Gas Smart Meters and Indicators of Fuel Poverty

[redacted]

Supervised by Dr [redacted]

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

August 2017

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

This dissertation examines how households consume gas throughout the day, using data collected from smart meters. The goal is to identify groups of households whose consumption behaviours suggest they are at higher risk of fuel poverty, using socio-economic factors as indicators of fuel poverty. To introduce the concept of consumption profiles and to aid understanding, there is an analysis of some sampled households' individual gas use, looking at differences between households, and differences between days. The analysis includes various ways to visualise consumption with 2D and 3D plots. The methods explored for grouping households are Principal Component Analysis, Self-Organising Maps (SOMs), and k-means clustering. SOMs are covered in the greatest depth with an comparison of parameter choices used in literature. k-means clustering is applied to a trained SOM, and the distributions of employment, social class, and education are analysed across the resulting clusters. Most notably from the analysis, one cluster is identified which has over twice the mean rate of unemployment, and another cluster is identified which has over twice the mean rate of retirement, evidence that clustering of consumption profiles can be used to highlight households who may be at risk of fuel poverty.

Contents

List of Figures

List of Tables

1 Introduction

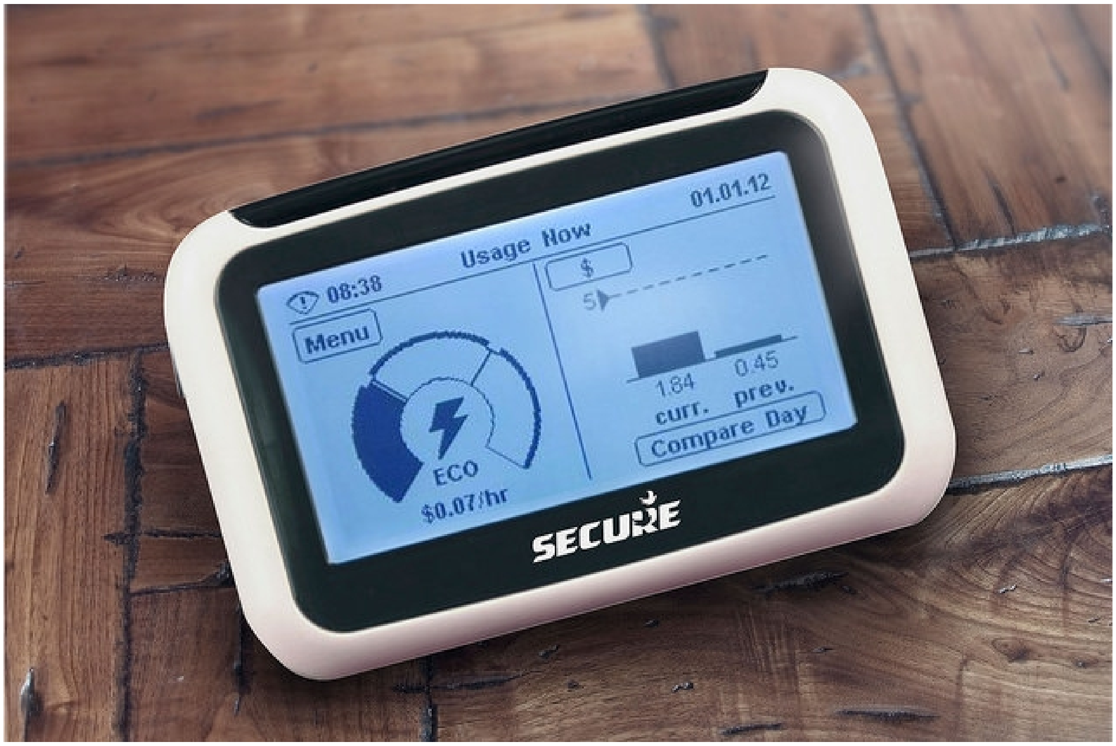
Smart meters can be considered the next generation of energy meter. The European Union (EU) advised all member governments to investigate the potential of smart meters for tackling climate change, and for updating energy supply (for example, moving towards a smart grid). In Annex 1 of Directive 2009/72/EC:

> "Where roll-out of smart meters is assessed positively, at least 80% of consumers shall be equipped with intelligent metering systems by 2020."

In the UK this equates to the replacement of around 53 million gas and electricity meters by the end of 2020. Smart meters mark a move from monthly, or quarterly, meter reading to half hourly; and with this a range of intelligent functions:

- near real time information on energy use through an In-Home Display (IHD);
- easier for the consumer to shift consumption to periods of less demand (potentially cheaper tariff);
- easier for the consumer to avoid unnecessary consumption, through caps set on IHD;
- more accurate bills for customers, no estimations;
- direct communication with supplier, so no physical meter reads necessary;
- reduced consumption, and optimisation of supply benefits the environment.

Figure 1: Smart meter In-Home Display (IHD) (Ovo Energy, 2017).



Some possible negative consequences of smart meters are:

- cost of installation to the supplier;
- disposing of old meters impacts on the environment;
- cost to supplier of upgrading systems and data storage, due to large increase in magnitude of data;
- data security and privacy concerns.

1.1 CER smart meter trials

The Commission for Energy Regulation (CER) is the independent energy regulator in Ireland. Following the EU recommendations a series of trials were performed on both electric and gas smart meter usage. The aim was to provide statistical information on the impact the initiatives have on overall and peak usage for residential and small business consumers. Specifically the trials were a "collaborative energy industry-wide project managed by the CER and actively involving energy industry participants including the Sustainable Energy Authority of Ireland (SEAI), the Department of Communications, Energy and Natural Resources (DCENR), ESB Networks, Bord Gis Networks, Electric Ireland, Bord Gis Energy and other energy suppliers" (ISSDA, 2012). The trials were not run concurrently and the sample consumers were different for each, but each consisted of a benchmark period for all consumers, followed by a test period with sample groups subject to different demand side management stimuli; a control group was billed on their normal tariff and were requested to continue using energy as normal (CER, 2011).

For the residential electricity trials:

- the benchmark period ran from 1st July 2009 to 31st December 2009;
- the test period ran from 1st January 2010 to 31st December 2010.

For the residential gas trials:

- the benchmark period ran from 1st December 2009 to 31st May 2010;
- the test period ran from 1st June 2010 to 31st May 2011.

CER have made both electric and gas datasets available, with 48 meter reads per day for the sample consumers, and they can be accessed via the Irish Social Science Data Archive (www.ucd.ie/issda). Recruitment for the trial was through a voluntary "opt-in" model using a tear off slip sent to a sample of energy customers (the response rate was 30% for electricity, and 25% for gas), and a small financial incentive of 25. Participants took part in Computer Assisted Telephone Interviewing (CATI), and the questions and answers of the final samples are included in the provided CER data. In other smart meter trials it had been noted that there is a significant risk of over-reputation of more highly educated or affluent consumers. To ensure the outcome of trials were robust, respondents who opted in were profiled to ensure they were representative of the national population; and after recruitment, those consumers who accepted were compared to those who had not, to confirm for representativity.

For the gas trial, consumers who were on a prepay tariff, those who were flagged as vulnerable by the supplier, and those on short-term tenancies (less than 12 months) were excluded from the trial, which may have an impact on identifying consumers at risk of fuel poverty.

1.2 Fuel poverty

In the same European Commission Directive that calls for the roll-out of smart meters, it calls on all Member States to

> “...develop national action plans or other appropriate frameworks to tackle energy poverty” (Directives 2009/72/ EC, 2009).

The EU definition of poverty is “Persons, families and groups of persons whose resources (material, cultural and social) are so limited as to exclude them from the minimum acceptable way of life in the Member State to which they belong” (EEC, 1984); but as of 2012, of countries participating in European Fuel Poverty and Energy Efficiency (EPEE), that is Belgium, France, Italy, Spain and the UK, only the UK has an official definition of fuel poverty.

As Moore (2012) outlines, it was defined in 1991 to cover households whose fuel expenditure on all energy services exceeds 10% of their income. At the time, this is what the poorest 30% of households were spending on fuel. At the time it only considered basic income, and in 1996 was updated to use full income including Household Benefit and Income Support for Mortgage Interest. Both 1991 and 1996 definitions considered the actual fuel prices of households, collected through a fuel consumption and tariff survey. This was dropped in in 2001, and fuel cost was based on regional average prices, split by payment types, and on modeled occupancy rates. In 2005, the computation of income was refined, and fuel costs for hot water and lights were updated based on actual occupancy. The income definition differs from those commonly used in poverty statistics.

Since the 2012 Hills Fuel Poverty Review, fuel poverty in England is measured using the Low Income High Costs indicator (GOV.UK, 2017), which considers a household to be fuel poor if:

- they have required fuel costs that are above average (the national median level);
- were they to spend that amount, they would be left with a residual income below the official poverty line.

In Ireland, the 10% threshold is used with the 2005 definition, and research by the Economic and Social Research Institute estimated 19% of households may have experienced fuel poverty in 2008 (Sustainable Energy Authority of Ireland, 2009).

Whichever definition is used, the literature suggests there are three main factors that affect the level of fuel poverty:

- fuel prices
- household income
- energy efficiency of the housing stock

1.3 Review of dissertation

This dissertation examines ways of identifying households that may be at risk of fuel poverty, using only the load profile (shape of consumption) data that is collected from smart meters. The structure of the dissertation is:

- a literature review covers other research done into classifying households using load profiles, some using the CER same data;
- a data description details the data provided by CER, and how data size can be managed;
- an examination of some sampled households' energy use, and different ways to visualise load profiles;
- Principal Component Analysis as a method for identifying relationships between households and between days;
- Self-Organising Maps and k-means clustering as methods for identifying similar households;
- an analysis of the clusters with respect to household income;
- a discussion of further work which could be done, and any alternative decisions which could have been made.

2 Literature Review

2.1 Beckel *et al.* (2012)

Beckel *et al.* (2012) investigate how household properties, such as its size and the number of residents can be inferred from electricity consumption data collected by smart meters. Their focus is on characteristics which may be interesting for personalised energy consulting services, and suppliers looking to attract and retain customers.

> “A SOM is an unsupervised learning method based on neural networks that can be used to automatically extract clusters out of an otherwise unstructured (and unlabeled) set of data. A SOM is an artificial neural network that relies on unsupervised learning to group input vectors into regions of a map. Each vector is assigned to a specific region depending on its Euclidean (or other type of) distance to already mapped vectors. Clustering procedures can then be applied to group vectors within neighboring regions into clusters.”

The authors use the CER electricity trial data, but for just one week, and just load factor indicators (referred to as features) rather than the raw 48 half-hourly reads per day. They separate the features into aggregate functions of daily consumption, ratios of aggregates, temporal properties (such as spikes) and statistical properties (such as how a household’s consumption profiles correlate over subsequent days). Once calculated, the values are normalised using the unit variance scaling method (that is to scale each column to have zero mean and unit variance), so they are clustering based on pattern of consumption rather than magnitude of consumption.

Once the households had been clustered, the authors examined for each characteristic, how the households were distributed across clusters. For the employment status of the chief income earner, they found that in three clusters there were 80% in employment, whilst another three were around 30% in employment. They conclude from this that employment status is likely to be predictable from the consumption data. Also likely to be predictable are number of bedrooms, floor area, social class and number of residents. The implementation of a classification system is part of a later work by the same authors, (Becker *et al.*, 2013).

In the case of employment status, the groups that are highly likely to be in employment, are on one side of the SOM; and on this same side, the ratio of mean consumption to maximum consumption is lowest. This suggests that those in employment have a greater difference between the mean consumption and maximum consumption.

2.2 Beckel *et al.* (2013)

Using the findings from their previous paper, Beckel *et al.* (2013) present a classification system that takes the consumption features as inputs, and outputs estimated values of the household properties. They compare the performance of a variety of classification algorithms:

- k-Nearest Neighbour (kNN) - classification of an object is based on the most common class of its k neighbours in the attribute space;
- Linear Discriminant Analysis (LDA) - maximizes the ratio of between-class variance to the within-class variance;
- Mahalanobis Distance - a measure of distance between a point and the mean of a distribution;
- Support Vector Machine (SVM) - constructs a set of hyperplanes to maximise the distance between each and any training point.

To measure the performance of classification algorithms they use accuracy (ratio of true positives plus true negatives to the full population); and precision, "the probability that a sample classified as class A truly belongs to class A", and recall, "the probability of a sample being classified as class A given that the sample belongs to class A." Rather than using all features for each classification, they determine the optimal subset of features. This is a computationally complex process, so feature selection algorithms were used, which approximate an exhaustive search. Particularly, two feature selection algorithms: Sequential Forward Selection (SFS) and Parallel Sequential Forward Selection (P-SFS) were implemented. Both algorithms iteratively increase the size of the feature subset, and only the feature that maximises the performance metric at each iteration is retained.

Again, the CER electricity trial data was used, this time for the second week of January (January 11th to January 17th 2010) and four-fold cross validation was used to separate training and testing data; k-fold cross validation splits the data in to k folds, then uses each fold as the test data, and the other folds as training data across k iterations.

The authors initially choose the classification algorithm with the highest accuracy for each of the properties. For the two-class properties, the accuracy ranges from 70% (employed or not employed) to 82% (single or not single, and house age greater than 30 or less than 30 years).

When comparing the classification algorithms, they found that SVM provides the highest accuracy for 8 of the 12 household properties. SVM is able to classify properties in the presence of non-linearly separable input data, which they found was very noticeable in predicting multi-class properties, where SVM clearly outperformed the other classification algorithms. For precision and recall the classification algorithm performance were less homogeneous, which the authors partially attribute to different number of households available for each class and property. They find LDA outperforms the other algorithms when the classification requires identifying households that are outnumbered by households of other classes.

2.3 Beckel *et al.* (2014)

Beckel *et al.* (2014) build on their previous work to build a robust estimation system for household characteristics. They use an increased list of features as inputs, notably including the first 10 principal components, and transform the features to improve normality before normalising to mean zero and unit variance. In terms of classification algorithms, they also now include the AdaBoost classifier (shorthand for Adaptive Boosting, this takes outputs of other algorithms, and combines them to give a weighted sum representing a final output). To address the problems they faced with unbalanced classes in the data (for example, age of chief income earner is split 436, 2819 and 953 for less than 35, 35-65 and over 65 respectively), under-sampling the data during training is used - this randomly removes samples from the overrepresented classes. Finally, for the continuous characteristics, they used an Ordinary Least Squares regression model.

2.4 McLoughlin *et al.* (2015)

McLoughlin *et al.* (2015) investigate three widely used unsupervised clustering methods:

- k-means - partitions observations to minimize the within-cluster sum of squares (WSS) of each cluster, with the centre of each cluster the mean of observations;
- k-medoid - similar to k-means, but uses an observation as the cluster centre;
- SOMs.

They used a Davies-Bouldin (DB) validity index (a method to indicate the similarity within, and distance between clusters) to identify the most suitable clustering method and number of clusters.

They begin with the current state of load profile classes (PCs, not to be confused with Principal Components) that exist in the electricity market. In the UK there are two domestic PCs- Unrestricted, and Economy 7; and in Ireland there are four domestic PCs- 24h and Night Saver, which are each split into Urban and Rural. The classes are derived from the average for all households in the class, and are not representative of how electricity is consumed in the home.

The authors identified a gap in the literature, as nothing had been done with clustering based on the load profile for a large sample; rather it has covered clustering with features of the load profiles, or aggregations, or on small samples. The CER electricity trial data is used here, and clusters computed for each day in isolation, and then diurnal, intra daily, and seasonal patterns are characterised. As household use electricity differently, the PC they are clustered in each day will change, but the PC they occupied for the majority of time was taken as the overall PC.

The SOM method showed a consistently lower DB index over some sample days, and so was chosen for applying to the full dataset; and the optimal number of clusters was between 8 and 10, after which is when the marginal decrease in DB index became minimal. Having trained the model, two clusters were larger than the others, accounting for 28% and 37% of the population, and so were further broken down into 4 clusters each; and smaller, similar clusters were grouped together, giving 10 profile classes in total.

They found in the majority of classes that there were characteristic primary and secondary peaks which the households in each class had in common; for some the morning was the primary peak, for others the afternoon was primary; that in most classes the first peak of the day was earlier for weekdays than weekends; and that most classes had an evident shift in profile across the months, that was related to sun rise and sunset times. The authors conclude with a multi-nominal logistic regression to predict the profile class, using household characteristics and find it is possible to classify households and their patterns of electricity use based on only these characteristics.

2.5 Anderson *et al.* (2017)

Anderson *et al.* (2017) assessed the feasibility of using electricity load profiles, collected from smart meters, to impute household characteristics; and whether these characteristics could be aggregated to give small-area indicators, such as indicators found in a census (which they criticise as a costly and frequently outdated source of population statistics).

The authors used the CER electricity smart meter trial data, but used only a four week period and included only Tuesday, Wednesday and Thursday (12 days in total); this was to avoid major holidays, seasonal and temperature variations, weekends and transitional days, Monday and Friday. Initially, they used a logistic regression approach to estimate the probability that the household response person was not in paid work, based on load profile indicators (evening consumption factor and load factor) and the assumption that number of residents and the presence of children were known. A within sample validation test gave a success rate of 65%. Just using load profile indicators gave a success rate of 60%.

The authors drew on McLoughlin *et al.* (2015) to develop clusters with similar load profiles. They used the weighted least squares and k-means clustering process to give six clusters, and the cluster membership was then used in the logistic regression together with load profile indicators - the success rate of this model was 64%. They also investigated habitual behaviour by calculating the autocorrelation coefficient for the 24 hour (48 half-hours) lag. In general the coefficient was highest in the immediately following periods, then a decline until a rise to the following 24 hour lag. They expected to see less regular habits for those out of work, but adding a 24 hour and 48 hour lag to the logistic model only gave marginal improvement, and only the 24 hour lag was statistically significant.

The load factor indicators used were:

- Evening Consumption Factor (ECF) - mean load during 4pm to 8pm, relative to the mean load at all other times in the day;

- Load Factor (LF) - ratio of mean load for the whole day to the maximum load for the whole day.

2.6 Chelmis *et al.* (2015)

Chelmis *et al.* (2015) aim to identify customers who would be suitable for demand response programs (price incentives for individuals to curtail demand when overall demand is high) using Principal Component Analysis (PCA) on 15 minute consumption data.

The authors discuss how they store the data in different representations to study different elements of the consumption:

- To study daily pattern per building, they store a matrix for each customer of dimension 365×96 -365 days in the sample and 96 periods.

- To study variations in demand for each day of the week, over time for each building, and to identify similarities for each building, they store a matrix for each day of the week (Monday-Sunday) which is $(52\times N)\times 96$ 52 instances of each day in the sample, N buildings, and 96 periods.

- To study coarse-grained similarities between buildings, and to statistically understand how each buildings demand changes through the week, they store a matrix for each day of the week with dimensions $N \times 96$.

Due to the magnitude of the data, which covers 5 years, 115 buildings, with 96 reads per day, the authors argue that directly applying clustering algorithms here would be inefficient and so propose first reducing the dimensionality, through PCA.

When plotting the first two primary components for the first set of matrices, for each building they found that there was a distinct separation between weekdays and weekends, but not between individual weekday. The buildings they looked at were on a University campus, so this observation makes intuitive sense.

When plotting the first two primary components for the second set of matrices, for each day of the week they found that each individual building's consumption values grouped closely, suggesting that energy needs remained similar across different weeks; and, they also found that building similar in nature (such as office blocks, or classrooms) clustered together.

Whilst the paper presents two dimensional plots of the first two primary components, the authors found that four dimensions described 96% of the data and the implicit patterns in it, which is a compression ratio of 4/96. With this reduced data they look at ways to cluster the buildings, using k-means, hierarchical, and k-medoids clustering.

2.7 Harold *et al.* (2015)

Harold *et al.* (2015) examine the determinants of residential gas demand using the CER gas trial data. Theirs is the only paper that uses the gas trial data, but they do not look at the shape of daily gas profiles, rather the overall daily consumptions.

In their data exploration they note that a large proportion of households have zero consumption each day, which as they are looking at over a calendar year, is attributed to the high seasonality of gas - many households will not have any gas use other than central heating, which is not used for a large part of the year. They also found that a similarly large proportion had consumption in every period of the day; this could be explained by pilot lights installed on some gas fires and central heating systems that burn continuously, and therefore results in a small gas consumption being recorded for every half hour period. In between these two extremes, they found that the frequency of the number of periods (half hours) with consumption increased from one to six with a distribution skewed to the right, then decreasing to 47 periods.

They do not include the price of gas in their analysis, as there was little variation over the sample period, and all households are subject to the same tariff, except for one sub-group subject to a variable tariff in the testing phase of the trial. Income is also not considered due to a low response rate but other socio-economic factors, namely education and employment status, which are highly correlated to income, are considered.

The geographical location of the households in the trial is not known, so in choosing where to take weather reading from, they found that Dublin has both the highest population density in Ireland, and the highest concentration of household gas meters in the country. The average daily temperature is about 7.5C across 539 days, and an outdoor temperature of 15.5C is considered the critical level above which temperature does not have an effect on heating requirements. They note a particularly cold spell over Christmas 2010 where an outdoor temperature of -7.5C was recorded. They also take into account cloud cover, sunshine hours, rainfall, and wind speeds.

The main model assumed daily gas consumption depends on socio-economic factors; the number of residents; the size, type and age of the dwelling; the ownership status; the weather, and seasons; and features of the dwelling that may increase energy efficiency. A range of models of increasing complexity were tested, with the most complex explaining almost 60% of the variation in daily gas consumption.

They expected that higher education levels would signal an increased awareness of energy efficiency concerns, possibly showing lower consumption, but the model suggested the opposite, with lower education households consuming less. This could be attributed to the income effect, with higher education comes higher income. The model also suggested that the self-employed and retired consume more than the employed, themselves in turn consuming more than the unemployed.

Overall, weather was found to be the most influential factor in determining daily gas consumption.

2.8 Summary

There were no specific objectives in the literature reviewed that addressed fuel poverty but employment status was featured often. Self-Organising Maps were used in two instances as a method for clustering households, with McLoughlin *et al.* (2015) finding it more successful than other classification algorithms when working with the 48-period load profile data. Chelmis *et al.* (2015) used Principal Component Analysis and several different arrangements of their data (96-period load profiles) to identify building types, and different behaviour on days of the week.

McLoughlin *et al.* (2015) and Chelmis *et al.* (2015) were the only two papers to use the the load profile data without reduction; Beckel *et al.* (2012), their subsequent works, and Anderson *et al.* (2017) reduced the load profiles to a series of features, such as aggregates and ratios before the main analysis.

Harold *et al.* (2015) was the only reviewed literature that analysed gas consumption, as opposed to electricity consumption. They investigated the determinants of daily consumption, disregarding the more granular period data.

3 Data Description

This section explains the data formats provided by CER, and the processing steps taken before analysis. Both electric and gas datasets were provided by CER for this analysis. The format of the files was similar for the two fuel types, but detailed separately in the following subsections for clarity.

3.1 Electricity data provision

Smart meter read data:
Six zipped files named `File1.txt.zip` to `File6.txt.zip` each containing one text file, each with three columns corresponding to:

- Meter ID
- Five digit code:
  - Day code: digits 1-3 (day 1 = 1st January 2009)
  - Time code: digits 4-5 (1-48 for each 30 minutes with 1= 00:00:00 00:29:59)
- Electricity consumed during 30 minute interval (in kWh)

Pre trial residential survey - questions:
One Word file named `RESIDENTIAL PRE TRIAL SURVEY` containing CATI coded survey
Pre trial residential survey - answers:
One Excel and one `.csv` file, each named `Smart meters Residential pre-trial survey data`. The first row has the question number and summary; and subsequent rows are the answers per respondent:

- Meter ID
- one column per survey question

3.2 Gas data provision

Smart meter read data:
78 `.csv` files named `GasDataWeek0` to `GasDataWeek77`, each with 3 columns corresponding to:

- **Meter ID**
- **Five digit code**:

  - Day code: digits 1-3 (day 1 = 1st January 2009)
  - Time code: digits 4-5 (1-48 for each 30 minutes with 1= 00:00:00 00:29:59)

- **Gas consumed during 30 minute interval (in kWh)**

Pre trial residential survey - questions:
One Word file named `RESIDENTIAL PRE TRIAL SURVEY - GAS` containing CATI coded survey **Pre trial residential survey - answers:**
One Excel and one `.csv` file, each named: `Smart meters Residential pre-trial survey data - Gas`. The first row has the question number and summary; and subsequent rows are the answers per respondent:

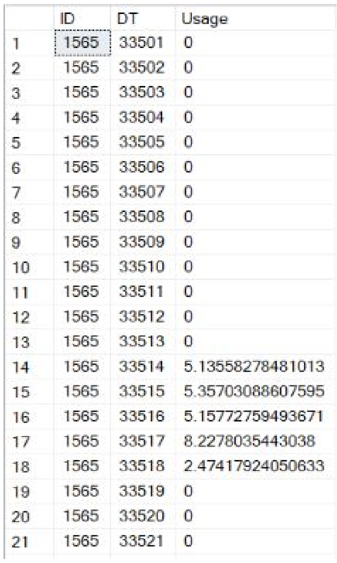
- Meter ID
- 1 column per survey question

3.3 Data size

Previous work using the electricity data, have not used the full data. The main reason for this is computational and resource constraints. Anderson *et al.* (2017) looked at only 3 days in a week over 12 week; Beckel *et al.* (2012) reduce the the 48 reads per day to a set of features, over one week; and, McLoughlin *et al.* (2015) use the full data but perform clustering on each day individually, then comparing the clustered data. Chelmis *et al.* (2015) use a different dataset (which covers 5 years, 115 buildings and 96 reads per day), and aim to reduce the dimensionality using PCA before clustering. Analysis of the gas data by Harold *et al.* (2015) was aggregated up to daily consumption, so size was not a problem.

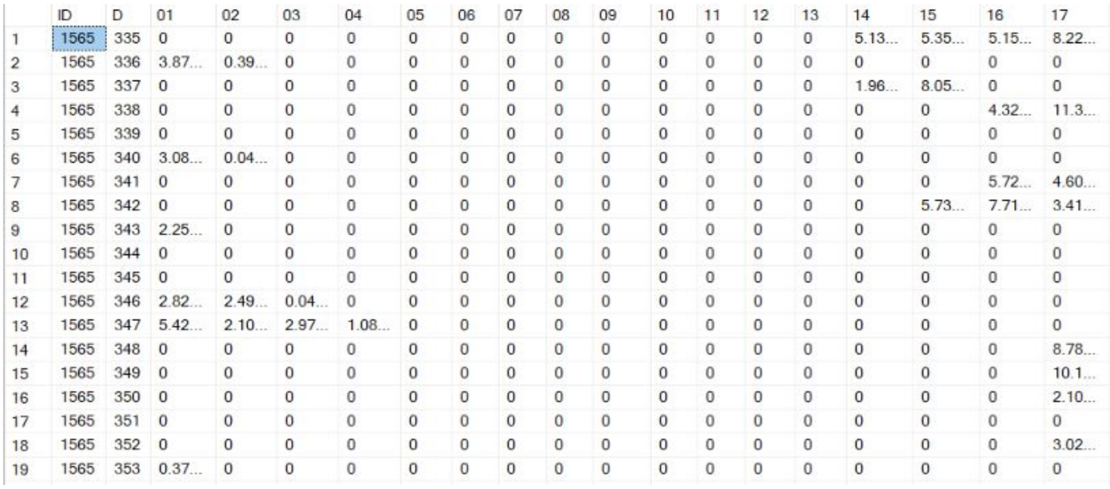
This dissertation only considers residential consumers, and to avoid any bias from the different demand side stimuli applied in the test period, it only considers the benchmark period of the trial. Table 1 shows the number of records in the source data and the resulting records once only residential customers in the benchmark period are considered. A sample of the gas data can be seen in figure 2; there is one line per consumer-date-period combination, but for some of the analysis a transposed version of the data, as in figure 3, was used.

Figure 2: Screen capture of gas data in SQL; there is one record per consumer-date-period combination.



	ID	DT	Usage
1	1565	33501	0
2	1565	33502	0
3	1565	33503	0
4	1565	33504	0
5	1565	33505	0
6	1565	33506	0
7	1565	33507	0
8	1565	33508	0
9	1565	33509	0
10	1565	33510	0
11	1565	33511	0
12	1565	33512	0
13	1565	33513	0
14	1565	33514	5.13558278481013
15	1565	33515	5.35703088607595
16	1565	33516	5.15772759493671
17	1565	33517	8.2278035443038
18	1565	33518	2.47417924050633
19	1565	33519	0
20	1565	33520	0
21	1565	33521	0

Figure 3: Screen capture of gas data in SQL after transposition; there is one record per consumer-date combination.



	ID	D	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17
1	1565	335	0	0	0	0	0	0	0	0	0	0	0	0	0	5.13...	5.35...	5.15...	8.22...
2	1565	336	3.87...	0.39...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1565	337	0	0	0	0	0	0	0	0	0	0	0	0	0	1.96...	8.05...	0	0
4	1565	338	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.32...	11.3...
5	1565	339	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1565	340	3.08...	0.04...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1565	341	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.72...	4.60...
8	1565	342	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.73...	7.71...	3.41...
9	1565	343	2.25...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1565	344	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1565	345	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1565	346	2.82...	2.49...	0.04...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1565	347	5.42...	2.10...	2.97...	1.08...	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1565	348	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8.78...
15	1565	349	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10.1...
16	1565	350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.10...
17	1565	351	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1565	352	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.02...
19	1565	353	0.37...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The individual read data files (8 electricity files, and 78 gas files) were concatenated using the batch command `copy GasData* GasData` (example given for gas only), and then imported to SQL Server 2014 for defining the date and time variables the (five digit code was extrapolated to give a date, day of week and time. For example, 33606 is 2nd December 2009, Wednesday, 02:30:00-02:59:59), reducing the data to relevant

Table 1: Counts of raw data and reduced data.

	Electricity	Gas
No. of records	157,992,997	38,698,559
No. of consumers	6,436	1,494
No. of residential consumers	4,232	1,365
No. of records - residential consumers in the benchmark period only	34,572,206	11,321,856
No. of individual days	720,254	235,872

consumers, and transposing the data. The SQL syntax used to define date and time is:

```
--split 5 digit code DT into date (day number) and time (period)
select ID
, left(DT,3) as D
, right(DT,2) as T
, Usage
into tmpReads3 from tmpReads2
--from the day number, define the date based on day 1 being 01/01/2009,
--and the day of week
select ID
, dateadd(day,cast(D as int),cast('2008-12-31' as date)) as 'Date'
, datename(dw,dateadd(day,cast(D as int),cast('2008-12-31' as date))) as 'Day'
, T
, Usage
into tmpReads5 from tmpReads3
```

4 Gas load profiles

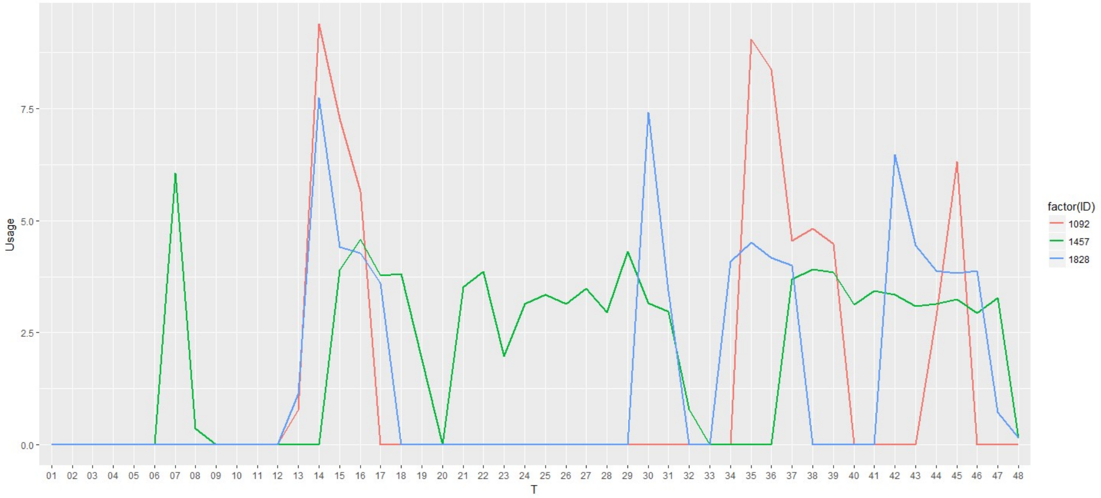
Prior research has mainly focused on consumer behaviour relating to electricity consumption, with only Harold *et al.* (2015) covering the CER gas trial. For this reason, and because the majority of households have gas central heating, this dissertation focuses on gas consumption. This section gives examples of some individual households' gas load profiles, how these differ between household and across days, and explores ways to visualise the energy consumption.

To demonstrate the how the shape of gas load profiles can differ between consumers, three sample households were chosen, each with two residents. The consumers chosen have the respondent to the survey being employed, unemployed and actively seeking work, and retired.

4.1 Wednesday 10th March 2010

In this first representation, the load profiles are each taken from Wednesday 10th March 2010. The three lines in figure 4 show employed (red), unemployed (blue), and retired (green) households.

Figure 4: Load profiles of three households on Wednesday 10th March 2010; employed (red), retired (green), and unemployed (blue).



The employed household (red) uses no gas from midnight until period 12 (05:30-06:00[1]) where it ramps up to a daily peak of 9.39kWh consumed in period 14 (06:30-

[1] The periods are inclusive of the 59th second before the next period, in this case 05:30:00-05:59:59. Here and in similar cases, it is referred to as 05:30-06:00 for easier comprehension.

07:00); this then reduces over the next two periods, to zero in period 17 (08:00-08:30). This is likely to be heating coming on by timer to heat the house before the residents wake up, then leave the house.

Consumption is zero through the day until period 35 (17:00 -17:30, where 9.05kWh is consumed; consumption continues of the next four periods, to zero in periods 40 to 43 (19:30 - 21:30), followed by two periods of consumption, and then zero again. This profile could be to central heating coming on by timer before the residents come home from work, or the residents turning it on manually; once the house is warm, the central heating is turned off until later when the house may have started to cool.

The unemployed household (blue) has a similar profile for the morning period, with a daily peak (7.74kWh) in period 14, but remains consuming for an extra period. So is also likely to be heating coming on by timer to heat the house before the residents wake up, then leave the house.

The afternoon profile is different, with earlier consumption, in periods 30 and 31 (14:30-15:30) up to 7.41kWh; and then a period of lower, more stable, consumption in periods 34 to 37 (16:30-18:30) around 4kWh to 4.5kWh. This may suggest the at least one resident in the household comes home in the mid-afternoon, when the heating is put on. Another spike comes in period 42 (20:30-21:00, followed by a lower, stable consumption period from 43 to 46.

The load profile of the retired household (green) is markedly different in that the peaks are lower, but there are more periods with consumption across the day. This could be due to the residents being in the house for most of the day so the house is never left to cool down, or there is a thermostat keeping a constant temperature, rather than a timer set for certain times.

The daily peak comes in period 7 (03:00-03:30) at with consumption of 6.05kWh. From period 15 to 47 (07:00 to 23:30) there are only 6 periods (3 hours) without some consumption.

Table 2: Statistics for three sampled households on Wednesday 10th March 2010.

Statistic	Employed	Unemployed	Retired
Mean (kWh)	1.324	1.503	2.047
Variance (kWh^2)	7.364	5.159	3.152
Periods on	11	18	31
Mean when on (kWh)	5.776	4.009	3.170
Variance when on (kWh^2)	6.423	5.873	2.918
Maximum (kWh)	9.389	7.741	6.052
Period of maximum	14	14	7
Time of maximum	06:30-07:00	06:30-07:00	03:00-03:30
Total consumption (kWh)	63.531	72.167	98.258

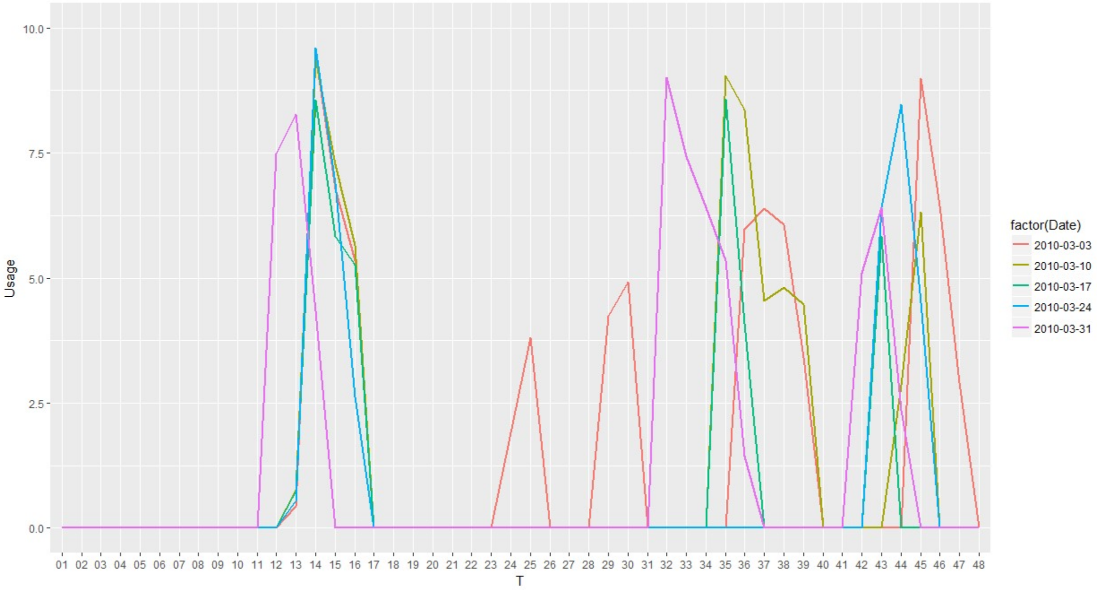
4.2 Wednesdays in March 2010

Figures 5, 6 and 7 show in turn how each household is consuming gas across the month of March 2010, looking at just Wednesdays. From these plots it becomes evident how consistently each household uses gas.

Example R script for sample household 1092:

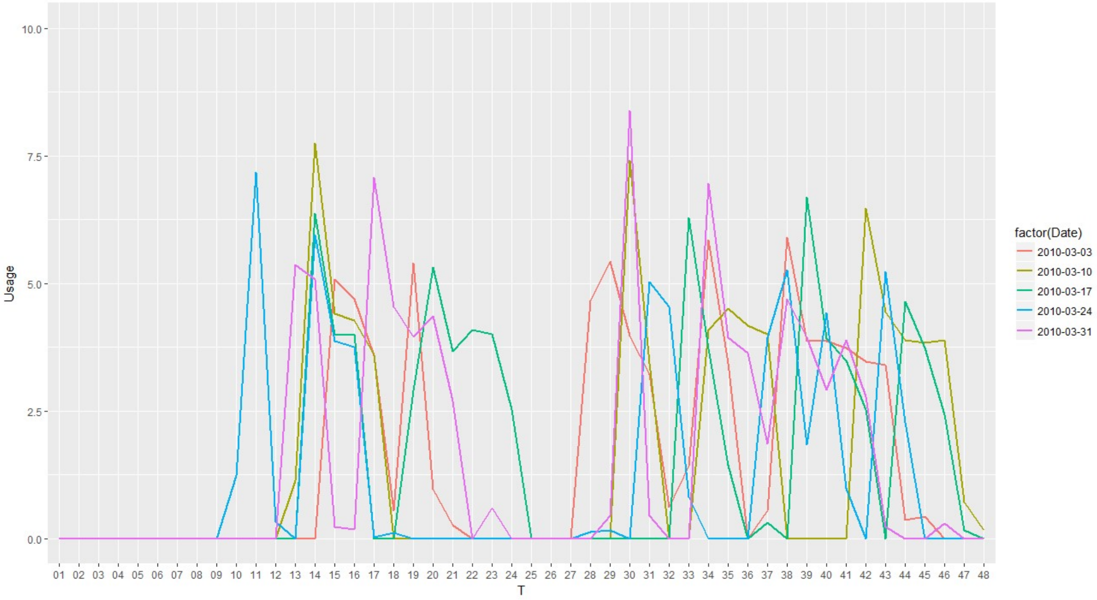
```
ggplot(data=GasLongWeds[GasLongWeds$ID=="1092"
& GasLongWeds$Date>’2010-02-28’
& GasLongWeds$Date<’2010-04-01’ ,]
,aes(x=T,y=Usage,group=Date,colour=factor(Date)))
+geom_line(size=0.8)+ylim(0,10)
```

Figure 5: Load profiles of sampled employed household on Wednesdays in March 2010.



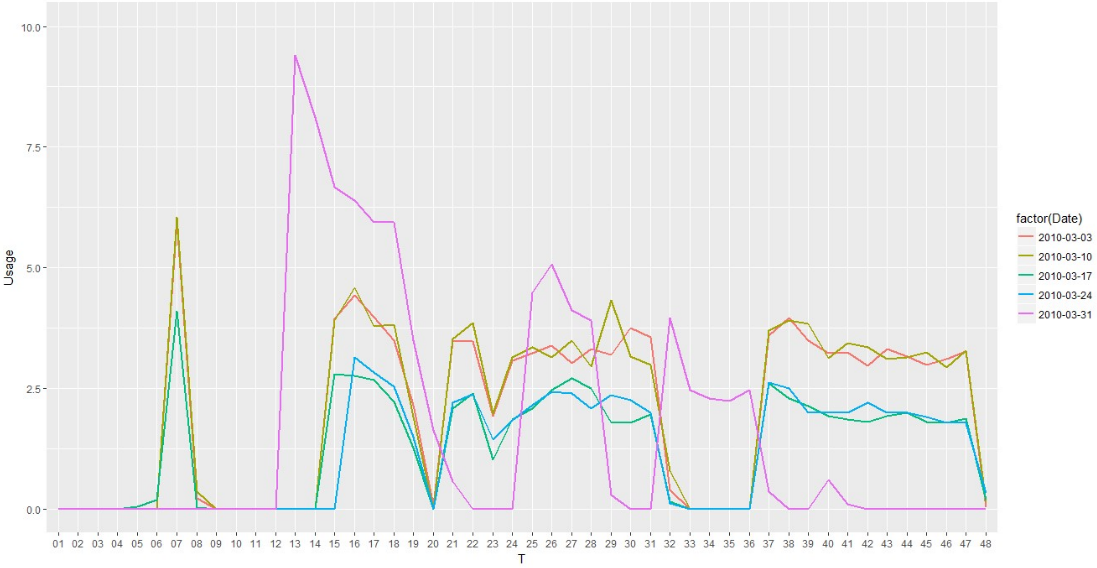
For the employed household, figure 5, in four of the five days, the gas comes on in period 13 (06:00-06:30), and goes off by period 17 (08:00); but, in the last Wednesday of the month, 31st March, these movements happen two periods (one hour) earlier. A possible explanation for this would be that the central heating timer had not been updated since the change to British Summer Time, on 28th March. After the morning period there is less distinct of a pattern, with 3rd March having gas use through the day, and the evening gas use not having a consistent start time. Two evening spikes are visible each week, but with different on and off times.

Figure 6: Load profiles of sampled unemployed household on Wednesdays in March 2010.



For the unemployed household, figure 6, the earliest gas use of the day varies from as early as period 10 (04:30-05:00) on 17th March, to as late as period 15 (07:00-07:30) on 3rd March; and on four of the five days there are two period of usage - where the half hourly consumption becomes zero, or near zero, to be followed by increased consumption. The afternoon and evening usage is also different week to week, and there is no evidence of the change to BST as in the employed household. The different load profile here could be due to a thermostat turning the central heating on and off to keep a constant temperature, or could be due to the residents not having a consistent daily routine.

Figure 7: Load profiles of sampled retired household on Wednesdays in March 2010.



The retired household, figure 7, shows three distinct load profile shapes. In the first two weeks of the month the profiles have the same on and off times and very similar magnitudes. In the second two weeks, the on and off times are largely the same as the first two weeks, but with a lesser magnitude. This could be due to warmer weather, meaning less heating needed, or due to the residents reducing the thermostat temperature. The last Wednesday, after the move to BST does not have the spike in period 7 (03:00-03:30), the gas is not used until period 13 (06:00-06:30). This is one hour earlier than other weeks, potentially due to the hour change, but the consumption through the day is notably different to the other weeks, not just offset by one hour.

4.3 Auto-correlation

Rather than just using a visual interpretation of the pattern week-to-week, the mathematical approach is to use autocorrelation to identify repeating patterns. Using the same three sample households, and all Wednesdays in March, the `acf` function is R was used to identify any habitual behaviour. Autocorrelation is the correlation of a time series with itself, which if stationary is expressed:

$$ACF = \frac{E[(x_t - \mu)(x_{t+k} - \mu)]}{\sigma^2}, \tag{1}$$

where k is the lag, and $x = x_1, x_2, ...$

Anderson *et al.* (2017) calculated the autocorrelation coefficient as part of their model to determine household characteristics from electricity load profiles, and expected lower autocorrelation for those not in paid work (derived from unpublished analysis). The actuality was that this only produced a marginal improvement to their model, and only at the 24 hour lag.

Figure 8 shows the autocorrelation for the employed household, figure 9 for the unemployed household and figure 10 for the retired household. In each plot the highest positive coefficient is in the half-hour period immediately following, then declines to a become a negative coefficient. This may be because following a period of time of high gas consumption, usually for heating a house, there is no need for gas consumption in the following periods as the house hold the heat. Also, the plots in subsection 4.2 showed distinct morning and evening periods of consumption for the employed and unemployed households, which were followed by no consumption due to a vacant house or night time (when residents are asleep). The retired household did not have these distinct periods, and may be why the negative autocorrelation coefficient is less.

For all the sampled households there are other high positive coefficients at 48 periods (24 hours, the same time of day). For the employed household this high positive coefficient is repeated at 96 periods, 144 periods - so for four consecutive Wednesdays they displayed very habitual behaviour. For the unemployed and retired households there are still positive coefficients at these points, but of a lesser magnitude. So for these sampled households, it agrees with the Anderson *et al.* (2017) expectation that those not in paid work will show less habitual behaviour.

Figure 8: Auto-correlation of sample employed household on Wednesdays in March 2010.

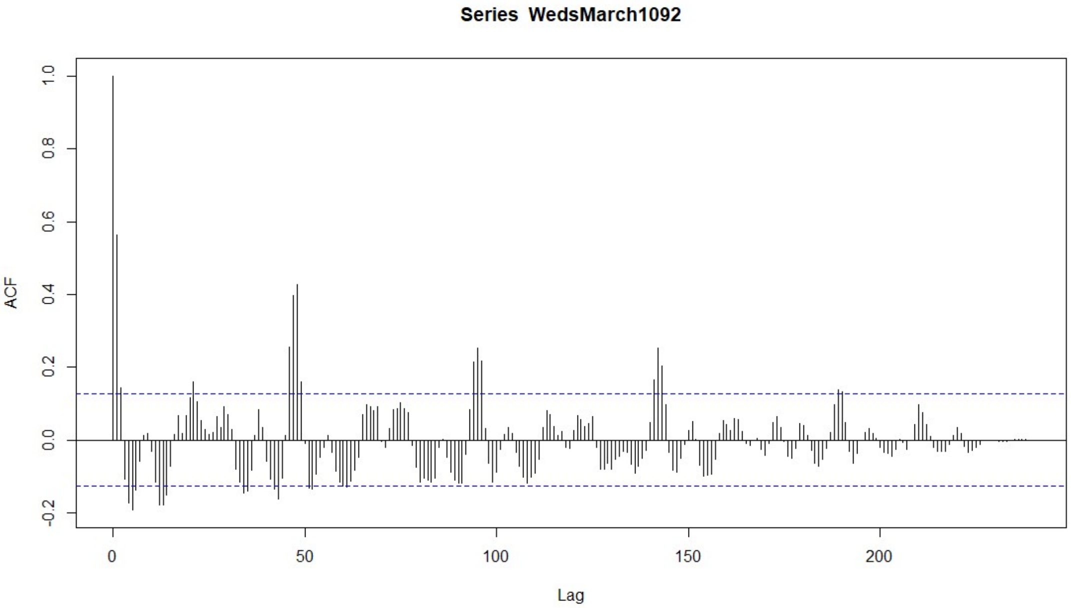


Figure 9: Auto-correlation of sample unemployed household on Wednesdays in March 2010.

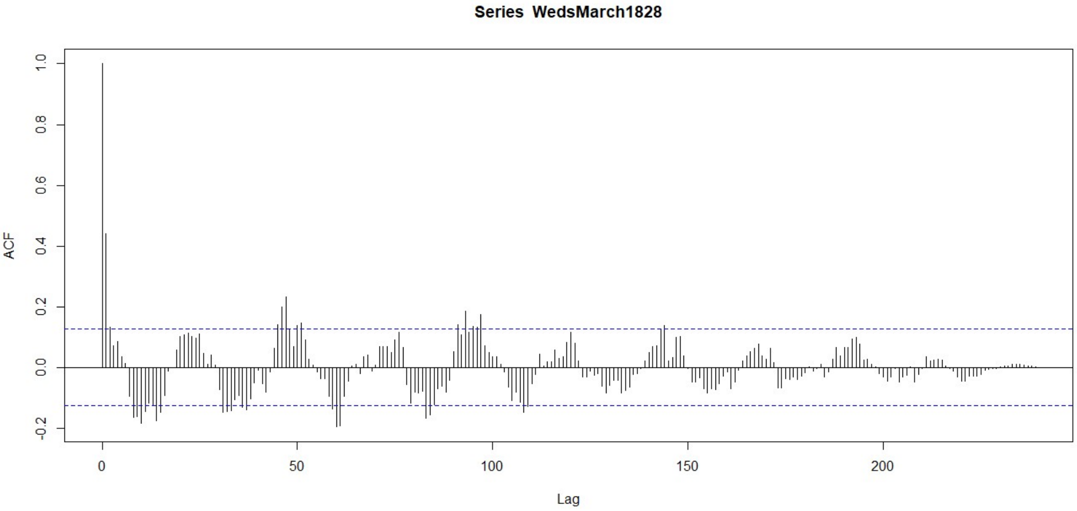
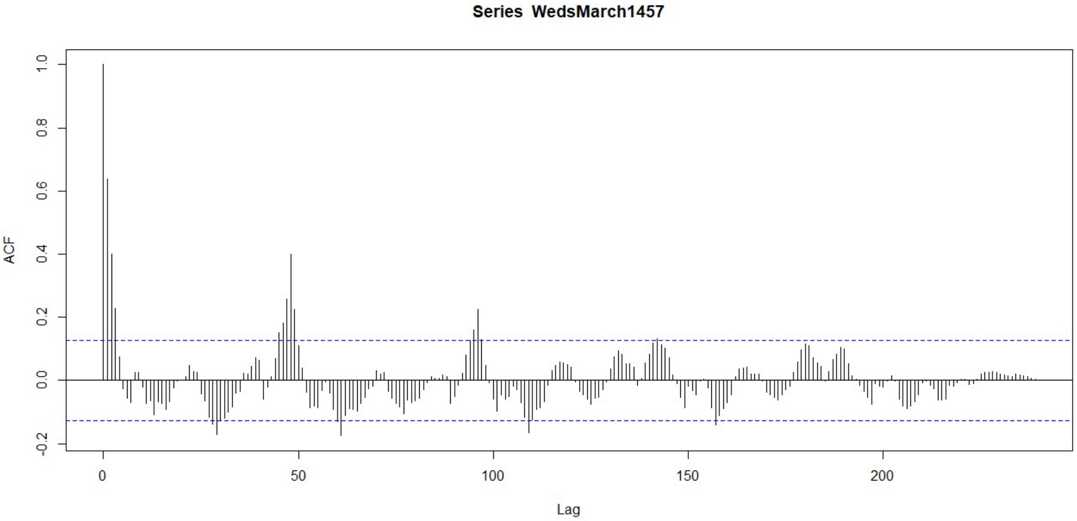


Figure 10: Auto-correlation of sample retired household on Wednesdays in March 2010.



4.4 Visualisation

This section explores some other ways to visualise the data, countering some of the weaknesses of plots such as figures 5, 6 and 7. With those plots it is difficult to display more than the 5 days, as lines begin to overlap and stack up on top of each other; there is also a difficulty in identifying differences in colour when the number of lines increases.

4.4.1 3D surface plots

The first type of plot to consider is a 3D surface over the $x-y$ plane. It is created in R using the `persp` function, where the x-axis displays the date (in days since 1900-01-01 format) and the y-axis displays the 48 periods. The figures 12, 11 and 13, are each showing all Wednesdays in the benchmark period - so covering 1st December 2009 to 31st May 2010.

The R code for this is:

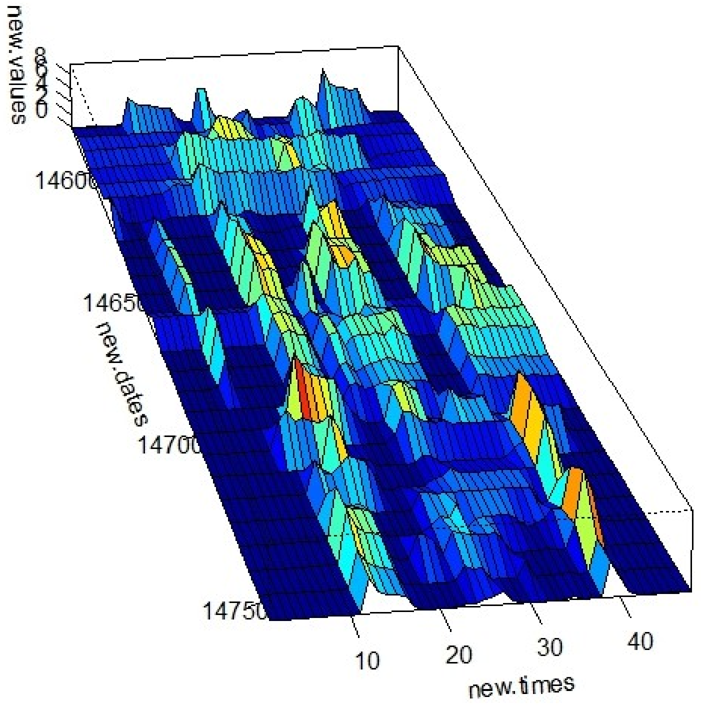
```
dat2=GasLongWeds[GasLongWeds$ID==1828,c(1:2,4:5)]

new.dates <- with(dat2, sort(unique(Date)))
new.times <- with(dat2, sort(unique(as.numeric(T))))
new.values <- with(dat2, matrix(Usage, nrow = 26, ncol = 48
, byrow = TRUE))
new.col=drapecol(new.values, col = femmecol(100), NAcol = "white"
, lim = NULL)

persp(new.dates,new.times, new.values, ticktype = "detailed"
,  r = 5,theta = 80, phi=20, scale = FALSE,col=new.col)
```

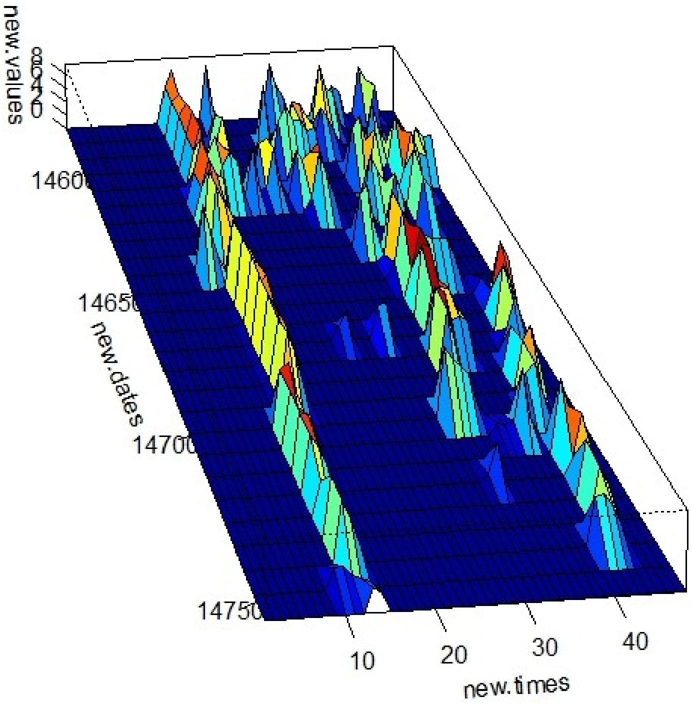
In each of the plots, towards the top of the plot, there are days where the gas is in use through the majority day time periods, and a generally different pattern to the rest of the days; this is due to the Christmas and New Year period, were people will alter their routine due to different working hours (both affecting those working and those using services), and potentially having more people in the house. This is particularly evident for the retired household, who usually have a very similar pattern of use week-to-week, but over at least two Wednesdays in December and January have gas use almost constant through the day.

Figure 11: 3D plot of sampled retired household's consumption on Wednesdays.



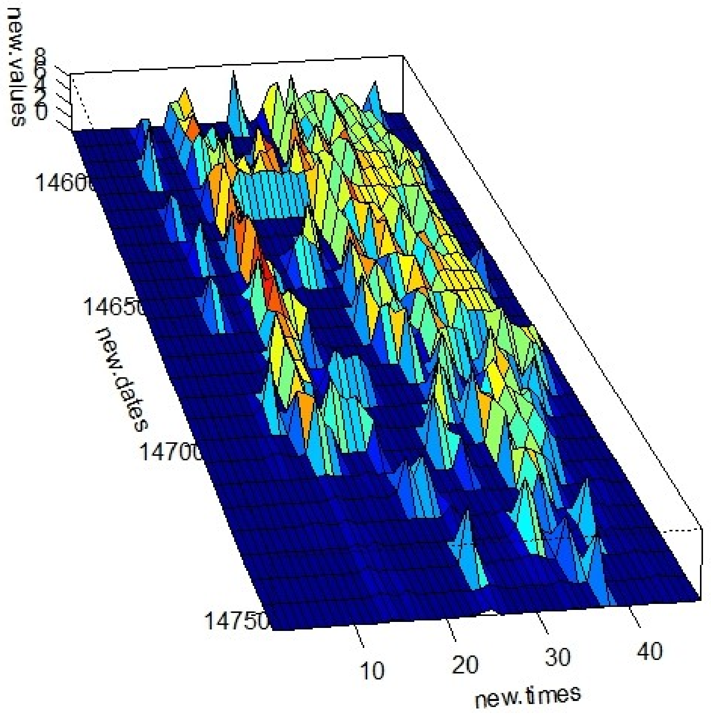
The retired household shows the effect that warmer weather can have on gas consumption. Up to the end of March there is a period overnight where it is assumed the heating comes on to warm the house. As seen in figure 7 this pattern was different in the last week in March, and this is confirmed in figure 11 to be a change for the rest of the sample period. The other change for the retired household is the evening consumption changes to be a shorter period of consumption but at a higher rate.

Figure 12: 3D plot of sampled employed household's consumption on Wednesdays.



The employed household, figure 12, continues with the earlier morning consumption (seen in figure 5), which was previously attributed to the central heating timer not being changed following the BST hour change; but as this behaviour continues, it may instead be based on temperature, or a change in routine. The afternoon and evening consumption tails off towards the end of the sample, with two spikes of consumption reducing to one, and eventually none (though as this is the end of the sample, it could be an anomaly, such as the house is vacant).

Figure 13: 3D plot of sampled unemployed household's consumption on Wednesdays.



The unemployed household in figure 13 probably shows the most obvious effect of temperature and the change from Winter to Spring, with morning gas consumption ceasing around mid-April, and afternoon and evening consumption reducing from a similar time.

4.4.2 2D overhead plots

The 3D plot is still limited in the number of days it can display, with a loss of clarity towards the back of the plot. To overcome this figures 14, 15 and 16 look directly from above. This is using the R package `plot3D` and the function `image2D`.

```
library(plot3D)


tmp.matr=GasWide[GasWide$ID==1828,]
tmp.matr=tmp.matr[order(tmp.matr$Date),]
tmp.matr=tmp.matr[,4:51]
tmp.matr=data.matrix(tmp.matr, rownames.force = NA)

image2D(tmp.matr)
```

Figure 14: 2D plot of sampled employed household's consumption on all days.

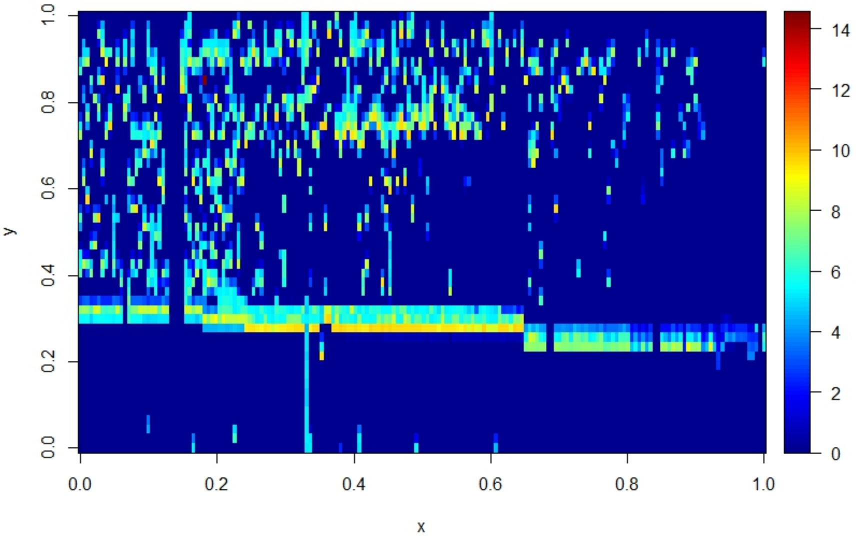


Figure 15: 2D plot of sampled retired household's consumption on all days.

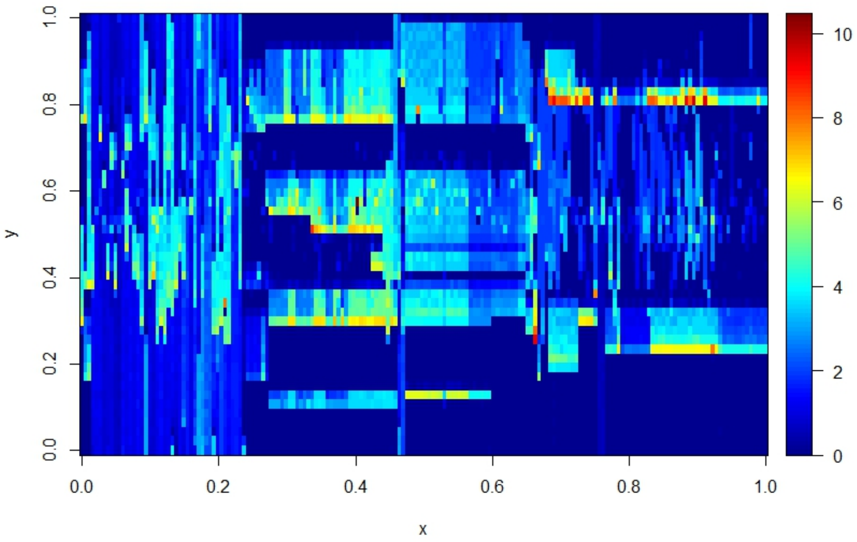
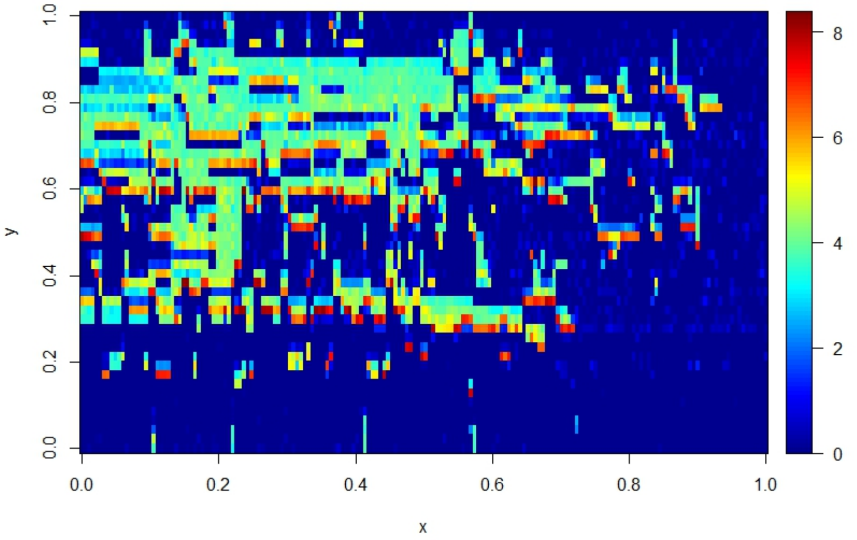


Figure 16: 2D plot of sampled unemployed household's consumption on all days.



In figures 14, 15 and 16 the x-axis displays all days (not limited to just Wednesdays) in the benchmark period, and the y-axis displays the 48 periods. By including the extra days of the week, there is no evident change in overall pattern for each of the households. They are still dominated by certain periods of consumption in the morning and evening, but other insights become apparent.

- Vacant days are visible for the retired household and for the employed household;
- Gas consumption into the early hours of the morning for the employed household and unemployed household;
- A period of months with very consistent on and off times for the retired household.

5 Principal Component Analysis

This sections looks at using Principal Component Analysis (PCA) to better understand the the gas data. The data has 48 variables, and as seen in subsection 4.3, there is some correlation between them. Chelmis *et al.* (2015) use PCA to reduce the dimensionality of the electricity data they are working with, and Beckel *et al.* (2014) use PCA to help describe different behaviour between households.

PCA can be used to find the important relationships between variables, so that all variables are not needed to understand the data; the aim is to capture a sufficient amount of information with as few variables as possible. It does this reduction by identifying the direction along which variations in the data is maximal - each of these directional linear combinations is a principal component. The first principal component shows the largest variation; the second principal component is the largest variation that is uncorrelated to the first, and so on. The outcome is a set of values of linearly uncorrelated variables, instead of the original possibly correlated variables. PCA is a benefit for computational resources, not needing to hold and process as much data; and, for interpretation and visualisation, it is easier for people to comprehend and understand lower dimensional data.

The ith principal component, Z_i is defined as

$$Z_i = \Phi_{i1}X_1 + \Phi_{i2}X_2 + \cdots + \Phi_{iK}X_K, \tag{2}$$

where Φ_{iK} is the loading vector, or eigenvector, of the ith principal component, and normalised variable X_K. To calculate the principal components the variables need to be normalised. With the gas load profiles, periods are all measured in kWh, but there will be certain periods which have higher variance, and this will cause a dependence of principal components on these variables. In other cases the units for each variable will be different, and so normalising becomes more important.

5.1 Relationship between households

In this subsection the focus is to investigate the variance between households on one day. The fist step is to organise the sample into a data matrix, of N rows and K columns, and in this instance the matrix is 1296×48. This is all 1296 households with 48 periods each for Wednesday 10th March 2010 (the same day used in subsection 4.1), similar to the third matrix type that Chelmis *et al.* (2015) used; they specified a matrix like this for each day of their sample, but here it is just one matrix for one day.

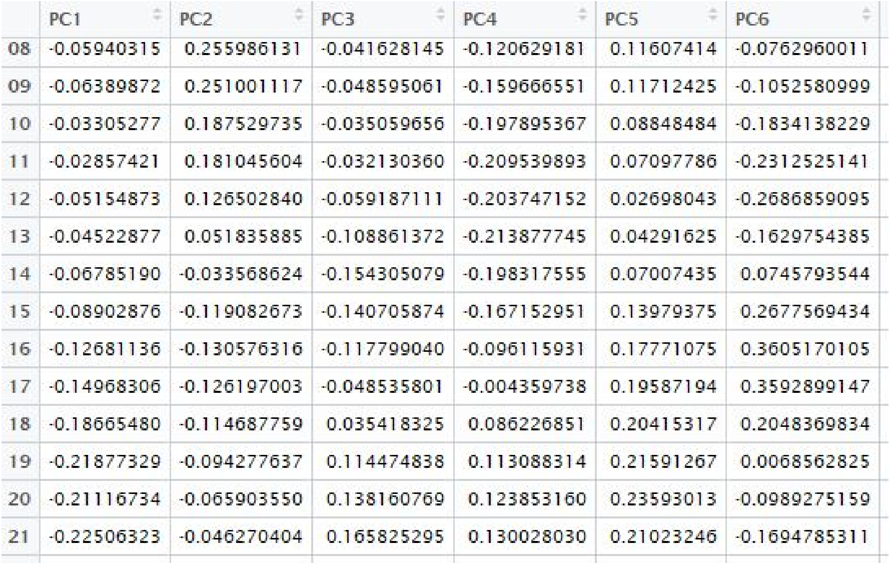
The `prcomp` function is used in R, and by specifying `center=TRUE` and `scale.=TRUE`, the variables are centred to zero mean, and scaled to unit variance. The scaled vector x' satisfies

$$x' = \frac{x - \overline{x}}{\sigma}, \tag{3}$$

where x is the original vector of variables, $\overline{x}$ is the mean of it, and σ is the standard deviation of it.

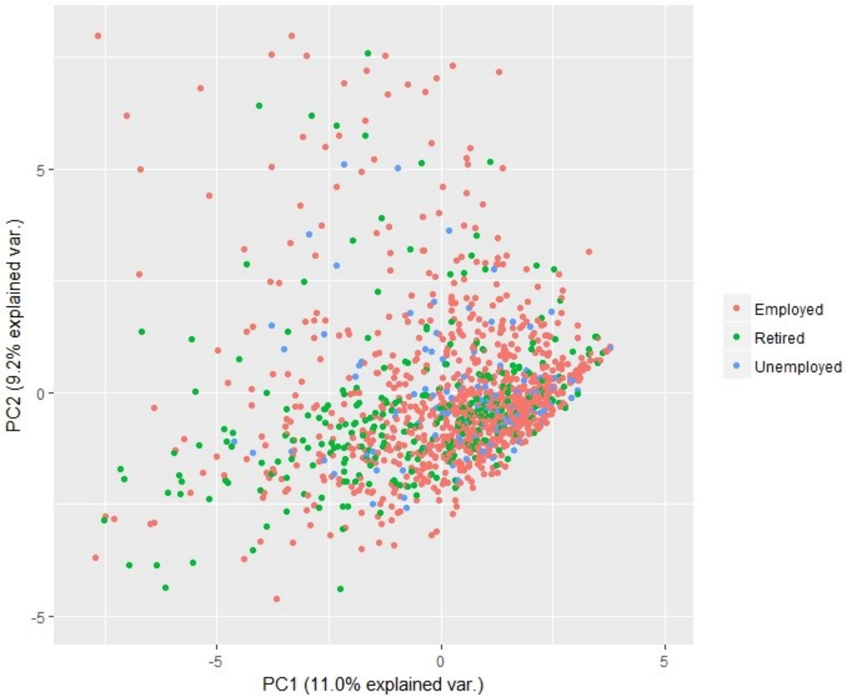
The principal component loadings for Φ_{iK} for $i = 1, \ldots, 6$ and $K = 8, \ldots, 21$ are shown in figure 17. By multiplying the normalised variables by the loadings, the result can be plotted and the direction of the principal components can be seen, see figure 18 for principal components one and two (PC1 and PC2). PCA is often used to identify groups of observations, such as the employment status of the household, but in this case just PC1 and PC2 alone do not provide enough information to do so; also labeled in the figure is the percentage of variance explained by each principal component, and together they explain 20.2%.

Figure 17: Principal component loadings for Φ_{iK} for $i = 1, \ldots, 6$ and $K = 8, \ldots, 21$. Screen shot from R - all households on one day.



	PC1	PC2	PC3	PC4	PC5	PC6
08	-0.05940315	0.255986131	-0.041628145	-0.120629181	0.11607414	-0.0762960011
09	-0.06389872	0.251001117	-0.048595061	-0.159666551	0.11712425	-0.1052580999
10	-0.03305277	0.187529735	-0.035059656	-0.197895367	0.08848484	-0.1834138229
11	-0.02857421	0.181045604	-0.032130360	-0.209539893	0.07097786	-0.2312525141
12	-0.05154873	0.126502840	-0.059187111	-0.203747152	0.02698043	-0.2686859095
13	-0.04522877	0.051835885	-0.108861372	-0.213877745	0.04291625	-0.1629754385
14	-0.06785190	-0.033568624	-0.154305079	-0.198317555	0.07007435	0.0745793544
15	-0.08902876	-0.119082673	-0.140705874	-0.167152951	0.13979375	0.2677569434
16	-0.12681136	-0.130576316	-0.117799040	-0.096115931	0.17771075	0.3605170105
17	-0.14968306	-0.126197003	-0.048535801	-0.004359738	0.19587194	0.3592899147
18	-0.18665480	-0.114687759	0.035418325	0.086226851	0.20415317	0.2048369834
19	-0.21877329	-0.094277637	0.114474838	0.113088314	0.21591267	0.0068562825
20	-0.21116734	-0.065903550	0.138160769	0.123853160	0.23593013	-0.0989275159
21	-0.22506323	-0.046270404	0.165825295	0.130028030	0.21023246	-0.1694785311

Figure 18: Principal components one and two, coloured by household employment status - all households on one day.



As stated previously, the first principal component explains the most variation in the data, with each subsequent principal component explaining less. Mardia *et al.* (1979, pp224-225) give some "rules of thumb" for choosing a suitable number of principal components, and what level of explained variance is appropriate:

- exclude principal components whose eigenvalues are less than the mean - eigenvalues measure the variability retained by each principal component
  - suggests first 15 principal components to keep
- examine a "scree graph" which shows the variance explained by each principal component, which often shows a distinct cut-off between large and small contributions
  - figure 19 shows the proportion of variance explained by each principal component
  - not particularly clear, could be at 5 principal components
- just enough principal components to explain 90% of the total variance

- figure 20 shows the cumulative proportion of variance explained each time an additional principal component is considered
- it takes 27 principal components to explain 90% of the total variance, reducing dimensionality by 21

Figure 19: Proportion of variance explained by principal components - all households on one day.

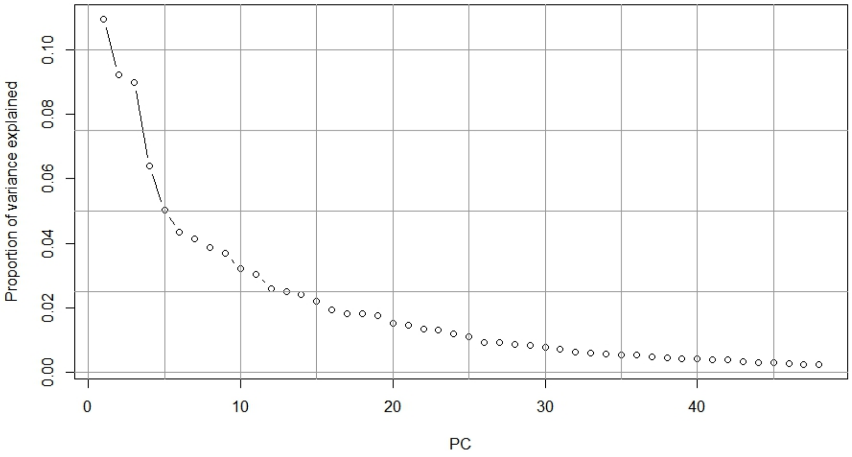
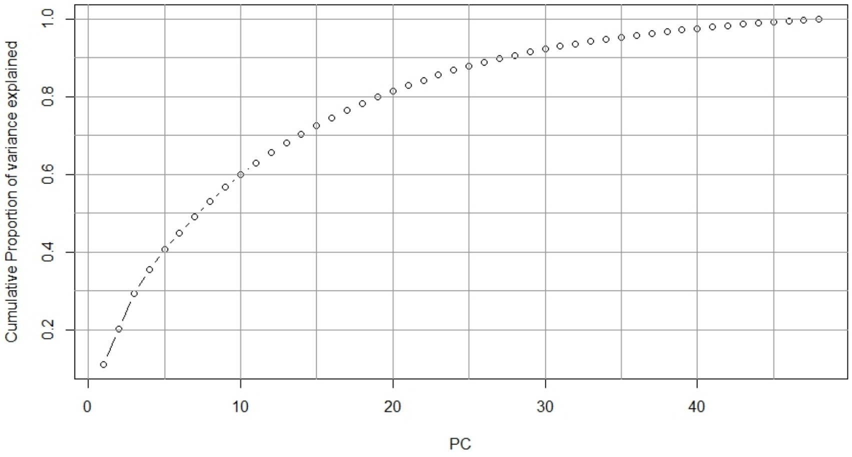


Figure 20: Cumulative proportion of variance explained by principal components - all households on one day.



The rules of thumb suggest to keep 5, 15 and 21 principal components, which explain 40.6%, 72.6% and 89.7% of the variance respectively, and the final choice comes down to the motivation for using PCA. Beckel *et al.* (2014) selected the first 10 principal components to help classify households when looking at electricity profiles. They do not state why the first 10 principal components were chosen, or to what "rule of thumb" they worked to, but they were held as variables along with other features they had picked out from the electricity profile data, so they may have been more concerned about reducing dimensionality than getting a more complete explanation of variance. Chelmis *et al.* (2015) were explicitly concerned about reducing dimensionality, it happened that the variation in their data could be 96% explained by four principal components.

In this dissertation the motivation is to get a more complete explanation of variance, and if 90% was achievable with a greater reduction in dimensionality then it would have been explored further, with clustering methods applied to the principal components.

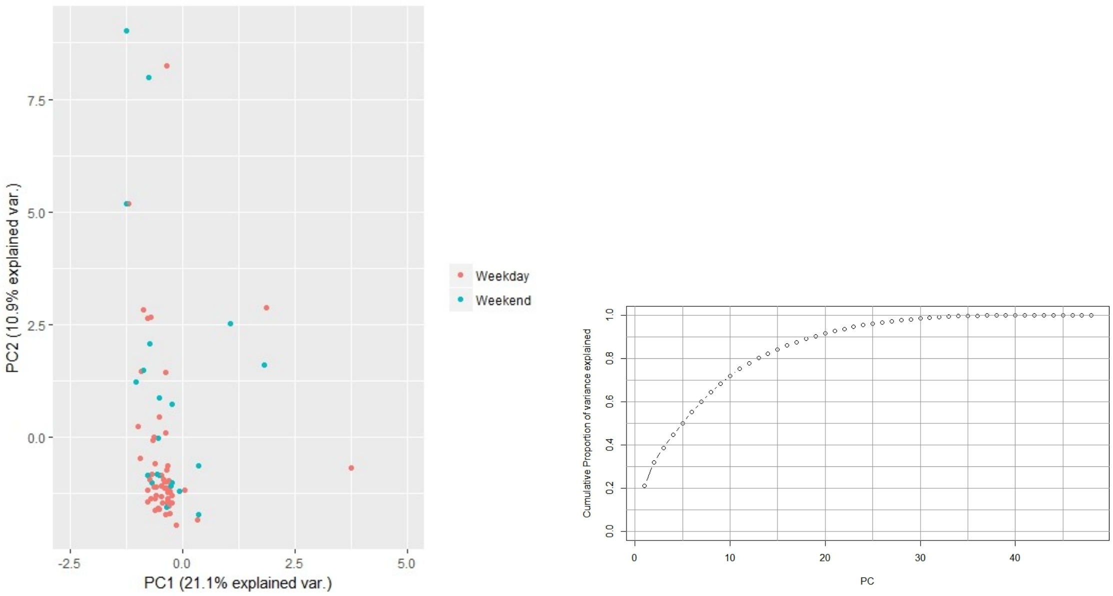
5.2 Individual households

The data is organised differently to the previous subsection, in that rather than each matrix being every household on one day, it is one household over many days; Chelmis *et al.* (2015) also considered this data arrangement to study the daily pattern per campus building. It is outside the objective of this dissertation, but by structuring the data this way, the principal components explain variances in each households individual behavior. External factors could be brought in, such as temperature or energy price bands, to see if households change their behaviour based on these external factors.

For this case study, the households to be analysed are the same three as in section 4; the period considered here is every day between Sunday 10th January and Saturday 27th March 2010, inclusive; and the grouping is on weekday and weekday. The period begins after the Christmas holiday season, and ends before the BST hour change. This means that each matrix is 77×48, representing 77 days of 48 periods, and as above each variable is normalised before the principal components are calculated.

The left-hand side of figures 21, 22 and 23 show the first two principal components, and the weekday groupings. The axis labels show the proportion of variance explained by PC1 and PC2; for the retired household, together they explain 55.4%, compared to 32.0% and 23.6% for the employed and unemployed households respectively. Complimenting this, on the right-hand side of the figures is a cumulative scree plot, and the retired household meets the Mardia *et al.* (1979) "rule of thumb" to explain 90% of the total variance with 10 principal components, while the employed and unemployed households take 19 and 21 principal components respectively. In section 4, the retired household was seen to be the most consistent consumer, with regular periods and magnitudes of consumption, and the unemployed household the least consistent.

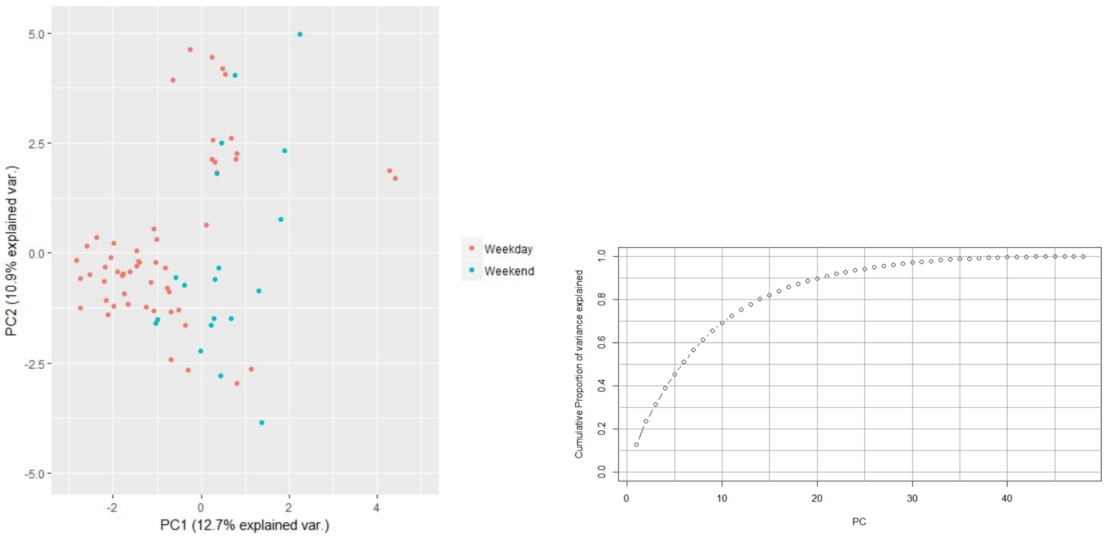
Figure 21: Principal component analysis - one employed household across many days.



Principal components one and two, coloured by day type

Cumulative proportion of variance explained

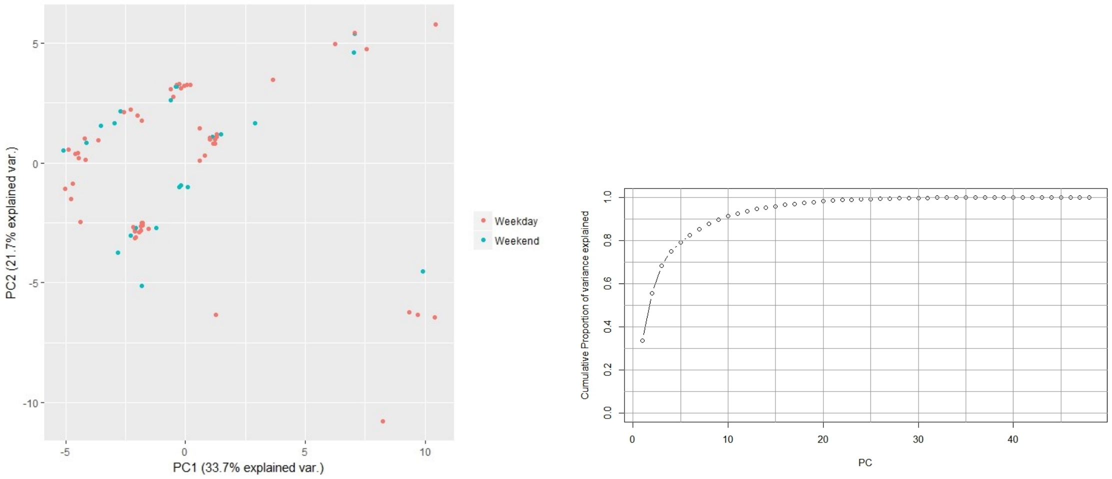
Figure 22: Principal component analysis - one unemployed household across many days.



Principal components one and two, coloured by day type

Cumulative proportion of variance explained

Figure 23: Principal component analysis - one retired household across many days.



Principal components one and two, coloured by day type

Cumulative proportion of variance explained

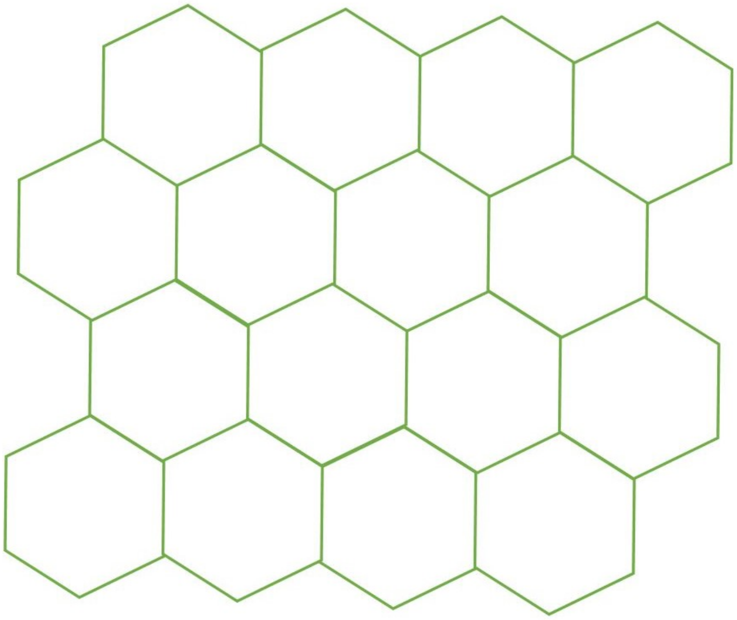
In terms of grouping the weekdays and weekends, it is the unemployed household which appears to have the most distinct clustering, with a group of weekdays with no weekends on the left-hand side of the plot. The employed household is less distinct, with the majority of weekdays grouped together, and days not in this group tending to be weekends; which could be due to a more stringent routine in the week due to work commitments. The retired household has two linear groups, but it is not weekday and weekend which is causing these groups.

If pursuing this, a revised method for centring and scaling of variables would have to be investigated. For some households there is the same period of zero consumption on every day of the sample, and so scaling to unit variance is not possible. In the case of the sample households there was a trace of consumption in almost every period in the sample, so unit variance could be calculated; but it may be acting in the opposite way as intended, giving these periods of essentially zero consumption equal weighting to periods of actual consumption.

6 Self Organising Maps

As seen with PCA, it can be difficult to interpret the 48 variables present in the gas data, so this chapter explore the SOM method, as used by Mcloughlin *et al.* (2015) and Beckel*et al.* (2012). SOMs, often known as Kohonen Self Organising Maps are useful for visualising relationships between high-dimensional input data. They were originally introduced by Kohonen (2001), and provide a classification of input data and maps of attribute space. From a data science standpoint, they are an attractive proposition due to relatively intuitive computation and easily interpretable results - that is the computation is based on Euclidean distances, and the closer observations are in the attribute space, the more similar they are. A SOM is a two-dimensional array of neurons, and each neuron has a vector of length p, where p is the number of variables (in this case $p = 48$). The neurons are initially assigned vectors of random variables, constrained by the attribute space. neurons are arranged in a grid of D neurons with each neuron having a non-zero number of neighbouring neurons, usually they are square or hexagonal. Figure 24 shows an example of a 4×4 grid of hexagonal neurons. The neurons are initially assigned vectors of random variables, constrained by the attribute space. There are versions of SOMs in a 3D space, where the edges are joined, but these are difficult to visualise on paper, and so will not be explored further, (Schmidt *et al.* 2011).

Figure 24: Example of a 4×4 SOM grid.



The input vectors can be represented as $\mathbf{x} = \{x_i : i = 1, ..., p\}$, and the neuron vectors represented as $\mathbf{m}_j = \{m_{ji} : j = 1, ..., N; i = 1, ..., p\}$. To train the SOM, each vector from the input data is presented in turn to the grid of neurons and a discriminant function is calculated as the squared Euclidean distance between the input vector $\mathbf{x}$ and the vector $\mathbf{m}_j$ for each neuron j, where the smallest distance is the winning neuron

$$d_j(\mathbf{x}) = \sum_{i=1}^{p} (x_i - m_{ji})^2, \tag{4}$$

or can be written

$$d_j(\mathbf{x}) = \left\| \mathbf{x} - \mathbf{m}_j \right\|, \tag{5}$$

and the winning neuron c satisfies

$$\|\mathbf{x} - \mathbf{m}_c\| = \min_j \{\|\mathbf{x} - \mathbf{m}_j\|\}. \tag{6}$$

The winning neuron learns from the input vector, but also the neurons surrounding the winning neuron learn from the input vector - within the brain, there is lateral interaction between neurons, there is a topological neighbourhood that decays with distance. In the SOM there is a neighbourhood function written in terms of the Gaussian function

$$h_{cj}(t) = \alpha(t) \times \exp\left(-\frac{R_{cj}^2(t)}{2\sigma^2(t)}\right), \tag{7}$$

where where $t = 0, 1, 2, ...$ is an integer, the discrete-time coordinate; R_{cj} is the lateral distance between the best matching neuron c and neuron j; $\alpha(t)$ is a learning rate factor where $0 < \alpha(t) < 1$; and, the parameter $\sigma(t)$ defines the width of the kernel - both $\alpha(t)$ and $\sigma(t)$ are some monotonically decreasing functions of time.

The learning from the input vector is shown by

$$\mathbf{m}_j(t + 1) = \mathbf{m}_j(t) + h_{cj}(t)[\mathbf{x}(t) - \mathbf{m}_j(t)], \tag{8}$$

This is showing that the new neuron vector is equal to the old neuron vector, plus a fraction of the difference between the old neuron vector and the input vector. At the edge of the neighbourhood, the fraction is less, so these neurons are less affected by the input vector.

Brunsdon and Singleton (2015, pp158-159) use the analogy of shooting data at the grid; the data hits the best matching neuron, and the deformation of the grid from the ‘impact’ causes nearby neurons to adjust their values to be more similar to the best matching neuron. The neurons are randomly intialised, but as data hits the grid, areas of deformation emerge. Continuing the analogy, as the training process continues, the speed of the data hitting the grid decreases, and the size of the deformation decreases too.

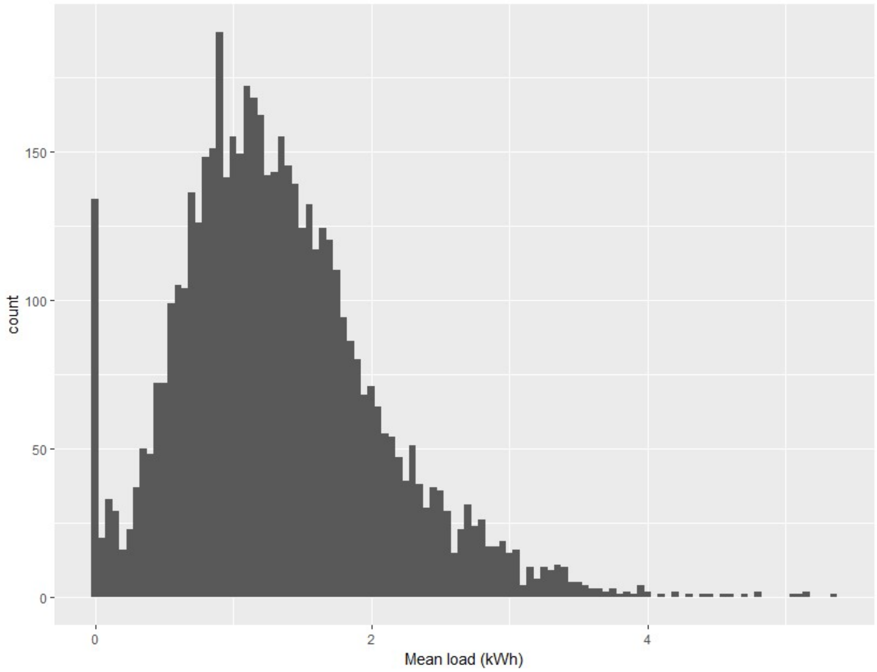
6.1 Defining the map

To use the SOM algorithm in R, the `kohonen` package is necessary. The first step is to define the dimensions of the SOM grid, and the shape of the neurons. There is no specific rule to follow in choosing the number of neurons - to achieve a fine distinction between categories, a larger SOM is needed. McLoughin *et al.* (2014) used between 8 and 10 neurons for one day of data; Beckel *et al.* (2012) did not specify, but used a much larger number of neurons, and then cluster the neurons; Brunsdon and Singleton (2015, p162) used 16,000 neurons for 211,000 input vectors on research for US Census Bureau (13.19 vectors per neuron).

To train the SOM, the data used has the following features:

- Tuesdays, Wednesdays and Thursdays
  - As in Anderson *et al.* (2017), this is to avoid the different consumption behaviour of weekends and transitional days, Monday and Friday.
- From Tuesday 12th January 2010
  - As seen in subsection 4.4, consumption behaviour over the Christmas and New Year holiday period was different to usual.
- To Thursday 25th March 2010
  - As seen in subsection 4.4, consumption behaviour was affected by the BST hour change, and then reduced through April.
- Excludes the lowest and greatest 5% of mean daily consumption
  - Training a SOM with a smaller data set resulted in one neuron having an imbalance of associated vectors compared to other neurons;
  - This neuron had a mean load of 0.0370, which is close to no gas consumption in the day, and are likely vacant on the sample days;
  - Figure 25 shows there is a spike at 0 to 0.05 kWh mean load in that sample data, and overall it is a gamma distribution.

Figure 25: Histogram showing the mean hourly gas consumption per household for Wednesdays in March.



R code that defines the training data:

```
##Select subset, mid Jan to hour change in March, Tuesday, Wednesday
##and Thursday
data_train_tmp = GasWideSurv[(GasWideSurv$Day=="Tuesday"
|GasWideSurv$Day=="Wednesday"
|GasWideSurv$Day=="Thursday")
                              & GasWideSurv$Date>="2010-01-12"
& GasWideSurv$Date<="2010-03-25",]

##Calculate and exclude 5% and 95%
data_train_tmp$ML= rowMeans(subset(data_train_tmp,select=c(4:51))
,na.rm=TRUE)
ggplot(data=data_train_tmp, aes(data_train_tmp$ML))
+ geom_histogram(binwidth=0.1)
+ labs(x = "Mean load (kWh)")

q05=quantile(data_train_tmp$ML, probs = c(0.05))
```

```
q95=quantile(data_train_tmp$ML, probs = c(0.95))
data_train_tmp=data_train_tmp[data_train_tmp$ML>q05 & data_train_tmp$ML<q95,]
ggplot(data=data_train_tmp, aes(data_train_tmp$ML)) + geom_histogram(binwidth=0.1) +

data_train=as.matrix(data_train_tmp[,4:51])
```

This results in 38,490 input vectors of 48 periods to train the SOM. Using a similar ratio as Brunsdon and Singleton (2017) gives a 50×57 of 2,850 neurons. The R code to define the map dimensions:

```
## Specify grid dimensions
som_grid = somgrid(xdim = 50, ydim=57, topo="hexagonal")
```

To train the model, the parameters to specify are (The R Project for Statistical Computing, 2017):

- `rlim` - the run length, specifies how many times the input data will be presented to the SOM;
- `alpha` - the learning rate factor, a range is specified as the rate monotonically decreases over the run length;
- `radius` - the size of the neighbourhood, a range is specified, decreasing to one (the best matching neuron only).

Similarly to setting the size of the grid, there is no specific rule to choose the run length other than too small a run length and the SOM may not be very well trained. Kohonen (2001, p211) gives a “rule of thumb” that for “good statistical accuracy the number of steps must be at least 500 times the number of network units (neurons)” and often use up to 100,000 steps - a step in this sense is one input vector presented to the SOM, not the full input data. As computing power has changed since this “rule of thumb”, to be at least 500 times the number of neurons for the chosen grid size, would be at least 1,425,000 steps. Given the number of input vectors, the run length should be at least 37 (that is to present all input vectors 37 times).

Running the SOM with `rlen = 37` took approximately 5 minutes (in R Studio, 2.4GHz Intel i7-5500 CPU, 7.88GB RAM). As more iterations lead to a better trained SOM, increase ten-fold to `rlen = 370` took 60 minutes. This still was not a significant pressure on computational power, so it was run with `rlen = 740` which took 170 minutes.

The default radius is set at two-thirds of the grid reducing to one. Kononen (2001, p112) states that if too small a radius is chosen, "mosaic-like parcellations of the map are seen", which can be avoided by starting with a large radius and reducing over time. As such, the SOM runs with the default radius.

Kononen (2001, p112) suggests the learning rate starts at about 0.9, decreasing to about 0.02, and that "an accurate time function is not important", but Brunsdon and Singleton (2017, p163) use the `kohenen` default of 0.05 and 0.01. Given the significant difference between the two parameters, the SOM was run for both sets of parameters, and the trained maps compared.

The R code that defines the training process of the SOM; `alpha` was changed to `c(0.05,0.01)` for the second training process:

```
set.seed(25817)
# Train the SOM
som_model3 = som(scale(data_train, center=TRUE, scale=TRUE)),
                 grid=som_grid,
                 rlen=740,
                 alpha=c(0.9,0.02),
                 keep.data = TRUE)
```

There are several plots to consider when considering the success of a trained SOM (The R Project for Statistical Computing, 2017):

- `changes` shows the mean distance to the nearest neuron for each iteration
  - Figures 26 and 27 show this plot for the two versions of the map.
- `counts` shows the number of input vectors assigned to each neuron
  - Figures 28 and 29 show this plot for the two versions of the map;
  - Figures 30 and 31 are histograms showing the number of input vectors per neuron for the two versions of the map.
- `quality` shows the similarity of input vectors to each neuron
  - Figures 32 and 33 show this plot for the two versions of the map.
- `dist.neighbours` shows the similarity between neurons, or the distance to each neighbour, and is often referred to as a 'u-matrix'. If the neighbour distance is low then that is a similar group of neurons, and any areas of dissimilarity are often what form cluster boundaries.

– Figures 34 and 35 show this plot for the two versions of the map.

Figure 26: SOM training process with 740 iterations, `alpha=c(0.9,0.02)`.

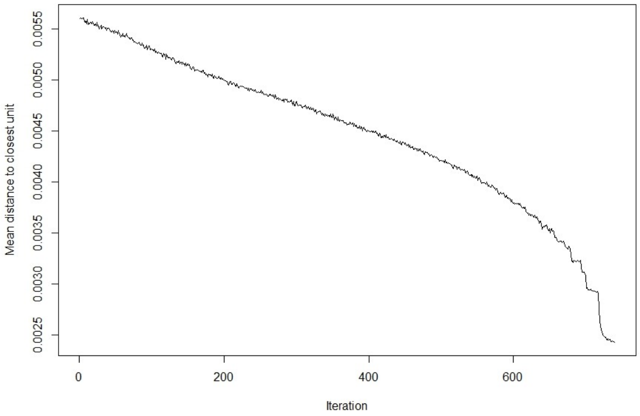
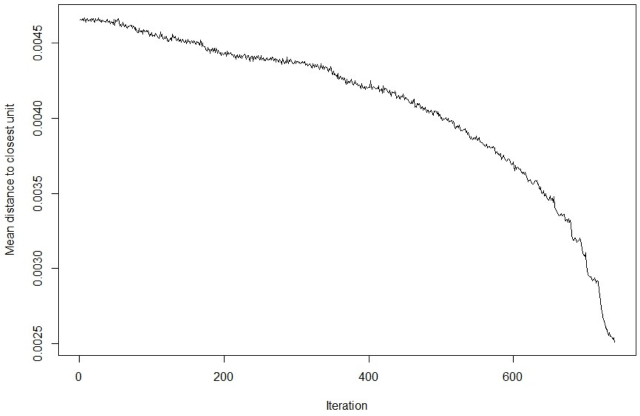


Figure 27: SOM training process with 740 iterations, `alpha=c(0.05,0.01)`.



Plotting `changes` shows the mean distance to the nearest neuron for each iteration, and for true confidence that the map is completed training, convergence would be expected, as each new iteration causes minimal changes to the shape of the map. In these cases seen in figures 26 and 27 there is not much evidence of convergence. To over come this more iterations can be presented, but on doubling the number for the first model the processing time became impractical. There is nothing between these two figures that would suggest one is a better trained model than the other.

Figure 28: SOM neuron allocation, `alpha=c(0.9,0.02)`.

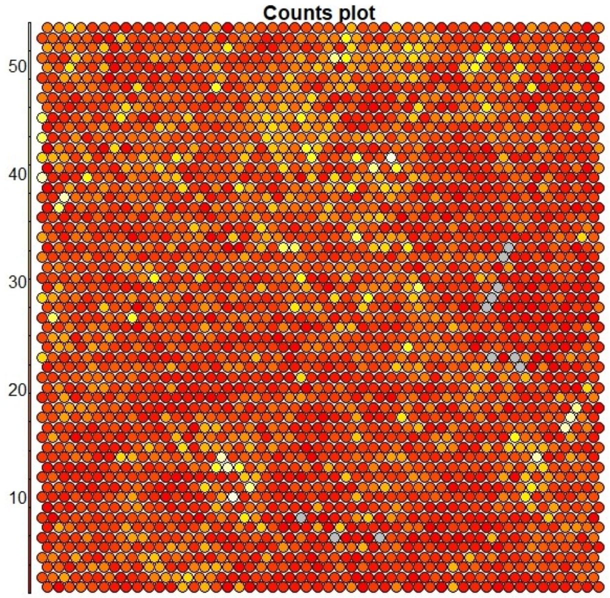


Figure 29: SOM neuron allocation, `alpha=c(0.05,0.01)`.

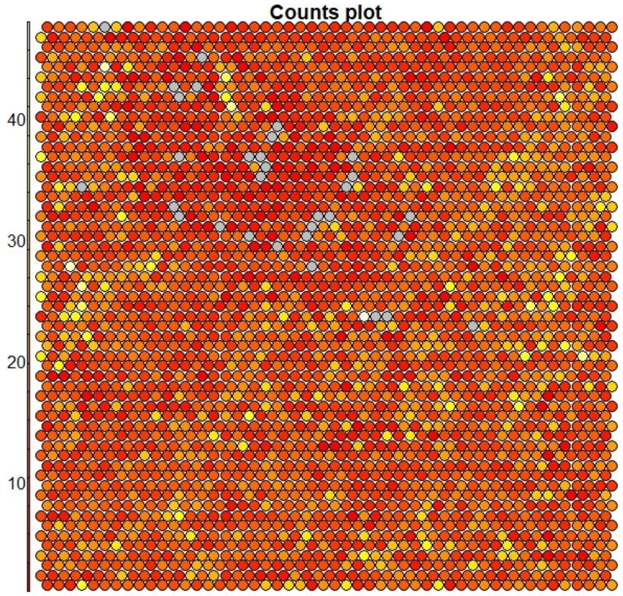


Figure 30: Histogram showing SOM unit allocation, `alpha=c(0.9,0.02)`.

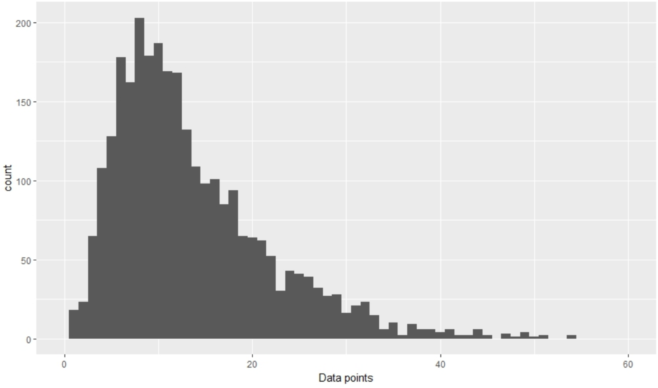
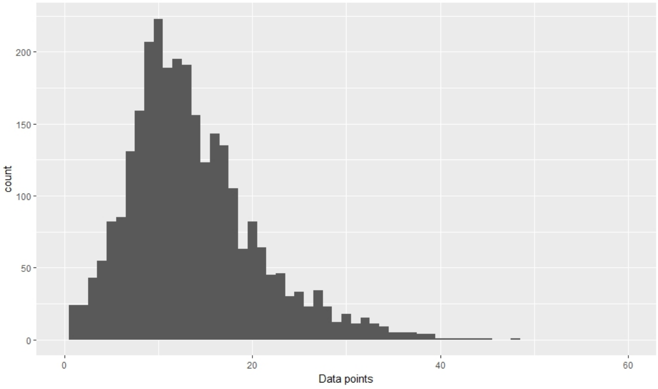


Figure 31: Histogram showing SOM unit allocation, `alpha=c(0.05,0.01)`.



Plotting `counts` shows the number of input vectors assigned to each neuron, and less variance is preferred. The histograms in figures 30 and 31 show similar gamma distributions, but figure 30 has a longer tail. This model has more neurons with high numbers of input vectors associated, and these can be seen as the white spots in figure 28. Figure 31 shows that this model has more 'empty' neurons - neurons with zero input vectors associated, and are seen in the allocations plots as grey spots. These neurons have still been subject to training and have been updated by being in the neighbourhood of the neurons around it, but suggest a smaller map might be suitable; whilst having too many over-allocated neurons suggests a larger map might be suitable.

Figure 32: SOM neuron quality, `alpha=c(0.9,0.02)`.

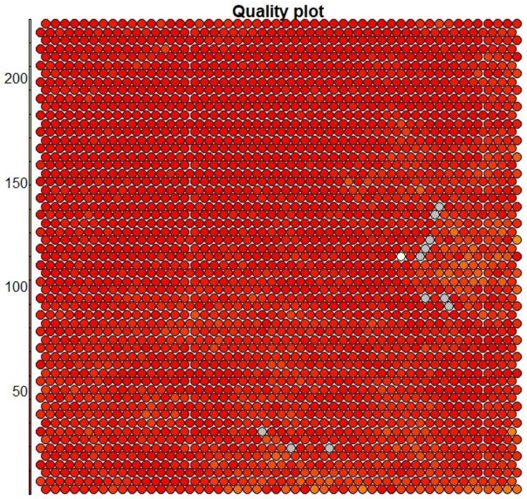
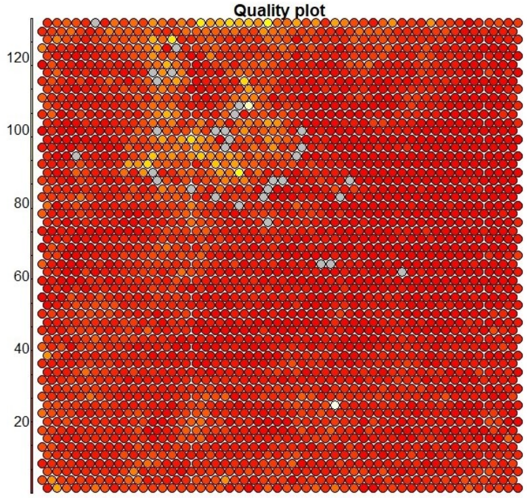


Figure 33: SOM neuron quality, `alpha=c(0.05,0.01)`.



Plotting `quality` shows the distance of input vectors to each neuron; the smaller the distances, the better represented the input vectors are. The colour scale runs from red for the smallest distances through yellow to white for the largest differences. Figure 32 and 33 are outputs from the `kohonen` package, but a limitation with this package is the scale given to each heat map is not programmable. Figure 32 looks to have a more consistent quality over the map, but this is caused by one neuron which is very low quality (high distance) compared to the other neurons. Figure 33 actually has better quality neurons and less variation, but there appears to be more variation based on the plot due to a lower maximum distance.

Figure 34: SOM neighbour distances, `alpha=c(0.9,0.02)`.

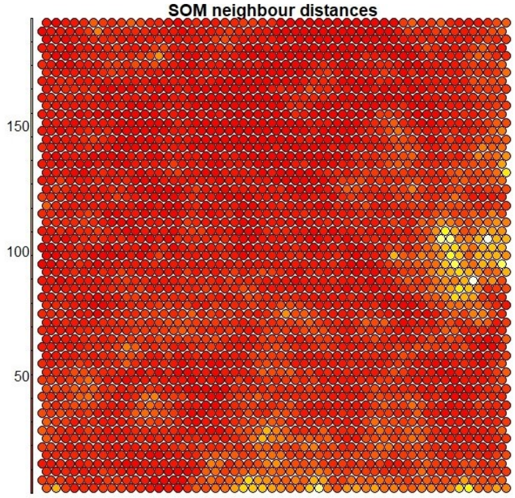
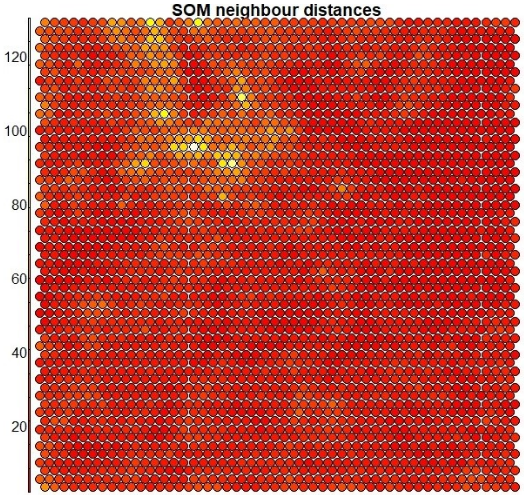


Figure 35: SOM neighbour distances, `alpha=c(0.05,0.01)`.



Plotting `dist.neighbours` shows the similarity between neurons, or the distance to each neighbour, and is often referred to as a ‘u-matrix’. If the neighbour distance is low then that is a similar group of neurons, and any areas of dissimilarity are often what form cluster boundaries. Figures 34 and 35 show that the two models have different areas of dissimilarity, and so would likely cluster differently. That SOMs are presented in 2D and as a fixed grid is a positive when trying to visually interpret the results, but this can be deceiving if there are areas of the map where neurons are less similar to their neighbours, so for this reason, less variation is preferred; figure 35 has the less variation in neighbour distance of the two.

The results from the SOM neuron quality, and neighbour distance plots suggest that the model run with the default parameters for alpha, `alpha=c(0.05,0.01)`, is the better choice and will be carried forward into the next sections. If more resources were available then making adjustments to the number of iterations and to the map size could improve the model, with potentially a convergence in training and fewer empty neurons.

6.2 Clustering the SOM

Having selected the most suitable map, this subsection looks at clustering the neurons. The data has been reduced from 38,490 input vectors to 2,850 neurons, each with its own unique load profile. k-means clustering is the chosen clustering method, for it's familiarity and ease of understanding, and as used in much of the literature. To select k, the Davies-Bouldin (DB) index is used, as it was by McLoughlin *et al.* (2015).

The DB index was introduced by Davies and Bouldin (1979) as a method to indicate the similarity of clusters, and can be used to infer the appropriateness of partitions of the data. It measures the ratio of the within-cluster scatter S of two clusters i and j, to the seperation between the two clusters M. These two measures are Euclidean distance functions, as used in k-means.

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}, \tag{9}$$

which is used to define D_i

$$D_i \equiv \max_{j \neq i} R_{ij}, \tag{10}$$

and this is used to calculate the DB index for N clusters

$$DB \equiv \frac{1}{N} \sum_{i=1}^{N} D_i. \tag{11}$$

The R code to calculate the DB index for 2 to 30 clusters, using the `clusterSim` package and `index.DB` function:

```
#Davies-Bouldin cluster separation measure
library(clusterSim)
##the load profiles for each neuron
som_codes=do.call(rbind, lapply(som_model3b$codes, data.frame
, stringsAsFactors=FALSE))
```

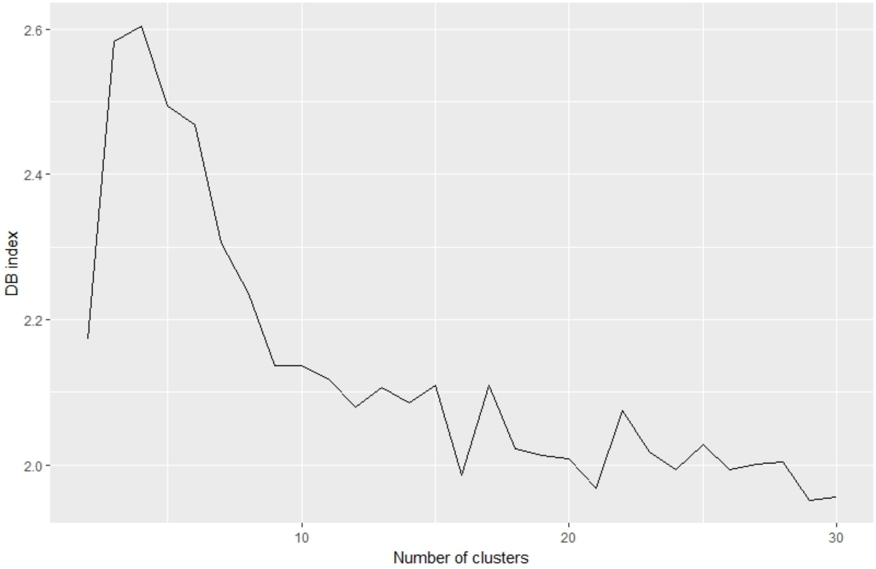
```
##Variables to store results
v1=NULL
v2=NULL

for (i in 2:30){
  nc=i
  ##k means
  SOMClusters=kmeans(x=(som_codes),
centers =nc,nstart=1000,iter.max = 10000)

  ##store cluster number
  v1=rbind(v1,nc)
  ##store the DB index for each cluster number
  v2=rbind(v2,index.DB(som_codes,SOMClusters$cluster)$DB)
  print(index.DB(som_codes,SOMClusters$cluster)$DB)
}

##Combine results and plot
DB.index=data.frame(cbind(v1,v2))
ggplot(data=DB.index,aes(x=X1,y=X2))+geom_line()
+xlab("Number of clusters")+ylab("DB index")
```

Figure 36: DB index for 2 to 30 clusters.



The aim is to choose the number of clusters with the minimum intra-cluster distance, so the smaller the DB index the better (Saitta *et al.*, 2007). The value of DB would continue to decrease until each neuron was its own cluster, so simply choosing the minimum DB is not applicable. In figure 36 there is only marginal change going from 16 clusters to 30 clusters, so k=16 was chosen.

The clusters are presented on the map in figure 37, and the composition of the clusters in terms of neurons and input vectors in table 3. The k-means clustering is performed in a 48-dimensional space, whilst the map is presented in a 2D format, which is why there is apparent overlap between the different clusters. The map in itself is not particularly interesting for understanding the clusters, but the load profile shapes of clusters are, and are introduced in the next section.

Figure 37: SOM with 16 clusters.

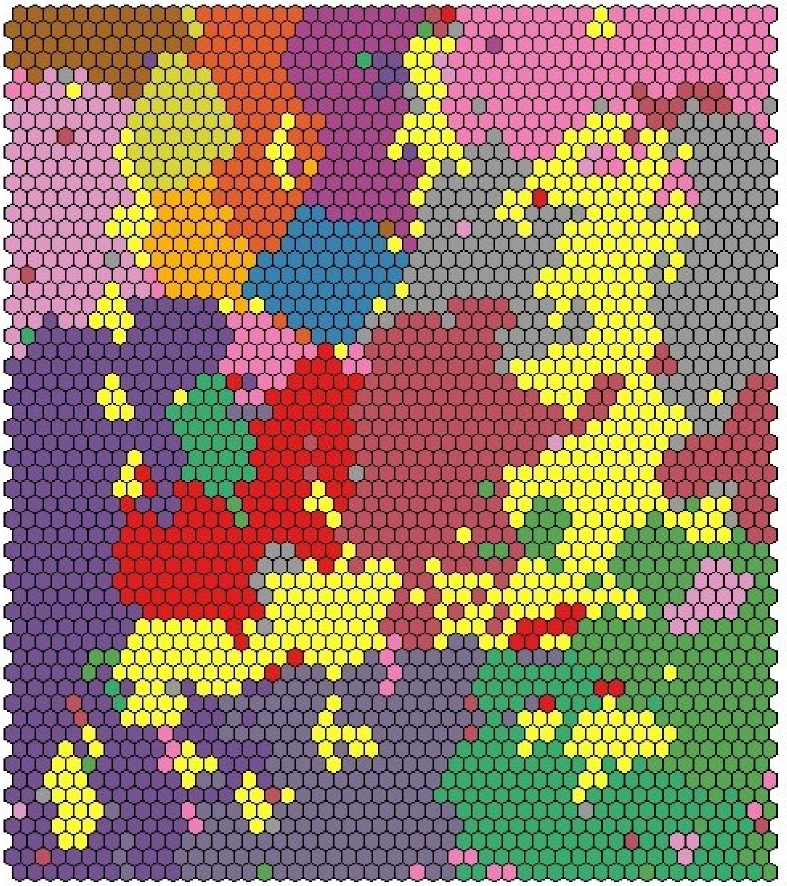


Table 3: Counts of neurons and input vectors per cluster.

Cluster	No. of neurons	No. of input vectors	% of neurons	% of input vectors
1	242	3,850	8.6	10.0
2	140	2,094	5.0	5.4
3	204	2,597	7.2	6.7
4	270	3,576	9.6	9.3
5	51	743	1.8	1.9
6	45	460	1.6	1.2
7	507	7,538	18.0	19.6
8	41	388	1.5	1.0
9	78	837	2.8	2.2
10	91	906	3.2	2.4
11	210	2,917	7.4	7.6
12	175	2,446	6.2	6.4
13	200	2,770	7.1	7.2
14	57	454	2.0	1.2
15	353	4,952	12.5	12.9
16	155	1,962	5.5	5.1

6.3 Cluster load profiles

To better understand the shape of the load profiles of each cluster, the mean consumption for each period was calculated for each cluster, and plotted on top of the load profiles associated with that cluster. The plots can be seen in figure 38, and all are to the same scale.

While initially the clusters look similar to each other, the majority have distinct features:

Two peaks:

- Cluster 1 has a morning peak smaller than evening peak;
- Cluster 2 has a morning peak greater than evening peak;
- Cluster 4 has a morning peak slightly smaller than evening peak;
- Cluster 7 has similar morning and evening peaks, of lesser magnitude than other clusters;
- Cluster 16 has similar morning and evening peaks, of a greater magnitude than other clusters.

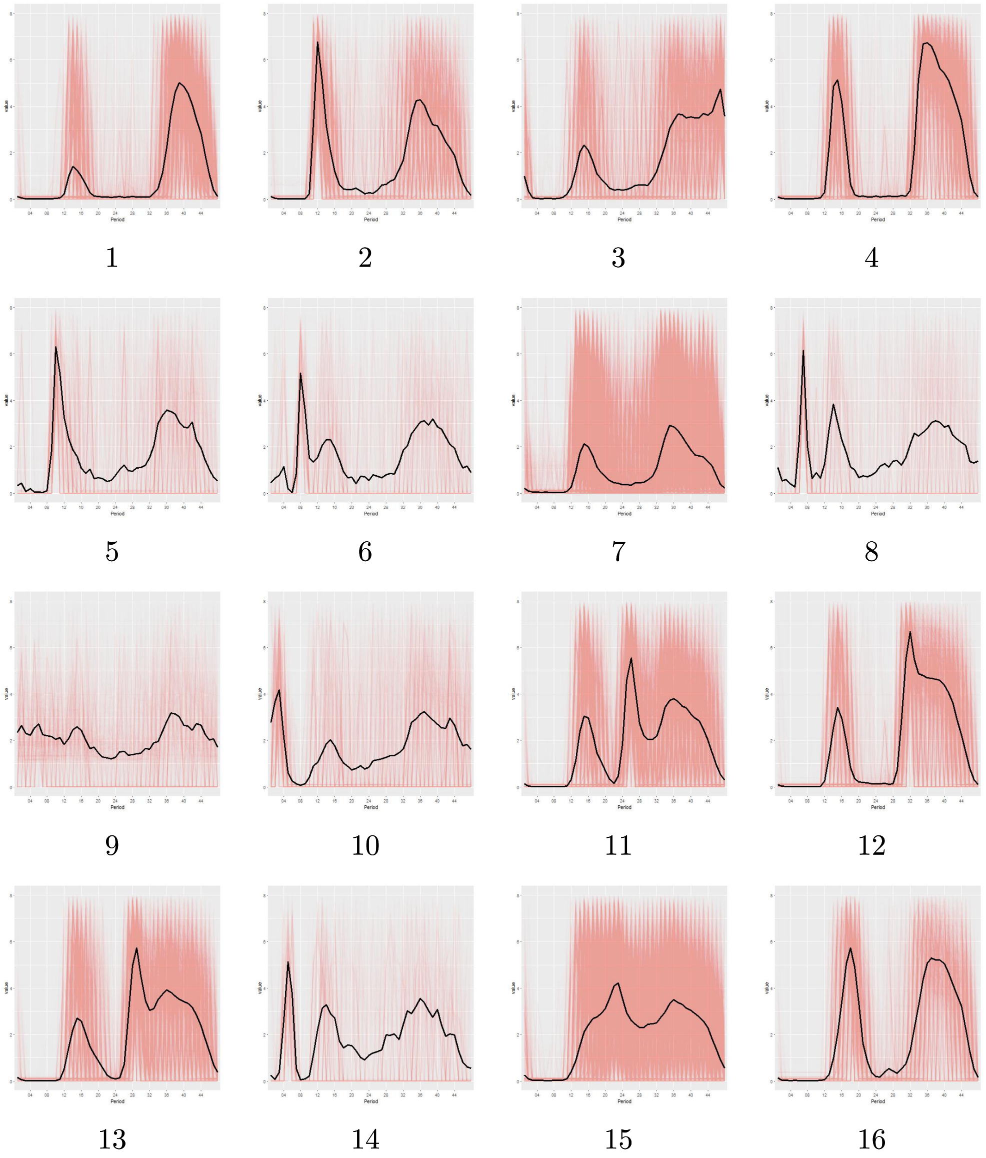
Multiple PM periods of consumption:

- Cluster 3 has a second increase in consumption up to period 47 (23:00-23:00);
- Cluster 11 has a peak at around period 26 (12:30-13:00) and at period 36 (17:30-18:00);
- Cluster 12 has a high plateau periods 34 to 39 (16:30-19:30) following a peak;
- Cluster 13 has a peak at around period 30 (14:30-15:00) and at period 36 (17:30-18:00);
- Cluster 15 has a steady rise from periods 12 to 23, a steady dip and rise to second peak at period 36 (17:30-18:00).

The above 10 clusters are those with the clearest and most interpretable shapes, and they also are the clusters which each contain over 5% of the neurons and over 5% of the input vectors. The remaining six clusters are less distinguishable, and may have been better treated as fewer clusters.

- Clusters 6, 8 and 14 all have the highest peak in the earlier morning, two morning peaks, and a shallow rise to an evening peak;
- Cluster 10 is similar in shape, but with a morning peak around period 5 (02:00-02:30) which is distinct from all other clusters;
- Cluster 5 is similar, but does not have a secondary morning peak; it is also similar to Cluster 2 but with an earlier start;
- Cluster 9 has consumption across the whole day and night, never less than 1kWh, with less severe spikes.

Figure 38: Mean load profiles for each cluster, plotted over actual load profiles associated with each cluster.



1 2 3 4

5 6 7 8

9 10 11 12

13 14 15 16

7 Cluster characteristics

One of the key factors driving fuel poverty is household income level, and the CER survey asked this question:

> QUESTION 4021
>
> Can you state which of the following broad categories best represents the yearly household income BEFORE TAX?
>
> 1 Less than 15,000 Euros
> 2 15,000 to 30,000 Euros
> 3 30,000 to 50,000 Euros
> 4 50,000 to 75,000 Euros
> 5 75,000 or more Euros
> 6 Refused

Unfortunately for the purpose of this dissertation, 73.9% of households refused to give an answer. Harold *et al.*(2015) used employment status and education level as proxies for household income, and so this section examines the clusters to see if any show a distinctly different compositions to the overall sample, for employment status, education, but with the addition of social class too.

7.1 Employment status

The CER survey had the following question about employment status:

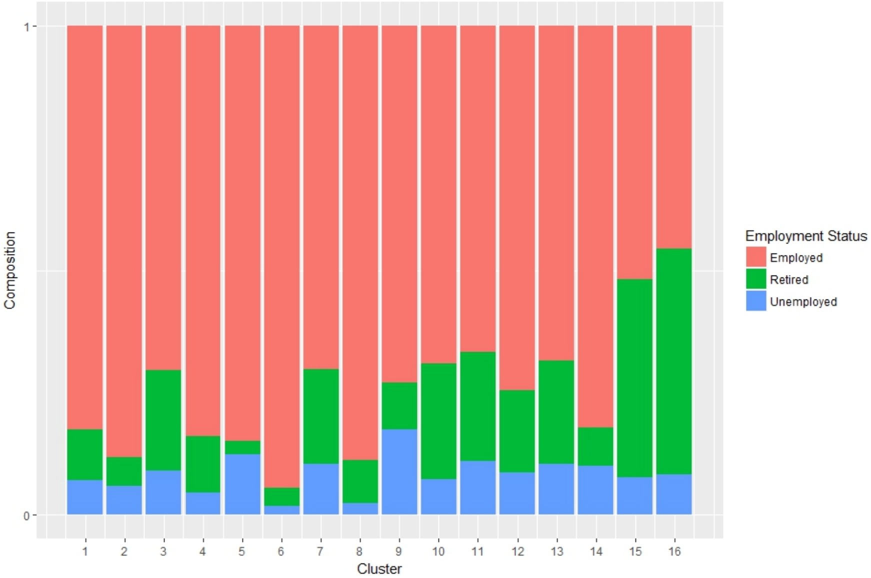
> QUESTION 310
>
> What is the employment status of the chief income earner in your household,
> is he/she
>
> 1 An employee
> 2 Self-employed (with employees)
> 3 Self-employed (with no employees)
> 4 Unemployed (actively seeking work)
> 5 Unemployed (not actively seeking work)
> 6 Retired
> 7 Carer: Looking after relative family

For simplicity the following analysis groups answers 1, 2, and 3 together as Employed; 4, 5, and 7 as Unemployed; and 6 as Retired. The distribution for the sample across all clusters is as in table 4, and is shown in figure 39 across all clusters.

Table 4: Employment status distribution across all input vectors.

Category	Percentage of sample
Employed	71.0
Unemployed	8.6
Retired	20.5

Figure 39: Composition of employment status across clusters.



Perhaps with more clarity, 40 shows the percentage of the sample that is unemployed at 8.6%, against the percentage for each cluster; and for cluster 9 it is 17.4% unemployed, over twice the rate. Seen in figure 38, cluster 9 has no severe peaks and quite constant consumption through the day compared to other clusters; this ties to the Beckel *et al.* (2012) observation of electricity profiles, that those in employment have a greater difference between the mean consumption and maximum consumption than those in unemployment. Also of note are clusters 6 and 8 which have less than half of the sample percentage at 1.7% and 2.3%, respectively.

Figure 41 is a similar plot but showing the percentage of retired households. Cluster 16 has over twice the sample percentage, at 46.1% against 20.5%; but also of note is cluster 15 which whilst not quite as significant at 40.5% retired, this cluster is the second largest, accounting for 12.9% of input vectors.

Figure 40: Percentage of each cluster which is unemployed, ordered by percentage; and the percentage of the sample shown by an intersecting line.

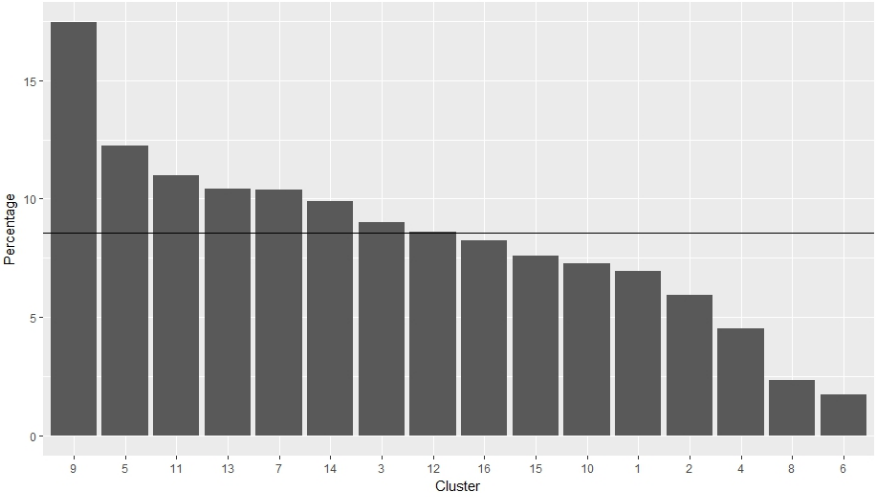
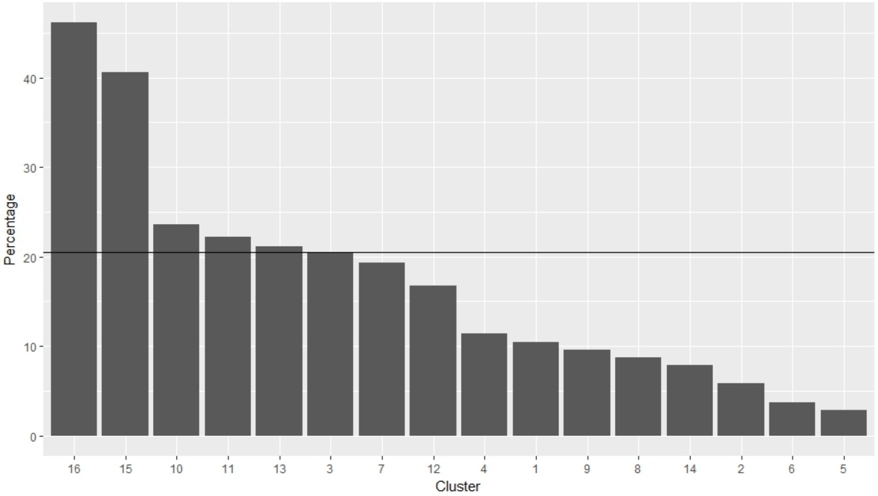


Figure 41: Percentage of each cluster which is retired, ordered by percentage; and the percentage of the sample shown by an intersecting line.



7.2 Social class

Social class is a way of grouping households, and is based on occupation. In the CER survey the respondent gives their occupation and the interviewer assigns a social class.

QUESTION 400

IF [Q310 , 1 , 2 , 3]

What is the occupation of the chief income earner in your household?

QUESTION 401

SOCIAL CLASS

Interviewer, Respondent said that occupation of chief income earner was....

Please code

1 AB
2 C1
3 C2
4 DE
5 F [RECORD ALL FARMERS]
6 Refused

There is no look-up provided with the data and manifest which says which occupations are assigned to which social class (other than that all farmers are assigned class F, and those not in employment are assigned to class E). A break down of the classes used in Ireland is, briefly:

- A - professionals, senior managers, top-level civil servants;
- B - middle management, principal officers in civil service, small business owners;
- C1 - junior management, owners of small establishments, all other non-manual positions;
- C2 - skilled manual workers, manual workers with responsibility for others;
- D - semi-skilled and unskilled workers, apprentices and trainees;
- E - entirely dependent on the state, long term;
- F - unique to Ireland, farmers and their dependents.

A common way to group these classes is as middle class - A, B, and C1; and as working class - C2, D, and E. As social class E is directly linked to employment status, that is it includes retired and unemployed households, in this instance it will be grouped separately to C2 and D. The distribution for the sample across all clusters is as in table 5, and is shown in figure 42 across all clusters.

Table 5: Social class distribution across all input vectors.

Category	Percentage of sample
ABC1	54.8
C2D	19.6
E	23.9
F	0.6
Refused	1.1

Figure 42: Composition of social class across clusters.

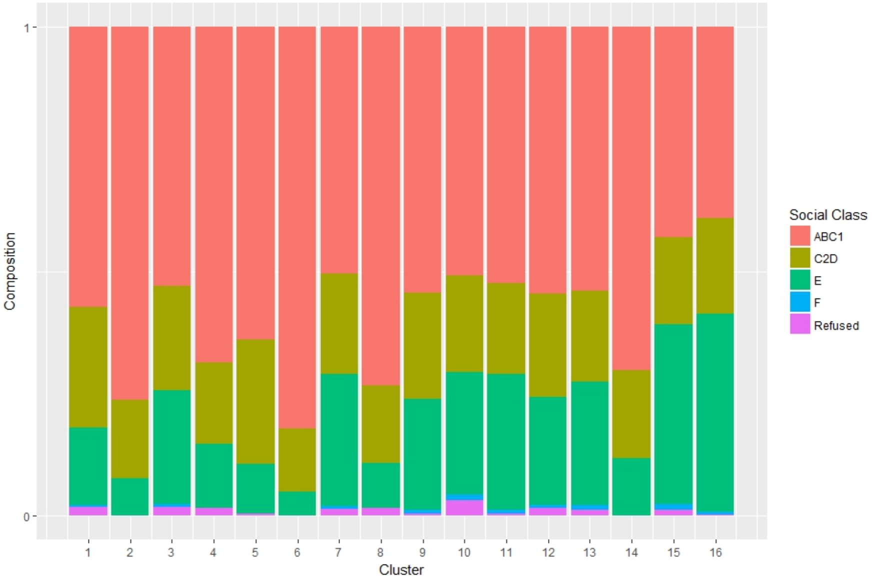


Figure 43 shows the percentage of the sample that is in social class C2 or D, so called 'working class', at 19.6%, against the percentage for each cluster. There is variation between clusters, but not as distinct as with unemployment and retired households.

Regarding social class E, when analysing the unemployment and retired rates of the clusters in the previous subsection, clusters with a higher rate of unemployment tended to have a lower rate of retirement, so it might be expected that these cancel each other out. Figure 44 shows that this is not the case, and the clusters on the left-most of figure 41 are also left-most here.

Figure 43: Percentage of each cluster which is in either social classes C2 or D, ordered by percentage; and the percentage of the sample shown by an intersecting line.

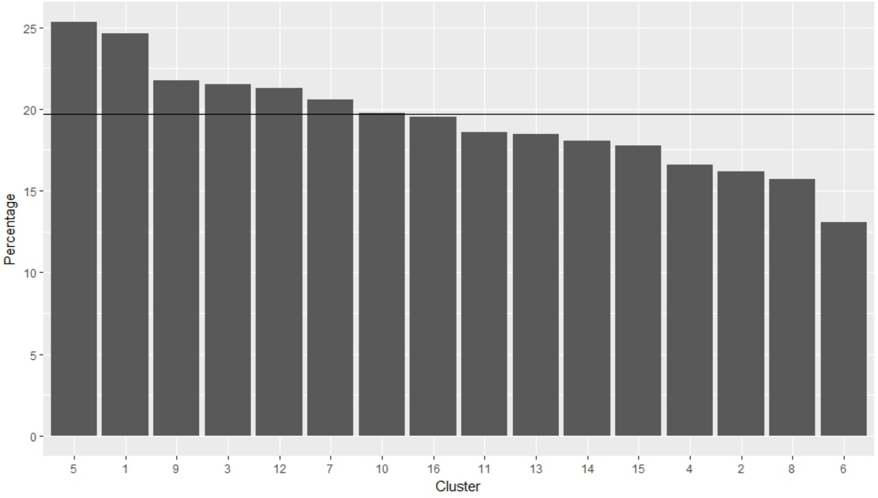
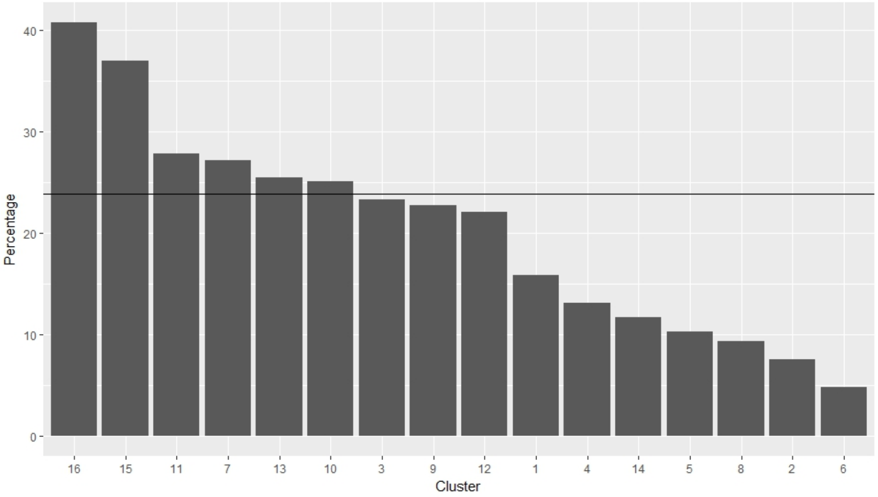


Figure 44: Percentage of each cluster which is in social class E, ordered by percentage; and the percentage of the sample shown by an intersecting line.



7.3 Education

Education is the final indicator considered as a proxy for household income. In Ireland, the Junior Certificate (which replaced the Intermediate Certificate in 1992) is completed by children ages 14 to 16, after three years of secondary school; and the Leaving Certificate completed by ages 17 to 19.

Table 6: Education level distribution across all input vectors.

Category	Percentage of sample
Intermediate Cert Level or less	21.5
Leaving Cert Level	23.1
Third Level	50.5
Refused	4.8

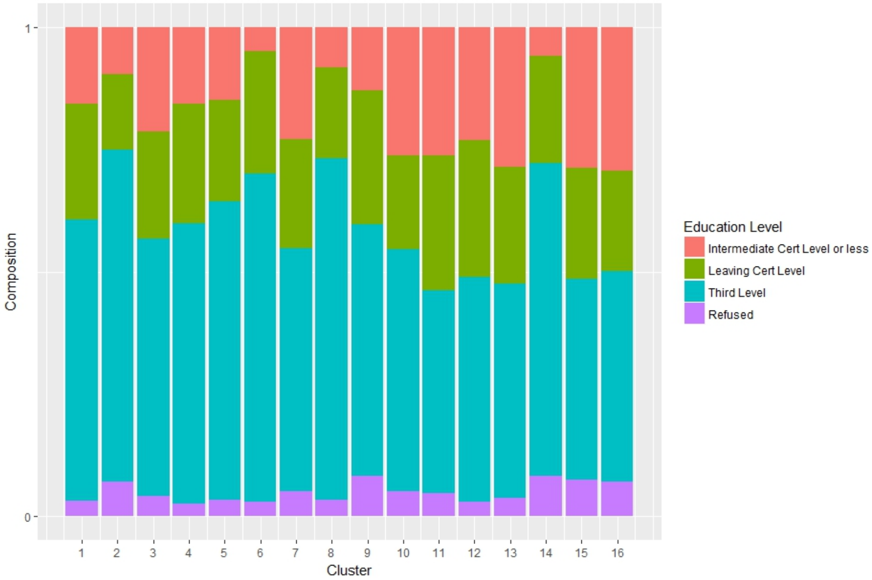
QUESTION 5418

Moving on to education, which of the following best describes the level of education of the chief income earner. Intermediate

1 No formal education
2 Primary
3 Secondary to Intermediate Cert Junior Cert level
4 Secondary to Leaving Cert level
5 Third level
6 Refused

As of 2000, the minimum leaving age is 16 years old, and under-18s must complete the Junior Certificate. For this reason, ‘No formal education’, ‘Primary’ and ‘Secondary to Intermediate Cert Junior Cert level’ were grouped together to incorporate all those who left education at the minimum age. The percentage of the sample in each group is in table 6, and of the clusters in figure 45.

Figure 45: Composition of education level across clusters.



Looking at just the ‘Intermediate Certificate Level or less’ group in figure 46, the two left-most clusters (those with the highest percentage), clusters 16 and 15, are also the left-most clusters in the retirement plot, figure 41 and the E social class plot, figure 44.

There is a higher percentage of respondents that have refused to give an education level than those who refused to give occupation (and hence get assigned a social class). There is a possibility that this is sensitive information to the respondent, especially if they have a lower level of education, so there is a risk of a bias in the sample to those with a higher level of education. The percentage per cluster of refusals is shown in figure 47, and cluster 9, which was the cluster with the largest percentage of unemployment, has the second highest rate of refusals - the mean is 4.8% and cluster 9 is 8.1%.

Figure 46: Percentage of each cluster which has education Intermediate Certificate Level or less, ordered by percentage; and the percentage of the sample shown by an intersecting line.

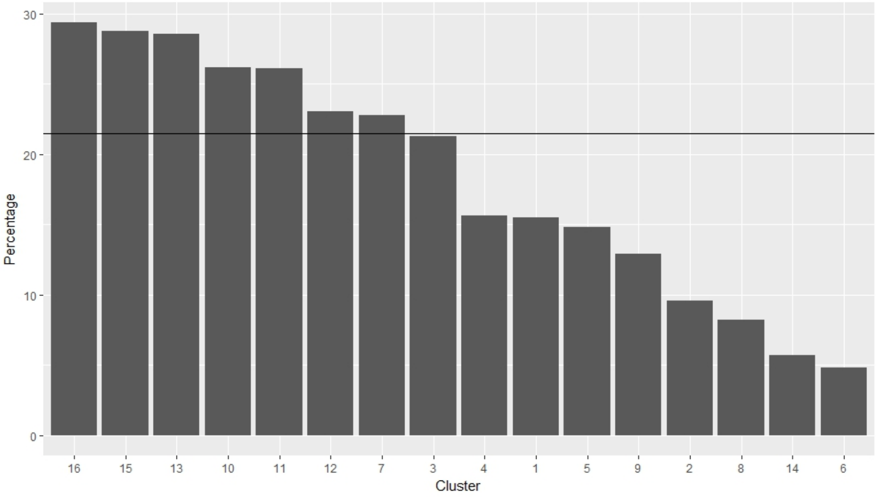
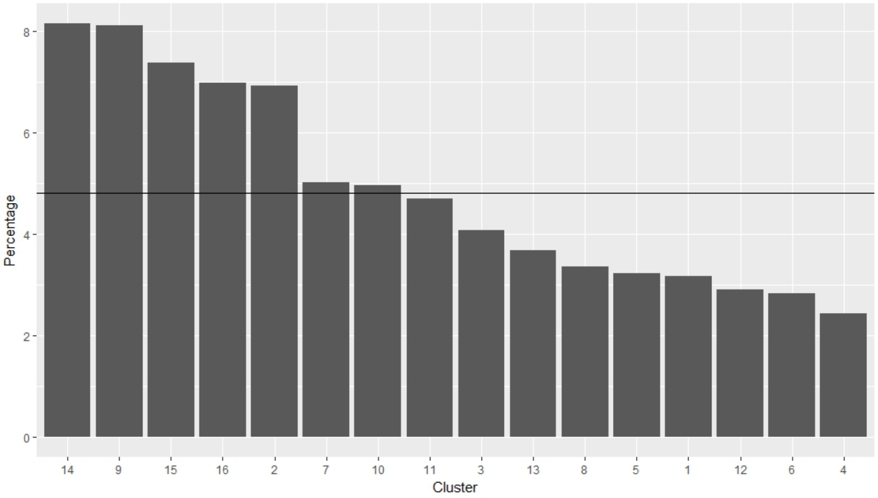


Figure 47: Percentage of each cluster which refused to give education level, ordered by percentage; and the percentage of the sample shown by an intersecting line.



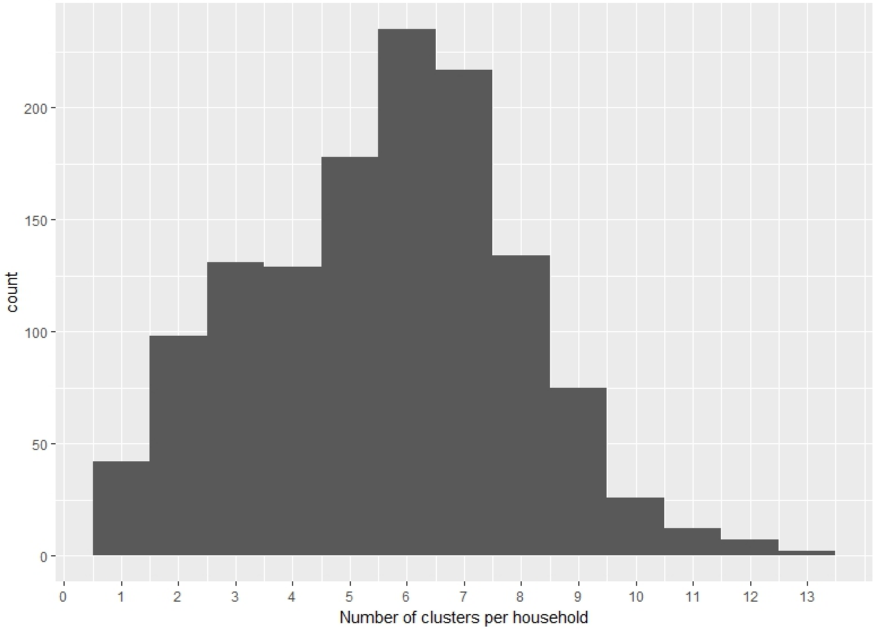
7.4 Movement between clusters

In training the SOM and the above analysis, each household had up to 33 different days as input vectors. There were no constraints when training to keep households in the same cluster, or to set a minimum number of households in each cluster. It is possible that a household could feature in all clusters, or that a cluster could be composed of just one household - the latter of these is not true due to least number of input vectors in a cluster being 388. McLoughlin *et al.* (2015) clustered the electricity load profiles in a similar way, and set the cluster (they refer to a cluster as a profile class) that each household occupies for the majority of the time as that household's overall profile class.

To find out the number of clusters each households featured in, the R code used:

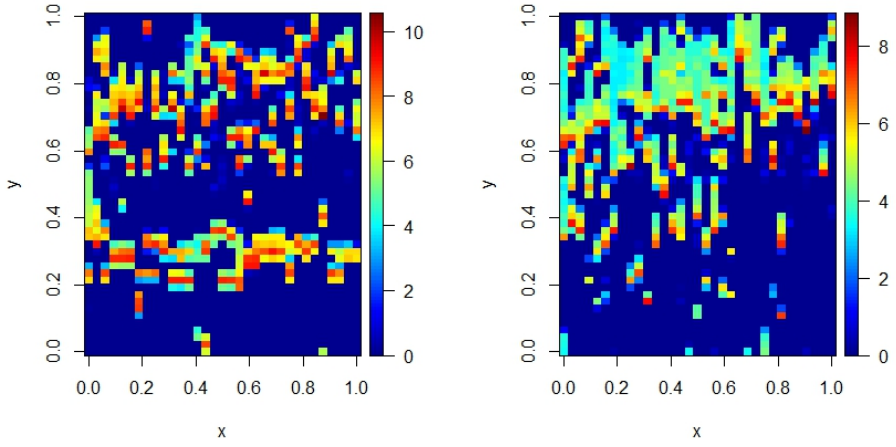
```
library(dplyr}
##How many times does each household go to different cluster
hh_diff_clusters=som_survey_m3b %>% group_by(ID,cluster)
%>% summarise(n=n())
##How many different clusters per household
hh_cluster_count=hh_diff_clusters %>% group_by(ID,Emp2,SC3,Edu2,FP)
%>% summarise(clusters=n())
##Over 33 days, how many different clusters does each household
##get attributed to
ggplot(data=hh_cluster_count, aes(clusters)) + geom_histogram(binwidth=1)
+ labs(x = "Number of clusters per household")
+ scale_x_continuous(breaks=seq(0,13,1))
```

Figure 48: Histogram showing the number of clusters each households featured in over 33 days.



The number of clusters per household is plotted in figure 48. There is a slight left-skew, with a mean of 5.6 and median of 6. This suggests that most households are relatively consistent with their gas use, though from this level of analysis it is not known how many days are spent in each cluster (for example it could be 28 days in one cluster, and 5 different days in 5 different clusters; or 5-6 days each in 6 different clusters). No household featured in all 16 clusters, but two households did feature in 13 clusters each. Using the 2D overhead plot that was showcased in subsection 4.4, these are plotted in figure 49 and it can be seen that there is variation day-to-day in both the on and off times, and the times of highest usage. For contrast, there were 42 households which were allocated to just one cluster across all the input vectors, and two are shown in figure 50, both showing very consistent use in terms of periods and magnitude.

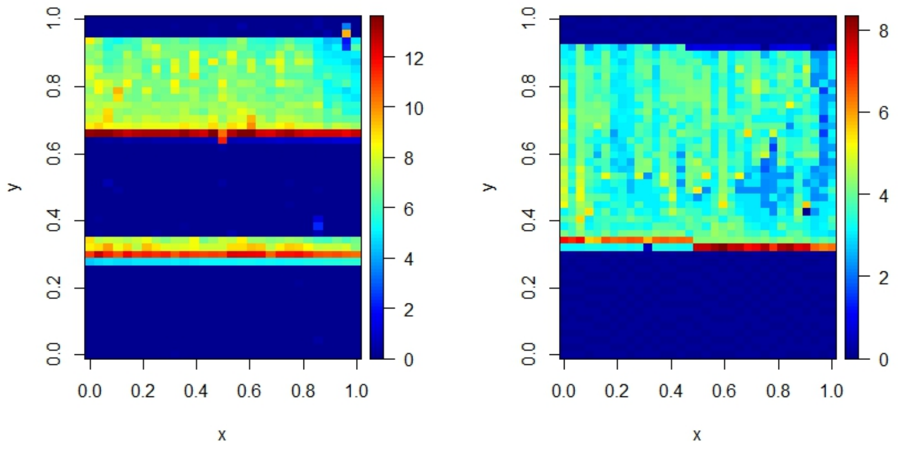
Figure 49: 2D plot showing gas consumption in the training data, for two households allocated to 13 different clusters.



ID: 1247

ID: 1520

Figure 50: 2D plot showing gas consumption in the training data, for two households allocated to one cluster for all input vectors.



ID: 1329

ID: 2187

The final piece of analysis takes what McLoughlin *et al.* (2015) did as part of their research, and finds for each household which is the most common cluster it is allocated to. This is a useful base for further research into what external factors might cause a household to move away from their most common behaviour. The R code to do this is as follows; a small random number is generated for cases where two clusters feature the same number of times for a household:

```
##Take most common cluster for each customer
##Some households are in two or more clusters the same amount of times
##Generate random number for each row and take the maximum of these
set.seed(25817)
hh_diff_clusters$rdm=runif(nrow(hh_diff_clusters), 0,0.01)
hh_diff_clusters$nrdm=hh_diff_clusters$n+hh_diff_clusters$rdm

hh_cluster_max= hh_diff_clusters%>%group_by(ID)
%>%filter(nrdm==max(nrdm))%>%arrange(ID)

###How many in each cluster
cluster_hh_count = hh_cluster_max%>%group_by(cluster)
%>% summarise(households=n())
```

Table 7: Counts of input vectors and households per cluster.

Cluster	% of input vectors	% of households	Difference
1	10	9.1	-0.9
2	5.4	6.1	0.7
3	6.7	6.1	-0.6
4	9.3	11.7	2.4
5	1.9	1.7	-0.2
6	1.2	1.2	0.0
7	19.6	24.7	5.1
8	1	0.7	-0.3
9	2.2	2.2	0.0
10	2.4	1.2	-1.2
11	7.6	5.2	-2.4
12	6.4	4.4	-2.0
13	7.2	4.3	-2.9
14	1.2	1.1	-0.1
15	12.9	14.8	1.9
16	5.1	5.4	0.3

Compared to the allocations of the input vectors of the SOM, the households per cluster is similar, see table 7, though the largest three clusters become larger.

7.5 Summary

There is definitely evidence that clusters can be used to highlight households who may be at risk of fuel poverty, or at least demonstrate factors associated with fuel poverty, namely employment status. The focus has been on those on the left-hand side of the plots, but those right-hand side of the plots are also important - the clusters with very low rates of each indicator, as they show where risk of fuel poverty is low.

Supplementary research could use the cluster information as an input in further analysis, such as Anderson *et al.* (2017) did, using a logistic regression to to estimate the probability a household being at risk. The analysis from this section can be used to consider to a certain day, the most common clusters, or how consistent a household is as inputs.

Using a similar approach as in this section, other factors of fuel poverty could be applied, some of which could be inferred from the CER survey data. A key factor of fuel poverty is the energy efficiency of the housing stock, and within the survey there are questions about house age, house type, and efficiency measures in place, for example.

8 Discussion

This dissertation has examined how households consume gas throughout the day, and methods to group households based on their consumption profiles. The data used for the analysis was provided by the Commission for Energy Regulation (CER) and the Irish Social Science Data Archive (IISDA).

Previous work has been done in grouping households based on electricity load profile, a lot of the work using the CER data, but not so for grouping based on gas consumption. This dissertation picked methods used for the electricity data and applied them to the gas data; a limitation found in the literature review was a lack of justifications for decisions - such as why Beckel *et al.* (2014) chose to use 10 principal components; or, how parameters were chosen - such as specifying learning rates of Self-Organising Maps. For these reasons it is not possible to assess directly how clustering gas consumption compares to clustering electricity consumption, as different parameters may have been chosen.

In general, electricity and gas are consumed differently by households, with electricity tending to have a non-zero base load powering appliances, and gas tending to be off for large portions of the day. It would be interesting to assess how reliable certain algorithms and learning methods are given this difference in base load.

Limitations were found with the R package `kohonen` in that it was a “black box”; it would have been good to examine the changing shape of the maps as they went through the training process as the neurons updated, and to have more control over plots. There is a concern that the maps used in the analysis had not been trained as fully as possible; the process was run on a single computer, and was limited in processing power and disk space. With more time and resource, a self-programmed SOM function could be developed, and make use of a distributed system for more iterations of training data. With more programmability it would be easier to compare changes in parameters, or changes in seed to test the robustness of the models.

Choosing 16 clusters was perhaps too many, as five of the smallest clusters showed similar mean consumption, and could have been grouped together, or allocated to other larger clusters. A penalisation could be applied to give favour to simpler models (as is done with the Akaike Information Criterion when choosing variables); but, going to too small a number of clusters could loose some of the individuality that has been seen. If the objective is to identify a certain type of household, it could be expected that they show different consumption profiles to the majority; to reduce the number of clusters by too much and these could be lost. For example, the cluster showing over twice the mean rate of unemployment was one of the smaller clusters.

Analysis was done by Chelmis *et al.* (2015) and McLoughlin *et al.* (2015) on how consumption changed over time for a certain building, or certain household. This was briefly addressed in subsection 7.4 which highlighted some sample household which moved between many different clusters over the three months but more could be done. Bringing in external factors such as the weather, or fuel price (though Harold *et al.* (2015) found this to be stable) and seeing if this causes changes in consumption could be the next step in this analysis.

References

[1] Anderson, B., Lin, S., Newing, A., Bahaj, A., James, P. 2017. Electricity consumption and household characteristics: Implications for census-taking in a smart metered future. *Computers, Environment and Urban Systems.* **63**, pp58-67.

[2] Beckel, C., Sadamori, L. and Santin, S. 2012. Towards Automatic Classication of Private Households Using Electricity Consumption Data. *In: Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings.* pp169-176.

[3] Beckel, C., Sadamori, L. and Santin, S. 2013. Automatic Socio-Economic Classication of Households Using Electricity Consumption Data. *In: Fourth International Conference on Embedded Sensing Systems for Energy Efficiency in Buildings.* pp169-76.

[4] Beckel, C., Sadamori, L., Staake, T. and Santin, S. 2014. Revealing household characteristics from smart meter data. *Energy.* **78**, pp397-410.

[5] Brunsdon, C. and Singleton, A. 2015 *Geocomputation: A Practical Primer.* London: Sage Publications Ltd.

[6] Chelmis, C., Kolte, J. and Prasanna, V. 2015. Patterns of Electricity Demand Variation in Smart Grids. *Working paper, University of Southern California Engineering Department.* Unpublished.

[7] Commision for Energy Regulation. 2011. *Smart Metering Information Paper. Gas Customer Behaviour Trial Findings Report.* [Accessed 11th July 2017]. Available from: http://www.cer.ie/docs/000340/cer11180(ai).pdf.

[8] Commision for Energy Regulation. 2011. *Electricity Smart Metering Customer Behaviour Trials (CBT) Findings Report.* [Accessed 11th July 2017]. Available from: http://www.cer.ie/docs/000340/cer11080(a)(i).pdf.

[9] Council decision 85/8/EEC of 19 December 1984 on specific community action to combat poverty.

[10] Davies, L. and Bouldin, D. 1979. A Cluster Seperation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **1**(2), pp224-227.

[11] Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC.

[12] GOV.UK, 2017. *Annual Fuel Poverty Statistics Report 2017 (2015 data)*. [Accessed 15th July 2017]. Available at: https://www.gov.uk/government/statistics/.

[13] Harold, J., Lyons, S. and Cullinan, J. 2015. The determinants of residential gas demand in Ireland *Energy Economics.* **51**, pp475-483.

[14] Irish Social Science Data Archive (ISSDA). 2012. *Data from the Commision for Energy Regulation (CER) - smart metering project.* [Accessed 16th May 2017]. Available at: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[15] Kohonen, T. 2001. *Self-Organizing Maps.* 3rd ed. Berlin: Springer.

[16] Mardia, K.V., Kent, J.T. and Bibby, J.M. 1979. *Multivariate Analysis.* London: Academic Press Inc.

[17] McLoughlin, F., Duffy, A. and Conlon, M. 2015. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy.* **141**, pp190-199.

[18] Moore, R. 2012. Definitions of fuel poverty: Implications for policy. *Energy Policy.* **49**, pp19-26.

[19] Ofgem. 2017. *Transition to smart meters.* [Online]. [Accessed 4th August 2017]. Available at: www.ofgem.gov.uk/gas/retail-market/metering/.

[20] Ovo Energy. 2017. *What is an in Home Display (IHD)?.* [Online]. [Accessed 4th August 2017]. Available at: https://www.ovoenergy.com/ovo-answers/topics/smart-technology/smart-meters/.

[21] The R Project for Statistical Computing. 2017. *Package kohonen.* [Accessed 6th August 2017]. Available at: https://cran.r-project.org/web/packages/kohonen/.

[22] Saitta, S., Raphael, B. and Smith, I. F. C. 2007. A Bounded Index for Cluster Validity. *Machine Learning and Data Mining in Pattern Recognition.* pp174-187.

[23] Schmidt, C.R., Rey, S.J. and Skupin, A. (2011) Effects of irregular topology in spherical self-organising maps. *International Regional Science Review.* **34(2)**, pp215-229.

[24] Sustainable Energy Authority of Ireland. 2009. *Effectiveness of Domestic Energy Efficiency Programmes. Fuel Poverty Action Research Report 1:Executive Summary.* [Accessed 4th August 2017]. Available at: http://www.seai.ie/Grants/