(4):1575–1583, 2010.

[45] Corinna Fischer. Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency*, 1(1):79–104, 2008.

[46] B. Neenan. Residential Electricity Use Feedback: A Research Synthesis and Economic Framework. Technical report, Electric Power Research Institute (EPRI), 2009.

[47] Rachelle M Willis, Rodney A Stewart, Kriengsak Panuwatwanich, Sarah Jones, and Andreas Kyriakides. Alarming visual display monitors affecting shower end use water and energy conservation in Australian residential households. *Resources, Conservation and Recycling*, 54(12):1117–1127, 2010.

[48] Phil Kotler and Kevin Keller. *Marketing Management*. Prentice Hall, 13 edition, February 1967, 1984, 1994 and 2008. ISBN 0136009980.

[49] S. Dibb. Criteria guiding segmentation implementation: reviewing the evidence. *Journal of Strategic Marketing*, 7(2):107–129, 1999.

[50] A.K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[51] Electricity Association. Load profiles and their use in electricity settlement. *UKERC*, 1997.

[52] SV Allera and AG Horsburgh. Load profiling for the energy trading and settlements in the UK electricity markets. In *Proc. DistribuTECH Europe DA/DSM Conference*, pages 27–29, 1998.

[53] J. Bailey. Load Profiling for Retail Choice:: Examining a Complex and Crucial Component of Settlement. *The Electricity Journal*, 13 (10):69–74, 2000. ISSN 1040-6190.

[54] L.G. Swan and V.I. Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, 13(8):1819–1835, 2009.

[55] Keith J. Baker and R. Mark Rylatt. Improving the prediction of UK domestic energy-demand using annual consumption-data. *Applied Energy*, 85(6):475 – 482, 2008. ISSN 0306-2619.

[56] V. Figueiredo, D. Rodrigues, and Z. Vale. Simulating DSM impact in the new liberalized electricity market. Technical report, Polytechnic Institute of Porto, School of Engineering, 2005, 2005.

[57] George Gross and Francisco D Galiana. Short-term load forecasting. *Proceedings of the IEEE*, 75(12):1558–1573, 1987.

[58] A.E. Cancino. Load Profiling of MERALCO Residential Electricity Consumers using Clustering Methods. *18th Conference of Electric Power Supply Industry (CEPSI*, 2010.

[59] JK Lin, SK Tso, HK Ho, CM Mak, KM Yung, and YK Ho. Study of climatic effects on peak load and regional similarity of load profiles following disturbances based on data mining. *International Journal of Electrical Power & Energy Systems*, 28(3):177–185, 2006. ISSN 0142-0615.

[60] Z.H. Zakaria and KL Lo. Load profiling in the new electricity market. In *Research and Development, 2002. SCOReD 2002. Student Conference on*, pages 278–281. IEEE, 2002.

[61] AM Ihbal, HS Rajamani, RA Abd-Alhameed, and MK Jalboub. Statistical predictions of electric load profiles in the UK domestic buildings. In *Energy, Power and Control (EPC-IQ), 2010 1st International Conference on*, pages 345–350. IEEE, 2010.

[62] A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi. A bottom-up approach to residential load modeling. *Power Systems, IEEE Transactions on*, 9(2):957–964, 1994.

[63] J.V. Paatero and P.D. Lund. A model for generating household electricity load profiles. *International journal of energy research*, 30(5): 273–290, 2006.

[64] J.A. Dominguez-Navarro, J.L. Bernal-Agustín, and R. Dufo-López. Data mining methodology for disaggregation of load demand. *Electric Power Systems Research*, 79(10):1393–1399, 2009. ISSN 0378-7796.

[65] G. Chicco, R. Napoli, and F. Piglione. Application of clustering algorithms and self organising maps to classify electricity customers. In *Power Tech Conference Proceedings, 2003 IEEE Bologna*, volume 1. IEEE, 2003.

[66] D. Gerbec, S. Gasperic, and F. Gubina. Determination and allocation of typical load profiles to the eligible consumers. In *Power Tech Conference Proceedings, 2003 IEEE Bologna*, volume 1, page 5. IEEE, 2004. ISBN 0780379675.

[67] S. Ramos, Z. Vale, J. Santana, and J. Duarte. Data Mining Contributions to Characterize MV Consumers and to Improve the Suppliers-Consumers Settlements. In *Power Engineering Society General Meeting, 2007. IEEE*, pages 1–8. IEEE, 2007. ISBN 142441296X.

[68] S. Ramos, JMM Duarte, J. Soares, Z. Vale, and FJ Duarte. Typical load profiles in the smart grid context - A clustering methods comparison. In *Power and Energy Society General Meeting*, pages 1–8. IEEE, 2012.

[69] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader. Emergent electricity customer classification. In *Generation, Transmission and Distribution, IEE Proceedings-*, volume 152, pages 164–172. IET, 2005.

[70] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader. A review of concepts and techniques for emergent cus-

tomer categorisation. In *TELMARK Discussion Forum European Electricity Markets, London*, 2002.

[71] R. Singh, B.C. Pal, and R.A. Jabr. Statistical representation of distribution system loads using Gaussian mixture model. *Power Systems, IEEE Transactions on*, 25(1):29–37, 2010. ISSN 0885-8950.

[72] G. Chicco and JS Akilimali. Renyi entropy-based classification of daily electrical load patterns. *Generation, Transmission & Distribution, IET*, 4(6):736–745, 2010. ISSN 1751-8687.

[73] SM Bidoki, N. Mahmoudi-Kohan, MH Sadreddini, Z. Jahromi, and MP Moghaddam. Evaluating different clustering techniques for electricity customer classification. In *Transmission and Distribution Conference and Exposition, 2010 IEEE PES*, pages 1–5. IEEE, 2010.

[74] G. Chicco and I.S. Ilie. Support Vector Clustering of Electrical Load Pattern Data. *Power Systems, IEEE Transactions on*, 24(3):1619–1628, 2009. ISSN 0885-8950.

[75] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader. Load pattern-based classification of electricity customers. *Power Systems, IEEE Transactions on*, 19(2):1232–1239, 2004. ISSN 0885-8950.

[76] M. Gavrilas, G. Gavrilas, and C.V. Sfintes. Application of Honey Bee Mating Optimization algorithm to load profile clustering. In *Computational Intelligence for Measurement Systems and Applications (CIMSA), 2010 IEEE International Conference on*, pages 113–118. IEEE.

[77] D. Gerbec, S. Gasperic, and F. Gubina. Comparison of Different Classification Methods for the Consumers' Load Profile Determination. In *17th International Conference on Electricity Distribution, CIRED, Barcelona, vol. Session*, volume 6, 2003.

[78] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina. An approach to customer's daily load profile determination. In *Power Engineering Society Summer Meeting, 2002 IEEE*, volume 1, pages 587–591. IEEE, 2002. ISBN 0780375181.

[79] V. Figueiredo, F. Rodrigues, Z. Vale, and J.B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *Power Systems, IEEE Transactions on*, 20(2):596–602, 2005. ISSN 0885-8950.

[80] M. Garamvolgyi and L. Varga. Electricity pricing in liberalized market using consumer characterization. In *Energy Market, 2009. EEM 2009. 6th International Conference on the European*, pages 1–6. IEEE, 2009.

[81] J.J. Lopez, J.A. Aguado, F. Martin, F. Munoz, A. Rodriguez, and J.E. Ruiz. Electric customer classification using Nopfield recurrent ANN. In *Electricity Market, 2008. EEM 2008. 5th International Conference on European*, pages 1–6. IEEE, 2008.

[82] H. Mori. State-of-the-art overview on data mining in power systems. In *Power Systems Conference and Exposition, 2006. PSCE'06. 2006 IEEE PES*, pages 33–34. IEEE, 2007. ISBN 1424401771.

[83] F. Gullo, G. Ponti, A. Tagarelli, et al. Low-voltage electricity customer profiling based on load data clustering. In *Proceedings of the 2009 International Database Engineering & Applications Symposium*, pages 330–333. ACM, 2009.

[84] J.A. Jardini, C.M.V. Tahan, MR Gouvea, S.U. Ahn, and FM Figueiredo. Daily load profiles for residential, commercial and industrial low voltage consumers. *Power Delivery, IEEE Transactions on*, 15(1):375–380, 2002. ISSN 0885-8977.

[85] I. Dent, C. Wagner, U. Aickelin, and T. Rodden. Creating Personalised Energy Plans: From Groups to Individuals using Fuzzy C Means Clustering. In *Digital Engagement 11, Newcastle*, November 2011.

[86] RF Chang and CN Lu. Load profile assignment of low voltage customers for power retail market applications. In *Generation, Transmission and Distribution, IEE Proceedings-*, volume 150, pages 263–267. IET, 2005.

[87] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice Hall College Div, 1988.

[88] B. Davito, H. Tai, and R. Uhlaner. The smart grid and the promise of demand-side management. *McKinsey on Smart Grid*, pages 38–44, 2010.

[89] W. Abrahamse, L. Steg, C. Vlek, and T. Rothengatter. A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, 25(3):273–291, 2005. ISSN 0272-4944.

[90] Paula Owen. Powering the Nation: Household electricity using habits revealed. Technical report, Energy Saving Trust, 2012.

[91] H.H. Chang, C.L. Lin, and J.K. Lee. Load identification in non-intrusive load monitoring using steady-state and turn-on transient energy algorithms. In *Computer Supported Cooperative Work in Design (CSCWD), 2010 14th International Conference on*, pages 27–32. IEEE, 2010.

[92] S. Firth, K. Lomas, A. Wright, and R. Wall. Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings*, 40(5):926–936, 2008.

[93] Rob Raine. Assessment of Electricity Demand-Side Management Technologies. Technical report, University of Sheffield, 2012.

[94] Rainer Stamminger, Gereon Broil, Christiane Pakula, Heiko Jungbecker, Maria Braun, Ina Rüdenauer, and Christoph Wendker. Synergy potential of smart appliances. *Report of the Smart-A project*, 2008.

[95] Jason Lines, Anthony Bagnall, Patrick Caiger-Smith, and Simon Anderson. Classification of household devices by electricity usage profiles. *Intelligent Data Engineering and Automated Learning*, pages 403–412, 2011.

[96] H. Allcott. Rethinking real-time electricity pricing. *Resource and Energy Economics*, 2011.

[97] D.S. Kirschen. Demand-side view of electricity markets. *Power Systems, IEEE Transactions on*, 18(2):520–527, 2003. ISSN 0885-8950.

[98] N. Mahmoudi-Kohan, M.P. Moghaddam, and MK Sheikh-El-Eslami. An annual framework for clustering-based pricing for an electricity retailer. *Electric Power Systems Research*, 80(9):1042–1048, 2010. ISSN 0378-7796.

[99] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader. Customer characterization options for improving the tariff offer. *Power Systems, IEEE Transactions on*, 18(1):381–387, 2003. ISSN 0885-8950.

[100] H. Allcott. Social norms and energy conservation. *Journal of Public Economics*, 2011.

[101] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[102] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Pranav Patel. Finding motifs in time series. In *Proc. of the 2nd Workshop on Temporal Data Mining*, pages 53–68, 2002.

[103] Gautam Das, King-Ip Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth. Rule Discovery from Time Series. In *KDD*, volume 98, pages 16–22, 1998.

[104] Frank Höppner. Discovery of temporal patterns. In *Principles of Data Mining and Knowledge Discovery*, pages 192–203. Springer, 2001.

[105] N.C. Castro. Time Series Data Mining. Semana de Engenharia, October 2010.

[106] J. Lin, E. Keogh, L. Wei, and S. Lonardi. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, 15(2):107–144, 2007.

[107] J. Shieh and E. Keogh. i SAX: indexing and mining terabyte sized time series. In *Proceeding of the 14th ACM SIGKDD international*

conference on Knowledge discovery and data mining, pages 623–631. ACM, 2008.

[108] H. Mori and Y. Umezawa. A SAX-based method for extracting features of electricity price in power markets. In *Transmission & Distribution Conference & Exposition: Asia and Pacific, 2009*, pages 1–4. IEEE.

[109] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[110] Myra Spiliopoulou, Irene Ntoutsi, Yannis Theodoridis, and Rene Schult. Monic: modeling and monitoring cluster transitions. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 706–711. ACM, 2006.

[111] Rosa Meo, Dipankar Bachar, and Dino Ienco. LODE: A distance-based classifier built on ensembles of positive and negative observations. *Pattern Recognition*, 45(4):1409–1425, 2012.

[112] James C Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

[113] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78 (9):1464–1480, 2002. ISSN 0018-9219.

[114] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster analysis*. Edward Arnold, London, 2001.

[115] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[116] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[117] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[118] Chris Fraley, Adrian E Raftery, T Brendan Murphy, and Luca Scrucca. MCLUST version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, Technical report, 2012.

[119] Chris Fraley. Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281, 1998.

[120] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[121] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[122] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[123] José Manuel Pena, Jose Antonio Lozano, and Pedro Larranaga. An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern recognition letters*, 20(10):1027–1040, 1999.

[124] AD Gordon and M Vichi. Partitions of partitions. *Journal of Classification*, 15(2):265–285, 1998.

[125] Kurt Hornik. A CLUE for CLUster ensembles. *Journal of Statistical Software*, 14(12), 2005.

[126] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[127] I. Dent, T. Craig, U. Aickelin, and T. Rodden. A Method for Cleaning and Storing Electricity Meter Data for Flexible Analysis. In *BeHave 2012, Helsinki*, 2012.

[128] R. Kimball, L. Reeves, W. Thornthwaite, M. Ross, and W. Thornwaite. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses with CD Rom*. John Wiley & Sons, Inc., New York, NY, 1998.

[129] D. De Silva, X. Yu, D. Alahakoon, and G. Holmes. A Data Mining Framework for Electricity Consumption Analysis From Meter Data. *Industrial Informatics, IEEE Transactions on*, (99):1–1, 2011.

[130] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[131] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. R package version 1.6-3.

[132] Wickham and Hadley. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 2007.

[133] R. Wehrens and L.M.C. Buydens. Self- and Super-organising Maps in R: the kohonen package. *J. Stat. Softw.*, 21(5), 2007.

[134] Christian Hennig. *fpc: Flexible procedures for clustering*, 2014.

[135] David A. James and Saikat DebRoy. *RMySQL: R interface to the MySQL database*, 2012.

[136] P. Roebuck. *matlab: MATLAB emulation package*, 2014.

[137] Lukasz Nieweglowski. *clv: Cluster Validation Techniques*, 2013.

[138] Ramon Diaz-Uriarte. *varSelRF: Variable selection using random forests*, 2010.

[139] Marcello D'Orazio. *StatMatch: Statistical Matching (aka data fusion)*, 2013.

[140] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering validity checking methods: part II. *ACM Sigmod Record*, 31(3):19–27, 2002.

[141] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[142] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, and M. Cordeiro. A comparative analysis of clustering algorithms applied to load

profiling. *Machine Learning and Data Mining in Pattern Recognition*, pages 73–85, 2003.

[143] GJ Tsekouras, FD Kanellos, VT Kontargyri, IS Karanasiou, AD Salis, and NE Mastorakis. A new classification pattern recognition methodology for power system typical load profiles. *WSEAS Transactions on Circuits and Systems*, 7(12):1090–1104, 2008. ISSN 1109-2734.

[144] Tao Zhang, Peer-Olaf Siebers, and Uwe Aickelin. A three-dimensional model of residential energy consumer archetypes for local energy policy design in the UK. *Energy Policy*, 2012.

[145] Dominique Gay, Nazha Selmaoui-Folcher, and Jean-Francois Boulicaut. Application-independent feature construction based on almost-closedness properties. *Knowledge and information systems*, 30(1):87–111, 2012.

[146] Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover. Exact discovery of time series motifs. In *Proc. of 2009 SIAM International Conference on Data Mining*, pages 1–12, 2009.

[147] I. Dent, T. Craig, U. Aickelin, and T. Rodden. Finding the creatures of habit; Clustering households based on their flexibility in using electricity. In *Digital Futures, Aberdeen, UK*, 2012.

[148] CER. Electricity Smart Metering Technology Trials Findings Report. Technical report, Commission for Energy Regulation, May 2011.

[149] J.E. Fischer, S.D. Ramchurn, M.A. Osborne, O. Parson, T.D. Huynh, M. Alam, N. Pantidi, S. Moran, K. Bachour, S. Reece, et al. Recommending Energy Tariffs and Load Shifting Based on Smart Household Usage Profiling. 2013.

[150] I. Richardson, M. Thomson, D. Infield, and C. Clifford. Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, 42(10):1878–1887, 2010.

[151] Ian Dent, Tony Craig, Uwe Aickelin, and Tom Rodden. Variability of Behaviour in Electricity Load Profile Clustering; Who Does

Things at the Same Time Each Day? In Petra Perner, editor, *Advances in Data Mining*, volume 8557 of *Lecture Notes in Computer Science*, pages 70–84. Springer, 2014. ISBN 978-3-319-08975-1. doi: 10.1007/978-3-319-08976-8_6.

[152] I. Dent, T. Craig, U. Aickelin, and T. Rodden. An Approach for Assessing Clustering of Households by Electricity Usage. In *UKCI 2012, 12th Workshop on Computational Intelligence*, 2012.

[153] Aristides Kiprakis, Ian Dent, Sasa Djokic, and Stephen McLaughlin. Multi-scale Dynamic Modeling to Maximize Demand Side Management. In *IEEE Power and Energy Society Innovative Smart Grid Technologies Europe 2011, Manchester, UK*, 2011.

[154] Ian Dent, Uwe Aickelin, and Tom Rodden. The Application of a Data Mining Framework to Energy Usage Profiling in Domestic Residences using UK data. In *Proc. Research Student Conference on Buildings Do Not Use Energy, People Do*, 2011.

[155] Ian Dent, Uwe Aickelin, and Tom Rodden. Application of a clustering framework to UK domestic electricity data. In *11th Annual Workshop on Computational Intelligence*, 2011.