Machine Learning with Domestic Energy Use Data

Sam Stern (s1134468)

March 23, 2015

Abstract

Contents

Chapter 1

Introduction

1.1 Introduction

Amidst international pressure on countries to reduce their carbon footprints [1] and the British public's becoming increasingly frustrated by rising energy bills with little to no explanation as to the reasons behind the increases [1], the UK Government is currently executing a plan to distribute smart meters to households across the country by 2020. Smart meters, which measure a household's gas and electricity consumption in real-time and regularly communicate the readings directly to the utility companies, are expected to help households reduce energy usage by displaying how much energy is actually being used. They should also increase transparency in the household's energy bills by eliminating the need for monthly meter readings and estimations by the energy providers. Instead, the energy companies will be sent documented accountings of their customers' real consumption, and as a result, will be able to invoice more accurately.

While there has generally been strong support for the smart meter program, there has also been resistance to the campaign, with fears that the energy companies will use the information as an opportunity to raise their customers' bills and increase their own profits [2]. Perhaps more interestingly though, and therefore the focus of this project, are concerns that have been raised regarding the security risks associated with measuring and storing energy consumption data [3] [4]. Specifically, how much other information about a household can be inferred from energy consumption readings?

In looking to answer whether these fears are well-founded, the aim of this project is to explore whether (and to what extent) it is possible to construct features that predict detailed personal information about a household based on its energy consumption readings, and if so, if the results would be reliable. Breach of privacy issues would include whether such intrusive knowledge of household habits could effectively be exploited for targeted marketing or advertising campaigns, Big Brother-type government "watching", or equally if not more maliciously, for timing burglaries or other crimes.

Using electricity consumption information collected by the Household Electricity Survey (HES), a DEFRA[1] sponsored national survey of energy use collected over a period from 2010 to 2011, classification models are created to predict two properties of households: (1) The presence (or absence) of children and

[1]Department for Environment, Food and Rural Affairs

(2) the Ipsos MORI social grade of the chief income earner. These properties are chosen because, of all the information gathered by the HES survey, they would logically be of interest to someone who might wish to intrude on a household.

This project has 3 main components:

1. Clean the data and create a database that stores each households energy-use information and any other relevant data;
2. Extract useful features from the data that can be used as inputs to a classification model;
3. Predict household properties using supervised learning methods.

It should be noted that although the terms *electricity, power* and *energy* are not synonymous, within the context of this paper, they all refer to the electrical power consumed by a household and are therefore used interchangeably.

1.2 Smart Meters

1.3 Related Work

Particularly in recent years, an increasing number of studies have applied machine learning and data mining techniques to model and analyse domestic electricity consumption. This field of research is of particular interest to energy providers as understanding who their clients are and how and when they use energy lets the providers optimise their resourses (providing more power during peak times and less during periods of low demand), and create and market products to specific client groups. The work done using household energy data can be broadly separated into two categories. Either, only consumption data is analysed to categorise households or relating it to additional information about the household. The first approach imposes fewer requirements on the data and has therefore been used in unsupervised tasks [5]. Chicco, for example, gives an overview of the clustering techniques used to establish suitable client groups for analysing electricity load pattern data [6]. Cao et.al also grouped consumers using electricity load profiles, however focusing on finding households with the same peak usage [7].

Another popualar problem is that of NILM (*non intrusive load monitoring*) which involves taking aggregated energy consumption data from households and disaggregating the consumption of the constituent appliances. Kolter and Jaakkola were able to use factorial hidden Markov models (FHMMs) to disaggregate energy readings with more that 90% precision on a synthetic data set [8]. A study performed by Lisovich et. al was able to use NILM to determine whether there are people present in a household, which appliances had been used (and when) as well as the sleep/wake cycle of households by looking at a dataset of households that had energy readings take at either 1 or 15 second intervals for between 3 and 7 days. Unlike the dataset used in this report, households that participated in the study performed by Lisovich et. a were more similar in the types of appliances they used (they didn't have electric showers or water heaters) [4].

Beckel et. al. used supervised learning methods to classify household properties of 4232 Irish households. Their work involved classifying the inhabitants, such as the age of the chief income earner , presence/absence of children and socio economic status of the household. They also looked to identify properties of the home itself, such as the number of appliances, the number of bedrooms and the type of cooking facilities [5]. While much of this of the work presented in the report overlaps with that done by Beckel et. al, we consider a different set of classifiers (random forest and logistic regression) as well as another class of features taken from the time-frequency transform of the data. Additionally, the study builds models that include features not given by the smart meter readings consumption to improve performance, which is not done here. Finally, McLoughlin et al., using the same dataset as Beckel et. al. explored correlation between electricity consumption data and household characteristics and investigated methods for clustering households based on their energy use.

1.4 This Project

Chapter 2

Data

2.1 Overview of the HES Dataset

The data used in this project comes from The Household Electricity Survey (HES), a study sponsored by DEFRA to monitor the electrical power demand and energy consumption of individual households in England over the period May 2010 to July 2011 [9]. The aim was to identify and catalogue the range and quantity of electrically-powered appliances found in a typical home, understand households' frequency and patterns of electricity usage, and collect 'user habit' data that emerges from recording a range of appliances [10].

The HES study monitored 250 households, of which 26 were observed for one year with the remaining 224 monitored for roughly one month. Not every household had the same number of appliances being monitored. The number was in the range of 13 to 85 appliances per home. When aggregated, (as outlined in section 2.2), the result could considered an estimate of a mains reading. Depending on the household, measurements were either taken in 2 or 10 minute intervals with units of deci-Watt hours (0.1Wh).

In addition to data regarding the appliance types and data readings, participating households also kept diaries of how they used their main appliances and provided supplemental information about the household, such as, the number of occupants, employment status, Ipsos social-grade and whether there were children present in the household.

2.2 Extracting the Data and Pre-Processing

As explained in Section 2.1, electricity readings of individual appliances and sockets were taken for each household (as opposed to total energy consumed by the household, as was required for this project). The HES study recorded measurements for the 250 possible appliances that a household could have (giving values of 0 to appliances that were not present in a household). The resulting raw data was held in large csv files with a significant number of redundant entries. In order to use the data to perform data mining, numerous pro-processing steps needed to be performed as explained here. Each of these steps were performed by writing python scripts with embedded SQL.

The first step in pre-processing the data was to create a MySQL database and import the appliance readings into a table. Cambridge Architectural Research

Ltd [11] provided additional files that mapped which appliances needed to be aggregated for each household in order to produce an estimate for the mains reading. This was often not simply the sum of all appliances readings. A table was therefore created for every household where each row contained the aggregated electricity measurements for a given date and time.

250 households in England participated in the HES study, a relatively small number for a machine learning task as there might not be enough data to build models that accurately sample the population. To help account for this, the 26 households that were monitored for an entire year were split into 12 instances that could be treated as separate households, resulting in an additional 281 household instances. While this does not create a more diverse group, it does add more instances to train, validate and test a classifier with. To avoid overfitting the classification models to the data, all instances from a given household were either in the training or test set, but never both.

Next, the inconsistency in measurement intervals was accounted for. While some households reported how much energy they used in 10 minute intervals, others were measured in 2 minute intervals. To create consistency in the data, for the '2-minute households', every five intervals were summed so that all households had 10 minute granularity. This step was important since some consumption features, would have been affected by differences in measurement intervals. Once all households were represented in 10-minute intervals with units of 0.1Wh (deci Watt hours), each reading was then multiplied by 0.6 to convert the data to Watts.

The last stage in pre-processing was to ensure that each instance was of the same length. As will be discussed in chapter 3, temporal structure was observed both intraday and intraweek. Therefore, the time series instances were manipulated so that each had a length of 28 days and started on the same day of the week. This was done by performing the following steps:

1. Ensure that each household has an integer number of days by topping and tailing the data.

2. Find the mode day of the week that the data starts from (which was found to be Sunday).

3. For the households that do not begin on a Sunday, chop the top few days so that the data begins on a Sunday.

4. If the household's data is now less that 28 days, append days to the end until it is of the correct length. If it is possible, use the days that were chopped off in the previous step, otherwise, reuse a days worth of readings.

Figure 2.1 gives a visual example of the data is made to be of uniform length. As the readings start on a Thursday (day 5), the first three days are chopped off the top. Since the data is now less that the required number of days, days are either reused or, is possible, taken from the days that have been chopped from the top.

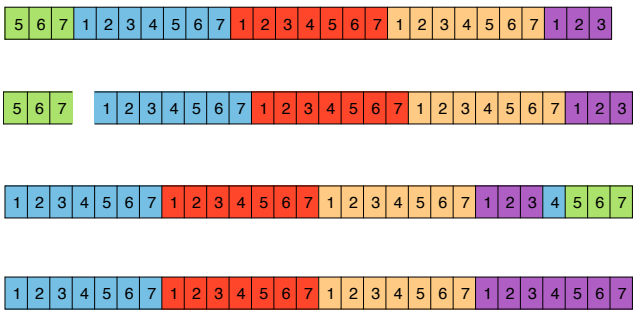


Figure 2.1

2.3 Household Classes

Each household that participated in the HES study completed a survey with questions about the building being occupied (such as the year the house was built), the household (such as the number of occupants) as well as the occupants' attitude towards climate change and energy consumption. The answers to these questions are used as labels for the households to perform supervised learning.

Social Grade	Description	Sample Size	% Sample	% Pupulation
A	High managerial, administrative or professional	33	6.4	4
B	Intermediate managerial, administrative or professional	95	18.3	23
C1	Supervisory, clerical and junior managerial, administrative or professional	197	38.0	29
C2	Skilled manual workers	128	24.7	21
D	Semi and unskilled manual workers	34	6.6	15
E	State pensioners, casual or lowest grade workers, unemployed with state benefits only	32	6.2	8

Table 2.1

Tables 2.1 and 2.2 show the sample sizes for each class of the two classification problems being concidered in this project. The distribution of households over each of the classes in our sample is similar to the true distribution, which means that the empirical prior probability of each class is a reasonable estimate of the true prior probability. However, there is a significant imbalance in the classes, especially in the socio-economic classes. This result in bias in the classification models that will need to be considered when evaluating them.

Class	Sample Size	% Sample	% Pupulation
Children	187	36	39
No Children	332	64	61

Table 2.2

2.4 Discussion

After the data had been extracted from the csv files, pre-processed and imported into MATLAB, plots of the data were made in order to visually gain insight into how households used energy and increase domainn knowlege. Figures 2.2 and 2.3 are examples of how some of the households consumed energy. Both figures show the data gathered from the same households, but over different time periods. Studying these plots gives valuable insight into the households which is used later to aid in feature extraction, as well as ensure that the data appears reasonable.

In figure 2.2, the first thing to be noted is that the consumption is not smooth. There are sharp peaks that vary in height, which can be used to make assumptions about which appliances are being used. For example, many of the peaks are around 1kW, which is roughly the amount of power used by a kettle. The next thing to note is that there is an obvious underlying daily repetition. The household tends to use more electricity at night than it does during the day time. Finally, it can be seen that the energy consumption on weekends is slightly different than that of week days, particularly, there are short periods of abnormally high electricity on Saturdays and Sundays which are observed less frequently during the week. To see this, note that both figure 2.2 and 2.3 start on a Sunday, and that each 'wavelet' is one day long.

It is a result of these observations that the data was made to be four weeks long. Ensuring that each day of the week appears exactly 4 times for each household means that features such as the total energy used is not influenced by which days of the week are present.

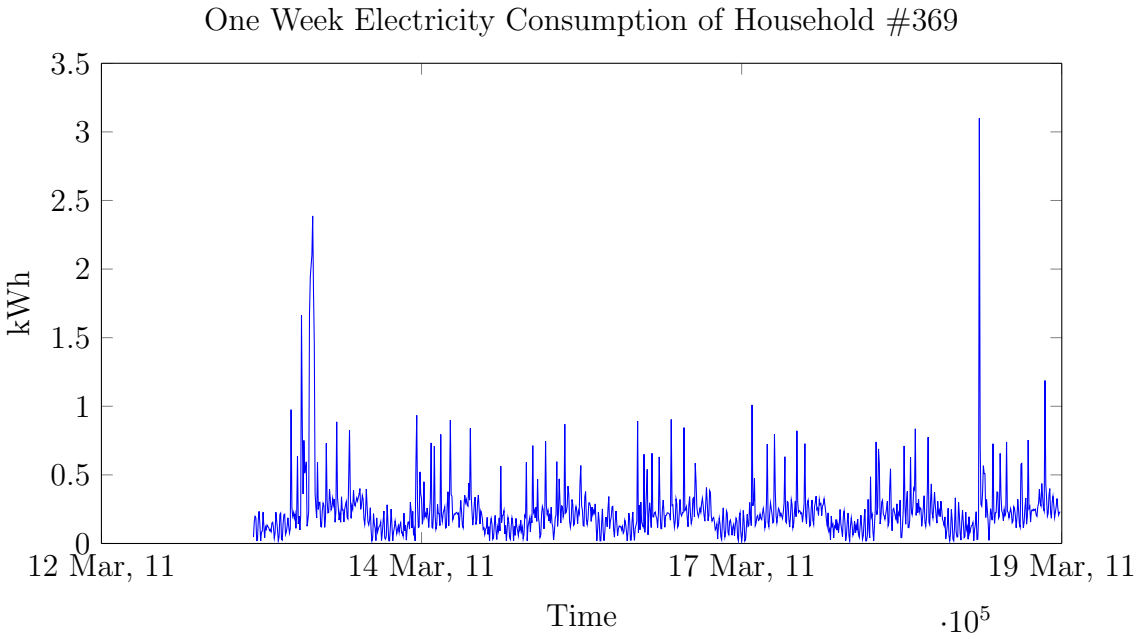


Figure 2.2: The electricity use of household No.369 over a week after being pre-processed

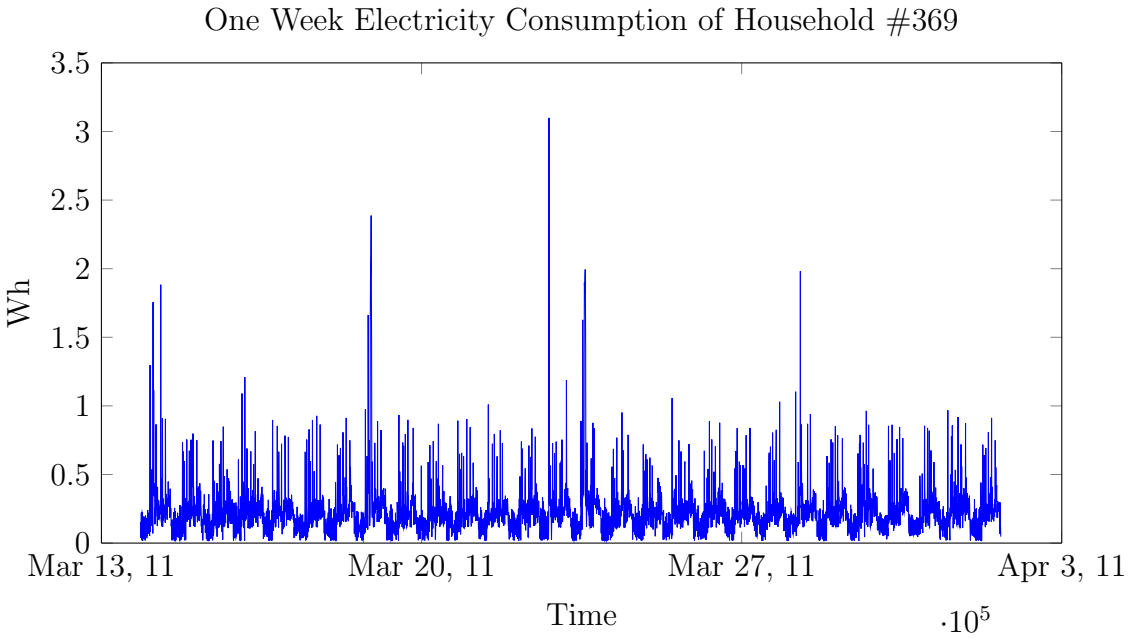


Figure 2.3: The electricity use of household No.369 over a four week period after being pre-processed

2.5 Issues

As with any study, numerous issues arose with the data that needed to be acknowledged. While some of these are simply the result of environmental issues, others relate more to the way the study was conducted. While those most relevant to this project are outlined below, others can be found in the CAR report [10]

The first problem with the data is the number of households that participated in the HES study. Only 250 individual households took part. Comparing this to the 4232 households that took part in the CER[1] (Commission for Energy Regulation) study of household electricity consumption in Ireland (used by Beckel et. al. and McLoughlin [12, 13, 5, 14]), it is less likely that the results generalise as well, particularly for the multi-class classification problem where there are as few as 32 households in a class. Moreover, only households in England took part in the study, and all of them were owner occupied. This means that the subset of homes considered in the HES study is not fully representative of the UK as a whole, since 84% of people in the UK live in England and only 64% of homes in England are owner occupied [15]. The aim of this project, however, is to determine *whether it is possible* to infer household properties from a household's electrical power consumption, not to build a classifier that can be used to infer British household properties from smart meter data. The distinction being that this project is a proof of concept and looks at whether information about a household is contained in they energy use patterns rather than to build a commercial product. Therefore, the quality of the sample population (or lack thereof) households is not detrimental to the aims of this project.

A more bothersome issue is the quality of data that gathered during the HES

[1] www.ucd.ie/issda/data/commissionforenergyregulationcer/

study. While household shown in figure 2.3 shows the characteristics of a typical home's consumption well, it is one of the 'better' households in the sample. There are many others that no not follow the same kind of trend such as those in figure 2.4. Either they do not have the same well-defined periodicity, or they may use significantly more (or less) energy than the average household. The task then becomes determining whether these discrepancies are reasonable differences that could be attributed to differences between households, or whether they are the result of poorly executed data collection. Since the HES study involved recording individual appliances, rather than the mains reading of a household, the total energy consumed by each household can not be given with certainty as it is not known if all appliances and sockets in a household were recorded. This is made evident in figure 2.4 where household #75 is always using at least some energy while household #121 sees its consumption drop to 0. Household #75 is a better estimate as it is reasonable that there will always be a small amount of electricity used by a home since the appliances are not 100% efficient and leak electricity.

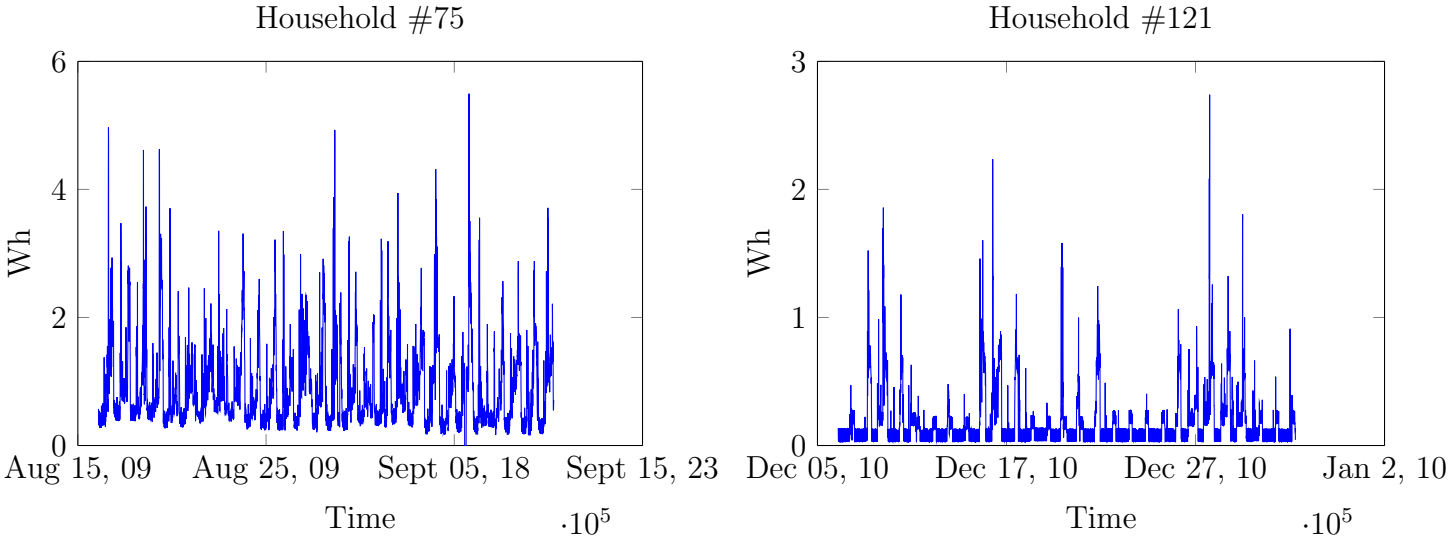


Figure 2.4: The electricity consumption of two households that do not show the same pattern of consumption as other households

Table 2.3, is taken from the Center for Sustainable Energy [16] and shows how much power various appliances use. Comparing these values to the data from the meter readings shows that it is reasonable to see a household use anywhere from 50W to upwards of 15,000W at a time. However, it is also noted that some of the expensive appliances (which will have large effects on a households consumption) will not be present in all households. These include electric cookers, electric showers, electric heaters and tumble driers. Only 38 of the 250 households used electric water hearing. While it can be expected that these factors will impact a households consumption and having knowledge of these would aid in classifying households, this is left to further work as the so called *disagregation problem* (see section 1.3) is a popular topic of research in and of itself.

While visualising the data, it was noted that some households had periods where their consumption vanished for several days. It was manually determined if these were instances of the household simply being unoccupied for a time, or if it was a case of erroneous data. Households were only discarded if the amount

Table 2.3: **Energy used by various household appliances**

Appliance	**Rating**		**Appliance**	**Rating**
Immersion heater	3,000W		Fridge	40-120W
Electric fire	2,000-3,000W		Fridge-freezer	200-400W
Oil-filled radiator	1,500-2,500W		Freeze	150W
Electric shower	7,000-10,500W		Electric mower	500-1,500W
Dishwasher	1,050-1,500W		Electric drill	900-1,000W
Washing machine	1,200-3,000W		Hairdryer	1,000W
Tumble dryer	2,000-4,000W		Heating blanket	130-200W
Toaster	800-1,500W		Games console	45-190W
Kettle	2,200-3,000W		Laptop	20-50W
Microwave	600-1,500W		Desktop computer	80-150W
Oven	2,000-2,200W		Tablet (charge)	10W
Grill/hob	1,000-2,000W		Broadband router	7-10W
LCD TV	125-200W		Smart phone (charge)	2.5-5W

of time where their consumption was 0kWh was a significant proportional of the total time for which they were being observed. If a household appeared to be on holiday (meaning their consumption patterns stopped but a small amount of energy was still being used by the house), then the data was kept. And if a given day's readings appeared to be erroneous, then that day was discarded and was replaced by an equivalent day of the week.

The next factor that needed to be considered is the effects of weather, particularly the time of year. Colder temperatures and shorter periods of sunlight cause households to use more electricity in during colder months [17]. Although CAR was able to provide a document outlining which appliances need to have their readings adjusted to account for seasonal factors, these did not appear to be well reasoned and didn't include many of the appliances used by households. Since most households were recorded in the colder months between November 2010 and April 2011, and those that were measured for a year didn't appear to significantly change their consumption in the warmer months, season adjustments were disregarded.

Chapter 3

Feature Exploration and Extraction

3.1 Types of Features

When data mining in time series, it is usually not sufficient to consider each point in time sequentially. In addition to ignoring the high dimensionality of the data, it does not account for the correlation between consecutive values [18]. It is therefore beneficial to transform and aggregate the data in such a way as to reduce the dimensionality as well as capture differences in the consumption patterns between classes.

According to Beckel et. al[13], possible features that are interesting for classification of households based on energy consumption are: consumption figures, ratios, temporal properties, and statistical properties. Consumption figures represent the average, maximum and minimum energy consumption over some time period. Ratios are features that calculate the ratio between consumption different figures, and can capture relevant patterns that occur through different time intervals. Temporal features capture the first or last time some event takes place, the time at which the daily maximum or minimum occurs or any periodicity within the household's electricity consumption. Finally, statistical properties, such as variance or correlation, give insight into the consumption curve.

Numerous statistical methods presume that input data follows a normal distribution. Therefore, the HES data was visualized and compared against a normal quantile plot in order to find the right non-linear transformations [19] [20]. Figure 3.1 shows the normal quantile plot of the average standard deviation of a household on Mondays (left) and the logarithm of this feature (right). The linearity of the sample quantiles of the features (x-axis) versus the theoretical quantiles of a normal distribution (y-axis) implies that the transformed features are (roughly) normally distributed. These transformations are important for classifiers, such as k-nearest neighbour, which rely on the distance between samples based on their features.

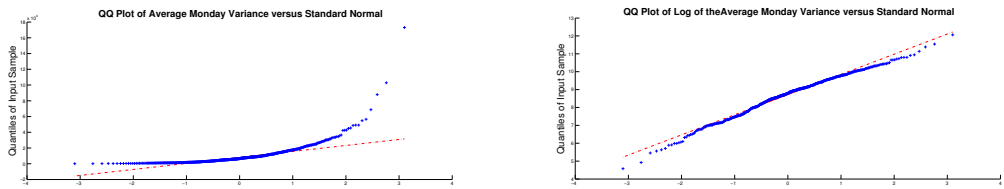


Figure 3.1

3.2 Creating Features

One method of extracting features is to compute as many different types as possible, compare them all and chose those that best discriminate the classes. Households can be further split into weeks, days and even hours. Consumption figures and statistical properties can then be measure for each of these intervals. While this method does provide more coverage and therefore a greater chance of finding the best features, it ignores any domain knowlege that we might have and is therefore potentially wasteful of the limited resources available to do the project.

Instead of creating features in an ad hoc manner, a more cost efficient approach was taken. Feature selection was done in the following way: 1) Assumptions were made regarding the distinction between classes (e.g., households with children use more energy overall). 2) Features were created to capture this distinction (e.g., the average energy over a 4-week period). 3) Tests were performed to evaluate the validity of the assumption. These tests varied in thoroughness as it was sometimes obvious from visualising the resultant features that they did/did not discriminate between classes. At other times, more sophisticated methods were used, as described in 3.3.

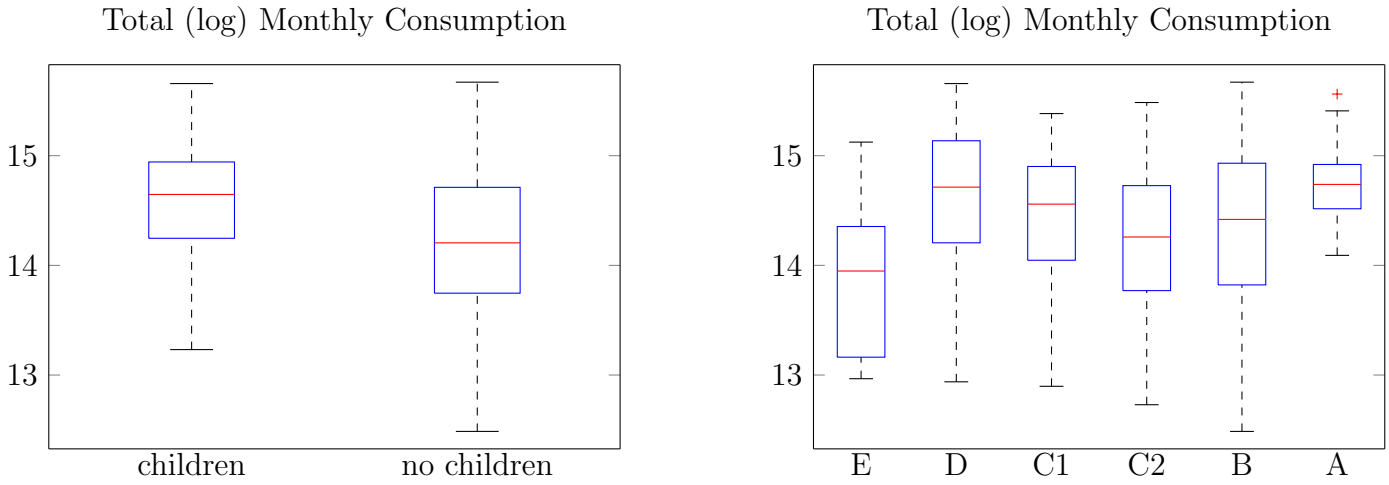
The remainder of this chapter describes features that were created from the energy reading data and justifies why it was assumed that they assumed would be able to discriminate between classes. The results of computing these features are then evaluated. Both classification problems (socio-economic classification and child classification) were considered when choosing features to evaluate.

Total Electricity

In visualising the data, it was noted that households had large differences in how much energy they used. While some households had a mean energy consumption rate of 1500 Watts per 10 minutes, others averaged as little as 65 Watts per 10 minutes; while one household consumed up to 19500 Watts in a 10 minute period, another never used more than 1190 Watts in the same time interval. To determine if these discrepancies can be attributed to different classes, the first feature that was explored was the total energy consumed within a given period of time. Since it was not known at this stage whether other factors, such as time of day, or day of the week, influence consumption, 28-day time frames were used to ensure independence of these factors.

Building a classifier using the total electricity as input assumes that some classes use more energy than others. This can be justified as there is a known

correlation between a household's disposable income and the amount of energy it uses [21].



(a) Total electricity used by households in a 28 day pe-riod, grouped by whether the household has chidren or not

(b) Total electricity used by households in a 28 day period, grouped by the IPSOS social grade of the household

Figure 3.2

Looking at Figure 3.2, it appears as though there is a difference in total electricity consumption between different classes. The left hand plot, which compares households with children against those without, shows that those with children do indeed tend to use more energy. The right hand plot, which compares total electricity, grouped by social grade, indicates that the highest socio-economic households do use more energy than those of the lowest social grade. It does not, however, distinguish well between intermediate social grades.

Average Daily Usage

As it has been established that some classes of households do indeed use more energy than others, it is worthwhile to dig deeper and determine whether there are any factors that influence these differences. With this in mind, the average energy used by each household for each day of the week was computed. This sort of feature explores not just if some classes use more electricity than others, but if the electricity consumption is dependent on the day of the week.

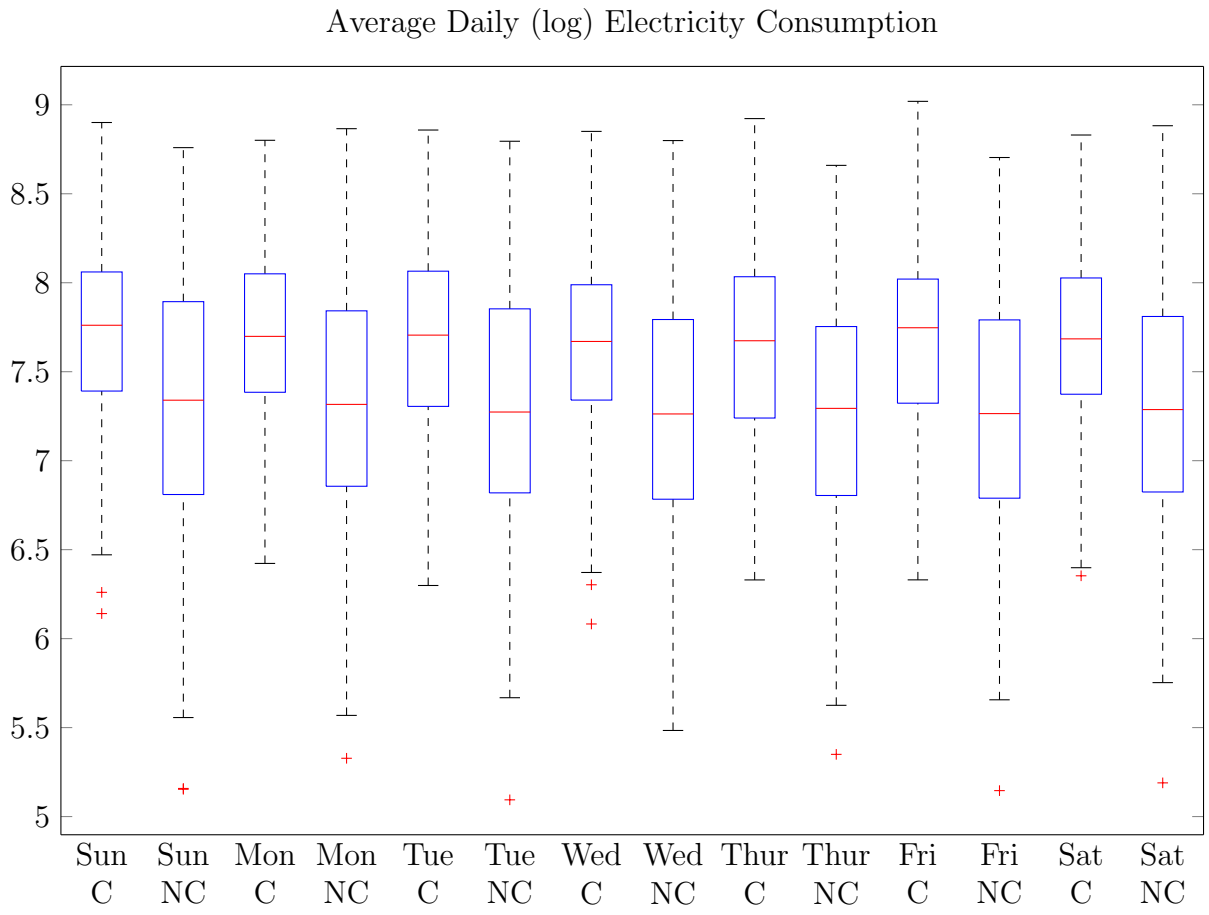


Figure 3.3: The average total energy used on each day of the week. Households are grouped by whether or not there are children present

Figure 3.4

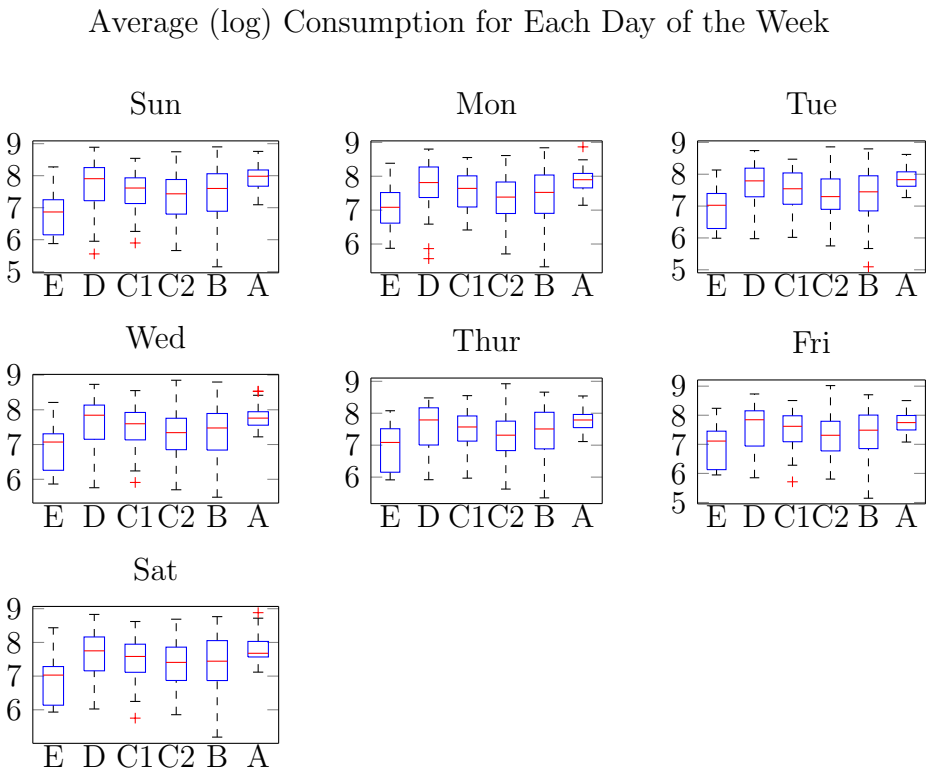


Figure 3.5: The average total energy used on each day of the week. Households are grouped by their IPSOS social grade

While Figure 3.4 does further show that households with children use more power than those without, it does not give any additional insight as to when, how or why this is the case. Households with children tend to use 1kW more electricity per day regardless of what day of the week it is.

Similarly, Figure 3.5, which compares the average daily usage of different socio-economic groups, does not offer any more insight into the differences between classes. There is no particular day where the differences in electricity consumption between classes is more visible than other days.

Average Part-Of-Day (APOD)

Going further, it could be that different classes use more or less energy at different times of the day. For example, lower socio-economic households might use more of their energy during the day than those of medium or high socio-economic status since they are more likely to be unemployed [22]. Similarly, it is reasonable to assume that the consumption gap between households with and without children might shrink when the children are at school and widen when they are at home.

Most schools days in England begin at 9:00 and finish between 15:00 and 16:00 [23]. Using this fact and the assumption that as children go to bed, the activity of the other members of the household will decrease and therefore electricity consumption will drop, then it is worthwhile to split each day into the following groups.

1. Morning (6:00-9:00): The time when members of the household would wake up and prepare themselves for work, school etc.

2. Daytime (9:00-15:00): The time that children are at school.

3. Evening (15:00-22:00): When a household can be presumed to be most active

4. Night (22:00-6:00): Depending on the type of household, people might be more of less active during this time period. For example, couples without children might stay up later.

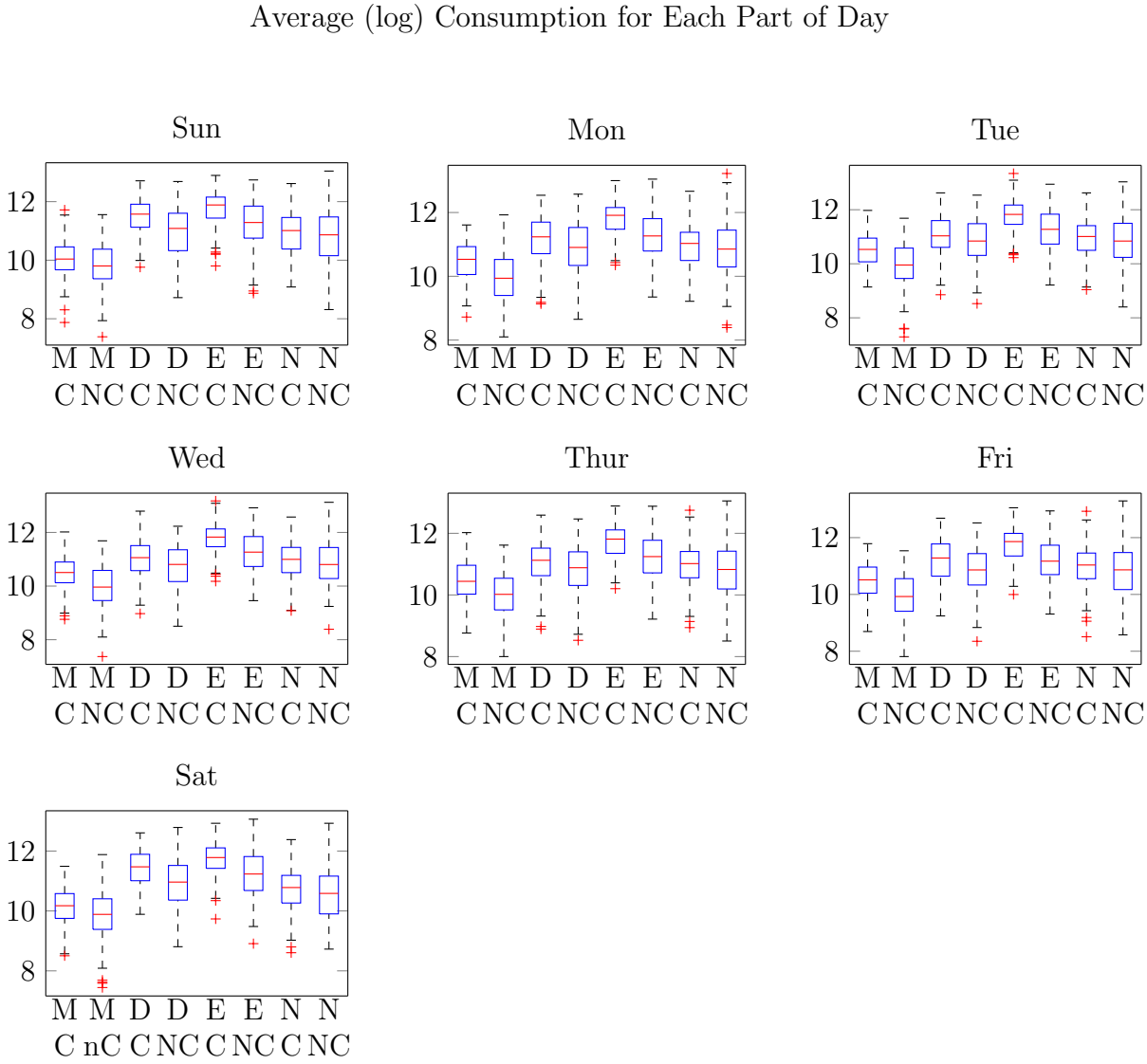


Figure 3.6

The data portrayed in Figure 3.6 indicates that energy use patterns are indeed different for households with and without children. We see that much of the difference in household electricity consumption can be attributed to household activity in the evenings (15:00-22:00), with the average household with children using 40kW more electricity during this period than households without. Furthermore, it can be seen that on weekday daytime (9:00-15:00, Monday-Friday) the two classes use similar amounts of electricity, however on Saturdays and Sundays, the gap widens and those with children tend to use more than those without.

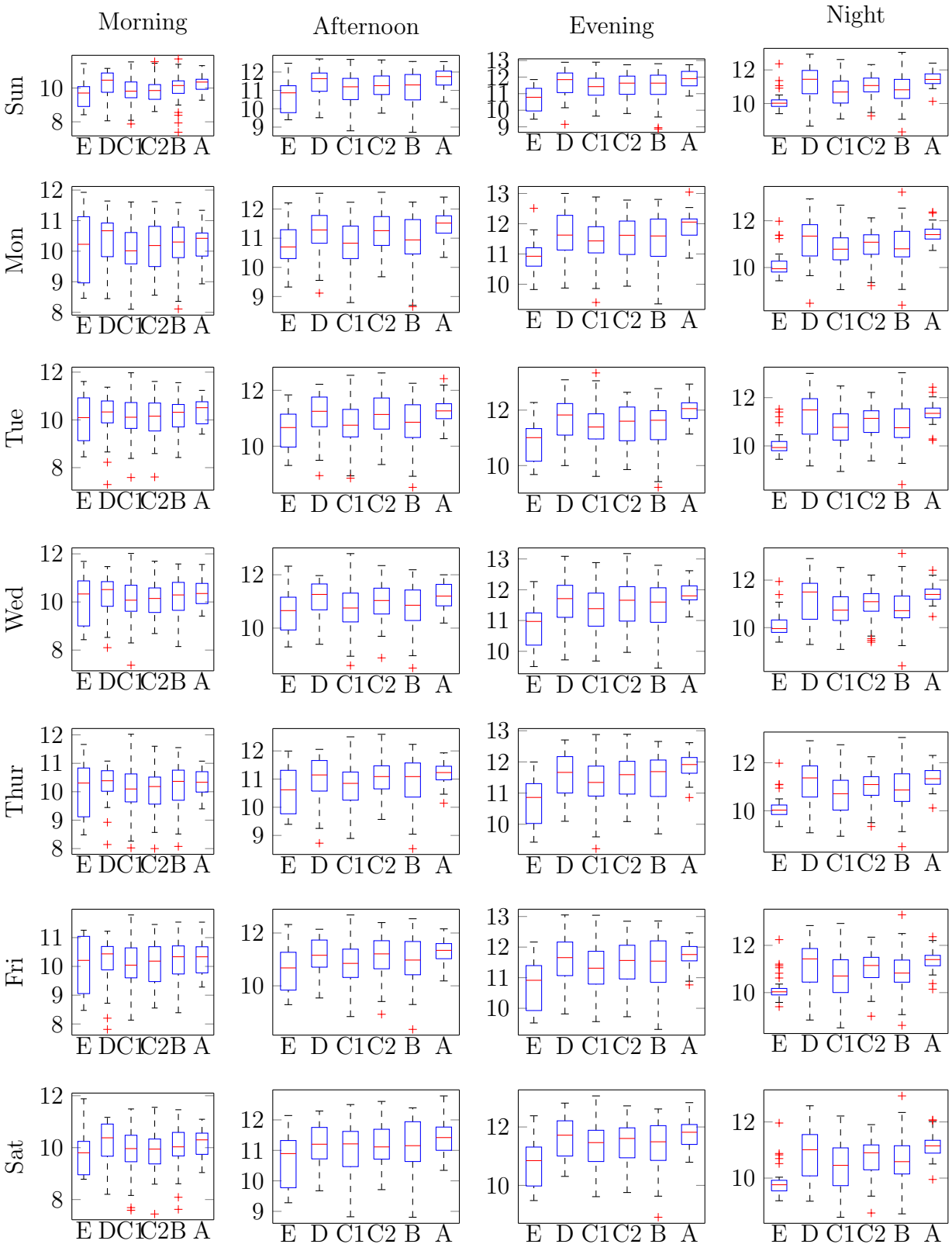
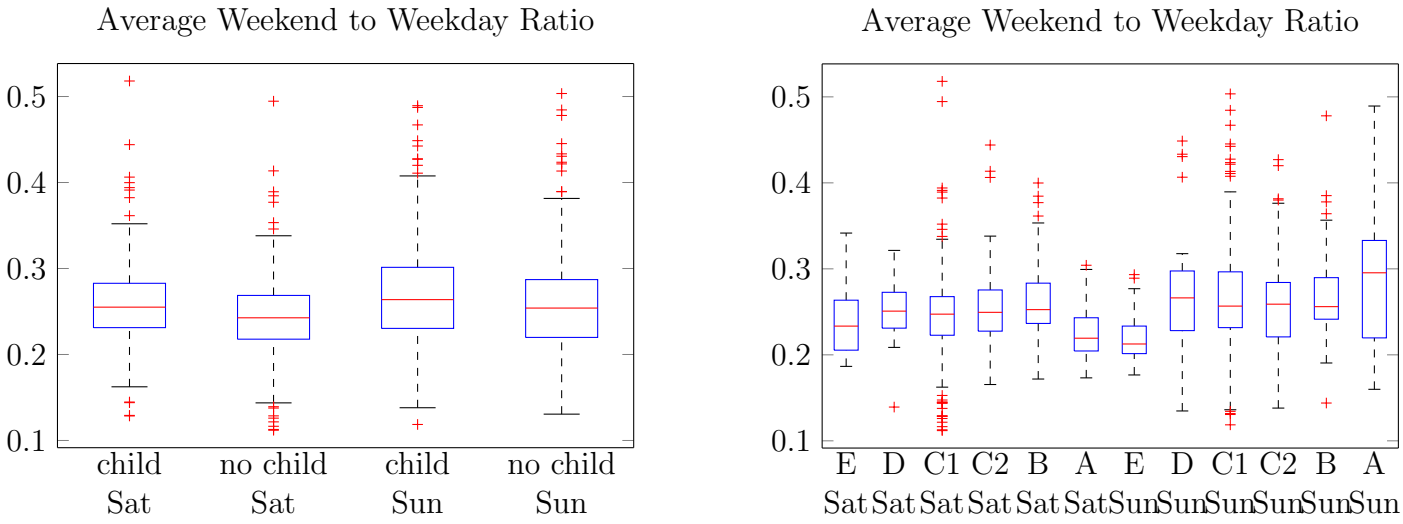

Average (log) Consumption for Each Part of Day
Morning
Afternoon
Evening
Night
Sun
Mon
Tue
Wed
Thur
Fri
Sat
E D C1 C2 B A

Figure 3.7

Figure 3.7 shows again the same results as the previously computed features. Households of social grade E appear to use relatively little energy at night than the households of other socio-economic groups, yet they seem to make up for it in the morning period where their consumption is more akin to the other groups. Households of group A show the opposite pattern, using more energy than others in the evenings but normal amounts (compared to the other classes) in the mornings.

Mean Weekday vs. Saturday and Sunday

In addition to looking at consumption features, ratios can also give insight into when a household is using its energy. Taking the ratio of the energy consumed on an average weekend day and an average weekday, one can determine if a household is using proportionally more of its energy during the week or at the weekend. The rationale being that households of social grades E,D and C2, whose chief income earner is either unemployed or a manual worker, is more likely to have a job that requires working on the weeknds than households of class C1,B or A who, given their supervisory and managerial professions, are less likely to work on weekends. It is therefore possible that the higher households will use a greater proportion of their energy on weekends than weekdays.



(a) The ratio between how much energy is used on the weekends and how much is used on weekdays. House-holds are grouped on whether or not there are childrenholds are grouped on their IPSOS social grade present

(b) The ratio between how much energy is used on the weekends and how much is used on weekdays. House-

Figure 3.8

After computing the ratio between weekend and weekday electricity consumption, classes seem to use similar proportions of their energy. And while Figure 3.8 suggests that households use more of their energy on Sundays than they do on Saturdays, this is independent of the both the household's socio-economic class and whether or not there are children present.

Variance on Weekdays

Thus far, the features that have been computed have been dependent on *how much* energy has been consumed. It is also worth considering how much volatility there is in the household's energy consumption. Continuing with the idea that energy usage will be different on weekdays versus weekends, the average daily variance for weekdays was computed separately from weekends.

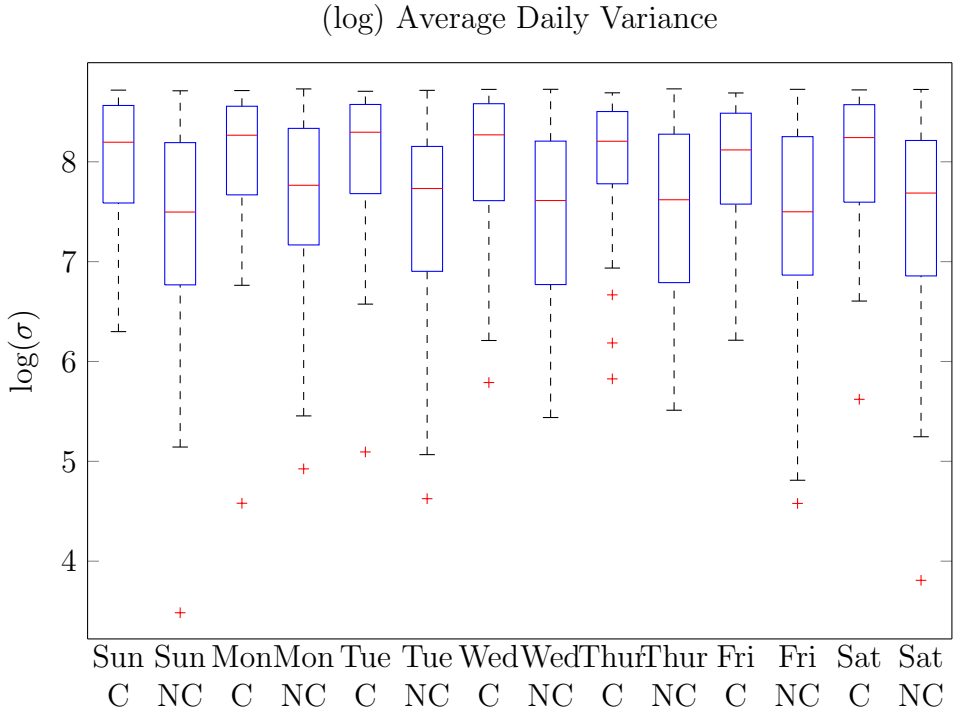


Figure 3.9: The variance of a household's daily consumption grouped by the day of the week and whether the household has children (C) or not (NC)

Although the average daily variance of households is volatile in and of itself, the results shown in Figure 3.9 indicate that the electricity use of households with children does tend to fluctuate more than those without children and therefore can give could be used to discriminate between households with and without children.

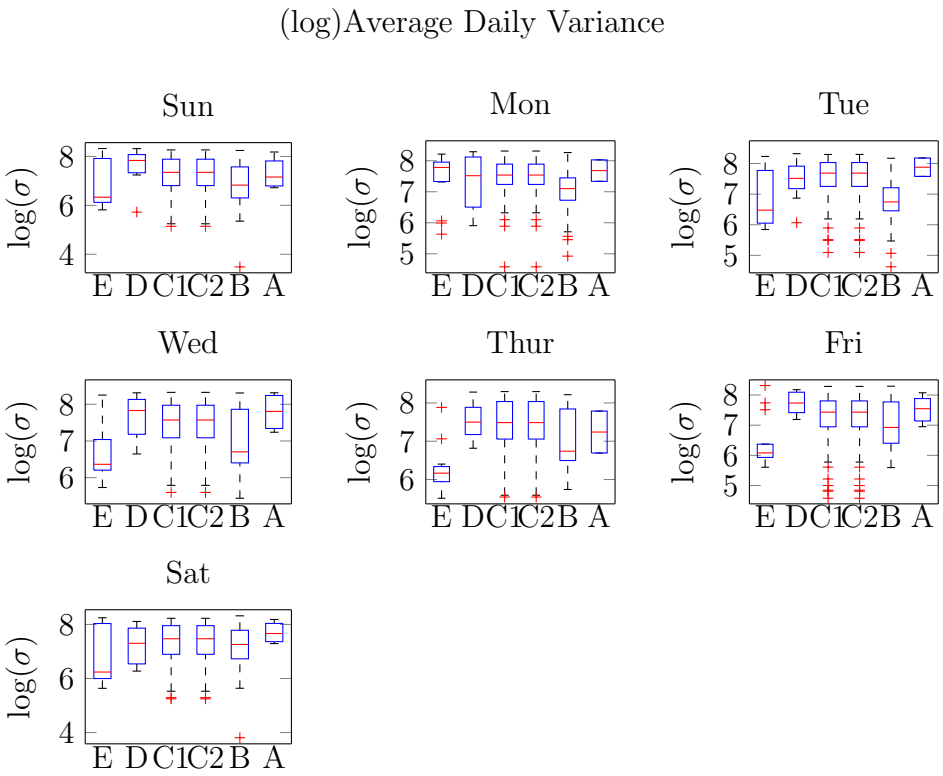


Figure 3.10: The variance of a household's daily consumption grouped by the day of the week and the Ipsos social grade

According to **??**, it is possible that the variance of a household's electricity consumption can be used to determine the socio-economic class of a household. It may be possible to seperate class E from the remaining classes based on the variance of a household's consumption on Thrusdays and Fridays, as well separating households of class B by the variance on Tuesdays.

It should, however, be noted that the range of features is itself relatively large and there are numerous outliers (represented by the red dots).

Correlation Between Weekdays

The average correlation coefficient between one weekday and every other weekday was calculated. Rather than using the 10-minute intervals, which appeared to be too granular to capture any covariance between days, electricity readings were summed into one-hour intervals.

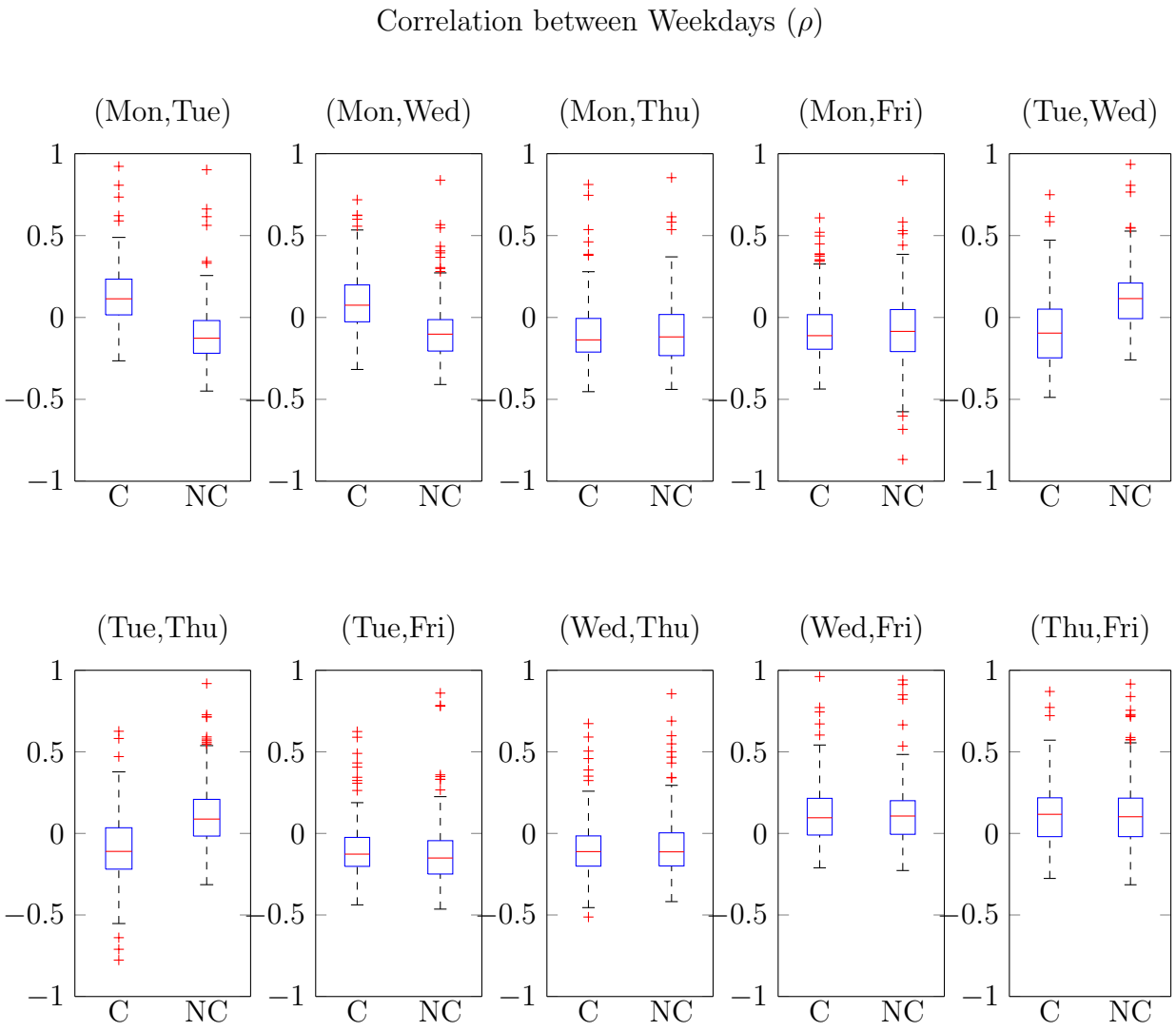


Figure 3.11: Average correlation coefficient between weekdays grouped by whether a household has children (C) or not (NC)

Looking at Figure 3.11, it appears that although the correlation coefficients are generally close to 0 (which means there is no correlation), there are differences between the two classes. Depending on which two days are being considered, the correlations of one class tend to be greater or smaller than that of the others. For example, it would appear that households with children demonstrate a slightly higher correlation between their Monday and Tuesday electricity use patterns than those without. Whereas for socio-economic classification, as depicted in Figure 3.11, the correlation between days does not result in features that separate classes.

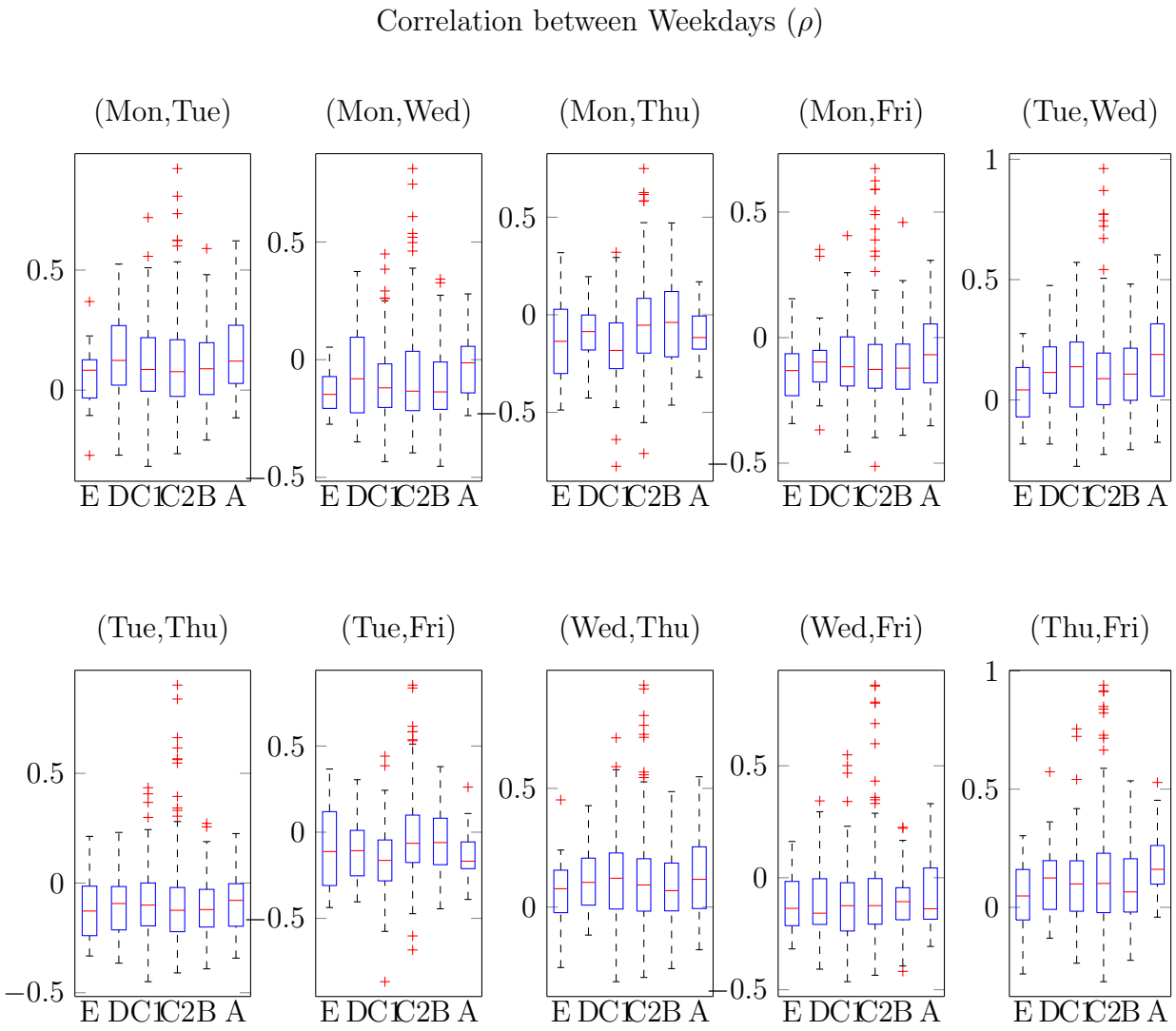


Figure 3.12: Average correlation coefficient between weekdays grouped by Ipsos social grade

3.3 Periodicity

Another approach used for feature extraction is to exploit the periodic consumption patterns exhibited by many households in order to search for temporal structures that are present in some classes but not in others. This method of feature extraction has been used successfully in previous studies involving forecasting and clustering. Methods outlined by Fabian Moerchen [18] for time series feature extraction are used to project each household's consumption into the frequency domain from which the most important frequencies are found. McLoughlin et. al. [14] showed in their research that temporal structure is present in household electricity consumption data and can be used to charachterise domestic energy demand.

Signal Smoothing

Before projecting the electricity consumption into frequency space, the Gaussian averaging operator was applied to each set of readings to filter noise whilst retaining the temporal structure of the data. Gaussian filtering (or Gaussian smoothing) is accomplished by convolving a time series with the Gaussian function. It can improve performance compared with direct averaging, as more structure is retained whilst noise is removed [24]. This is done because the time-frequency transformation used (the discrete Fourier transform method) has difficulty characterising small intervals of large electricity demand [25].

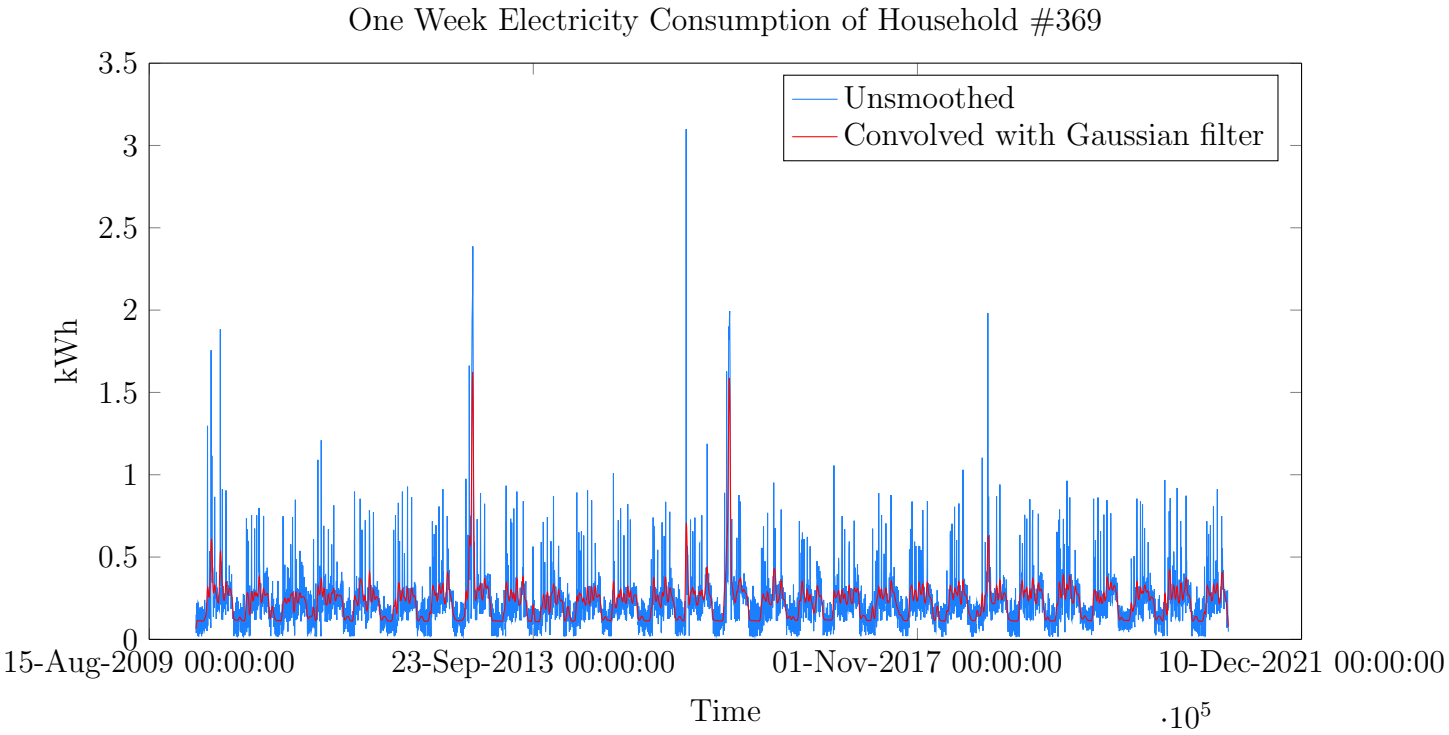


Figure 3.13: The electricity use of household No.369 shows that households may have both a daily and weekly pattern. The clusters of peaks represent individual days while the regions without peaks are the indicative of night time. Additionaly, the large spikes are observed roughly every seven days, on either Saturdays, Sundays or both. After applying the Gaussian filter, the time series maintains its temporal structure however the sharp peaks are smoothed, which would not be handled well by the Fourier transform

Fourier Transform

For uniform samples $[f(1)..., f(n)]$ of a real signal $f(x)$, the *Discrete Fourier Transform* (DFT) is the projection of a signal from the time domain into the frequency domain by

$$c_f = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} f(t) \exp \frac{-2\pi i f t}{n}$$

where $f = 1, ...n$ and $i = \sqrt{-1}$. The c_f are complex numbers and represent the amplitudes and shifts of a decomposition of the signal into sinusoid functions [18].

Issues do present themselves when using this method. As already mentioned, the Fourier transform measures global frequencies and the signal is assumed to be periodic. This assumption can cause poor approximations at the borders of the time series [18].

Energy Preservation

For l time series of length m, the DFT produces an $l \times m$ matrix C of coefficients, such that element $c_{i,j}$ is the j^{th} coefficient of time series i. In our case, since the number of households, $l = 519$, is small compared to the length of each time series, $m = 4032$, the number of coefficients must be reduced in order to minimise redundancy, noise and computational time. According to Moerchen [18], the best subset of k columns is found by selecting those that optimize energy preservation E, defined as

$$E(f(t)) = \sum_{j=1}^{m} a_j c_j^2$$

where c_j is the j^{th} column and a_j is an appropriate scaling coefficient corresponding to signal $f(t)$.

Let I be a function measuring the importance of coefficient j on all values of l, and let $J_k(I, C)$ be a function that chooses a subset of $M = 1, ..., m$ of the k largest values of I. Moerchen [18] proves that $J_k(mean(c_j^2), C)$ is optimal in energy preservation.

The MATLAB fast Fourier transform function (fft) was used to find the discrete Fourier transform; the five best features were chosen, based on the energy preservation method.

3.4 Dimensionality Reduction

Even though the success of a classifier is dependent on several variables, which may differ from one classifier to another, all classifiers are dependent on the quality of their input data. To achieve accurate results with the least amount of computational time, it is necessary to ensure that as little noise and redundancy as possible is present in the input. This may involve dimensionality reduction, the process of identifying and filtering out as much irrelevant and redundant information as possible [26].

As mentioned, different classification algorithms will be affected by overparameterisation in different ways. In the k-nearest neighbour classifier, additional features can largely affect the distance between two points. While redundant features (i.e, those that don't change the distance between points) would only influence computational cost, added noise to the system can impact the distance between points, likely in a negative way.

Like k-nearest neighbour, the need for feature reduction in logistic regression has less to do with removing redundancy than with reducing noise and computational cost. Logistic regression accounts for highly correlated features by lowering their weights. Uninformative features, however, would cause weights to be learned that do not improve the performance of the classifier.

Random forests are not as susceptible to the problem of overparameterisation as other methods. When training each tree, since the 'best' features will be branched on towards the top of the tree, pruning could be used to limit the size of each tree (thus avoiding overfitting). An issue would only start to arise when the number of redundant or noisy features is much larger than the number of good features. This is because, when training a tree, a random subset of features is selected when creating a branch. If the number of bad features is much larger than the number of good ones, then the probability of choosing a subset where no good features are present becomes significant.

Dimensionality reduction can usually be characterised as one of two tasks: *feature selection* and *feature transformation*. Feature transformation methods involve performing a transformation of the data (such as a rotation or projection) to create a new set of features (of smaller size) that has more descriptive power than the original set. A commonly used example of this is *principal component analysis* (PCA) which finds a set of orthogonal unit vectors that point in the directions of greatest variance of the data. The features are given by projecting the data onto this basis. While these sorts of methods are popular and do tend to perform well, the resulting features are usually not interpretable [27].

It might be of interest to see which features are most responsible for differences between classes. Therefore, instead of using feature transformation methods, feature selection is used to find a subset of features for which a classifier achieves its best performance. There exist numerous methods for performing feature selection, such as nested subset methods, filters or direct objective optimisation [27], as well as adaptive boosting [28].

We use *sequential floating selection* (SFS) [29] to find the optimal set of features. SFS is a greedy algorithm that works in the following way: Starting with an empty list, sequentially consider each feature selection and assesses its impact on a given evaluation score. Choose the feature that scores best and adding it to the list. Then, again, go through each of the features that have not been added to the list, and assess their impact in combination with the features already added to the list. Find the best one and add it to the list This is repeated until the list is full [30]. A superior method, *sequential forward floating selection*, has been proven to perform better [29], which backtracks after a new feature is included to solve the *nesting* problem, it proved inefficient to implement for the multi class.

3.4.1 Implementation

Since it is not necessarily the case that the best features are the same for each classification proplem, or even for each classification algorithm, the best features are found for each classifier irrespective of the others. The figure of merit for each, which is optimises the classifier is found by using cross-validation and training a classification model model with training data and then evaluating it on a validation set. If at any stage, the feature being considered improves the figure of merit, then the feature will be added to the set of 'kept' features.

Different evaluation scores are used depending on the classifier. In the k-nearest neighbors classifiers, the *mincost* is, which is the predicted label with the smallest expected misclassification cost. The expectation is taken over the posterior probability, and cost as given by the Cost property of the classifier (a

matrix). The loss is then the true misclassification cost averaged over the observations. For the random forest implementation, the cumulative misclassification probability of the entire ensemble is used as the cost to evaluate combination of features. In the case of logistic regression, the deviance of the fit is used. These methods were used for two reasons, firstly because efficient implementations exist with MATLAB's stats toolkit, and they produced the sets of features that performed best on when tested on a validation set.

Chapter 4

Models

4.1 Overview

There are several classification algorithms that can be used to perform supervised learning tasks and vary in their computational complexity, implementation and assumptions that they make about the distributions of the data [5]. Three well known methods are used to classify the data: Logistic regression, random forrest and k-nearest neighbour.

All three methods are examples of discriminative classifiers. The discriminative approach is appealing in that it it directly models $p(y|\mathbf{x})$. Also, density estimation for the class-conditional distributions is a hard problem, particularly when $\mathbf{x}$ is high dimensional, so if we are just interested in classification, then the generative approach may mean that we are trying to solve a harder problem than we need to[31].

4.2 Logistic Regression

For a binary classification problem $y \in \{0, 1\}$, such as discriminating between households with children ($y = 1$) and households without ($y = 0$), the logistic regression model learns a weight vector $\mathbf{w}$ such that given some new household with feature vector $\mathbf{x}$, the posterior probability of that household being in class, $p(y = 1|\mathbf{x}) = g(\mathbf{x}; \mathbf{w})$ where $g(x)$ is the logistic (or sigmoid) function.

$$g(\mathbf{x}; \mathbf{w}) = \sigma(\mathbf{x}; \mathbf{w}) = \frac{1}{1 - e^{-(b+\mathbf{w}\cdot\mathbf{x})}}$$

There are numerous advantages to using logistic regression for the household classification task. Firstly, logistic regression is interpretable. After the model has been trained and the weight vectors established, they can be used to determine how important each feature is to the classifier. Secondly, the confidence of a prediction can be inferred, resulting in interpretable results. There are, however, also drawbacks to logistic regression. Since the maximum likelihood function does not have a closed form solution, an iterative process must be used instead to learn the weights, which is not guaranteed to converge.

4.2.1 Multi-class Logistic Regression

To extend the problem of logistic regression to the multi-class case, often times the the *softmax* is used as a generalisation of the logistic function (σ), the predicted class of an instance is then given by

$$P(y = Y_i|\mathbf{x}) = \frac{\exp^{-(b_i+\mathbf{w}_i\cdot\mathbf{x})}}{\sum_{j=0}^{J}\exp^{-(b_j+\mathbf{w}_j\cdot\mathbf{x})}}$$

Although this is a valid method of classifying the data, it fails to acknowledge the ordinal property of the classes and assumes the data to be nominal. Ideally we would be able to build a model that exploits the fact that some classes are more similar than others. For example, if the true label of a household is B, then we would rather misclassify the instance as A or C1 than as D or E. Luckily, ordinal logistic regression (or ordered ligit) can be used to build a model that incorporated the ordering of the classes.

Using McCullagh's proportional odds model [32], p_i is defined as the proportion of instances that are in class i, then the ordered logit model has the form

$$\begin{aligned}
logit(p_1) &= \log(\frac{p_1}{1-p_1}) = b_1 + \mathbf{w}\cdot\mathbf{x} \\
logit(p_2) &= \log(\frac{p_1+p_2}{1-p_1-p_2}) = b_2 + \mathbf{w}\cdot\mathbf{x} \\
&\vdots \\
logit(p_6) &= \log(\frac{p_1+p_2+p_3+p_4+p_5+p_6}{1-p_1-p_2-p_3-p_4-p_5-p_6}) = b_6 + \mathbf{w}\cdot\mathbf{x}
\end{aligned}$$

The model assumes *proportional odds* which is that each explanatory variable exerts the same effect on each cumulative logit. This translates to the assumption that the weights are the same for each cutoff, but rather the classes have different intercepts b. A pragmatic description of regression models for ordinal data in the context of machine learning is given by Herbrich et. al. [33].

4.2.2 Implementation

To build the binary logistic regression classifier, MATLAB's fitglm tool which fits a generalised linear model to the data. To make the resulting model logistic, a binomial distribution and logit link function were specified. For the case of multi-class classification, two models were created. One McCullagh's proportional odds model (treating the data as ordinal) and one treating the data as nominal. Both methods were implemented in matlab using the mnrfit tool

4.3 Random Forest

Random Forest is a classification method that grows an ensemble of decision trees from a set of training instances and determines the class of a new instance by allowing the trees in the ensemble to vote on the most popular class. For N training sets and M features, each tree is grown by:

- Randomly sample n training instances from the N training with replacement (this will be the training sample to grow the tree).
- At each node, m features are selected at random (where $m < M$). The best of the m features is used to split the node.
- The trees are grown to the largest possible size (no pruning takes place).

A new instance is then classified by running it through each tree, allowing each of the trees to assign the instance a class. The predicted class of the test instance is then taken given by the vote of each tree.

Although (in contrast to building a single decision tree), it is not easy to visualise a random forest, it is still possible to gain an estimate of the variables that are most important for classification and can be used on data sets with a large number of features (see section 3.4). Random forests have been shown to perform particularly well on unseen data compared to other classification methods as they avoid overfitting by only ever looking at a random subset of features and data [34].

4.3.1 Implementation

MATLAB's builtin treeBagger class was used to build the random forest. Because bootstrap aggregation is used to randomly sample the training data, the out-of-bag estimates were used to optimise the model's parameters, instead of using cross-validation. The parameters to optimise are the number of trees and size of features m to consider for splitting each node.

4.4 K-Nearest Neighbour

K-nearest neighbor is a fundamental method for classification as it is intuitive and requires little *a priori* knowledge about the data. It is a non-parametric model that classifies an unlabeled input by finding the k-nearest training points in feature space, using the classes of the nearest points to predict the class of the unlabeled point [35].

4.4.1 Implementation

MATLAB's fitknn tool was used to build a nearest neighbour classification model and the optimum parameters were found using 5-fold cross-validation. The parameters to find were the distance measure, search method and k (the number of neighbours).

Chapter 5

Results

This section discusses the quantitative evaluation methods used to determine the potential for each of the classifiers to reveal household characteristics and then analyses the results from training and running each classifier.

5.1 Evaluation Methods

For each classifier, a *confusion matrix* (CM) is produced using the MATLAB tool `confusionmat`, which, for a K class classification problem, returns a $K \times K$ matrix where each element (i, j) contains the number of times an instance of class i has been classified as j. The diagonal elements elements of CM contain the number of instances of households that have been classified correctly for each class. [36]

The accuracy of a classifier is defined as the sum of the diagonal elements of CM, divided by the total number of samples,S.

$$ACC = \frac{\sum_{i=1}^{K} CM_{i,i}}{S}$$

This is compared to the accuracy of performing a random guess (RG), which assigns a household to one of the K classes at random.

$$ACC_{RG} = \frac{1}{K}$$

To account for the imbalances in classes, we also calculate the most probable class (MPC) which uses knowledge of the prior probability of each class in the training data to find a baseline by assigning all samples to the most probable class.

$$ACC_{MPC} = \frac{argmax(S^K)}{S}$$

where S^K is the number of samples from the test data that are in class K.

For socio-economic classification problem, the ordinal structure of the classes should also be taken into account i.e it is worse for our classifier to predict a household of social grade B as D, then it is to predict it as C1 or A. Therefore, the *accuracy within* n[37].

Particularly for unbalanced classes, reporting the accuracy alone is not satisfactory in determining the quality of a classifier. The obvious and well known example being; constructing a classification problem where 99% of instances are in class A and only 1% in class B. A classifier that simply predicts all new data as class A would be correct 99% of the time, but would still not be a good classifier.

A widely applied method for evaluating a classifier is to compute the *true positive rate* (TPR) and *true negative rate*(TNR). The TPR gives the proportion of positives that are correctly identified as being positive, while the TNR gives the porportion of negatives that are correctly identified as negative.

$$TPR = \frac{TP}{TP + FN} \qquad TNR = \frac{TN}{TN + FP}$$

From these statistics, it is common to plot an ROC curve, which is a plot of the TPR against the *false positive rate*(FPR), which is defined as 1-TNR. The evaluation criterion (the area under the ROC curve) is preferred over the accuracy, particularly when considering unbalanced classes as the impact of skewness can be analysed [38]. To create the ROC curve, a value is found for each classifier which acts as the threshold above which an instance is classified as positive. Typically for logistic regression, this is the probability of an instance being assigned to class 1.

This is not as straight forward for random forests and knn as they are not probabilistic classifiers. Probabilities can, however be generated from the classifier results. For random forest the decision boundary may be the ratio of number of trees that vote in favor of assigning an unseen instance to class 1 and the total number of trees. In knn it is the number of nearest neighbors that are of class 1 divided by the total number of nearest neighbors.

In computing the ROC curve to evaluate the binary classification task of discriminating between households with and without children is straight forward, it is straightforward te determine which class is 'positive' and which is 'negative' (has children is positive). However for multi-class classification it is unclear what is 'positive' and what is 'negative'. When evaluating their socio-economic classifier, Beckel et. al. group nearby groups together and then use a one-versus-all approach[13, 5]. A similar method is used, analogous to the *accuracy within n* method described above, where classes within n are considered positive and all else are negative.

5.2 Classifiers

Bibliography

[1] Office for National Statiscics. *Full Report: Household Energy Spending in the UK, 2002-2012*. 2014.

[2] Stop Smart Meters! (UK). Stop smart meters! (uk), 2015.

[3] Elias Leake Quinn. Privacy and the new energy infrastructure. *SSRN Journal*, 2014.

[4] Mikhail A. Lisovich, Deirdre K. Mulligan, and Stephen B. Wicker. Inferring personal information from demand-response systems. *IEEE Security and Privacy Magazine*, 8(1):11–20, 2010.

[5] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, 2014.

[6] Gianfranco Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68–80, 2012.

[7] Hong-An Cao. *Are domestic load profiles stable over time? An attempt to identify target households for demand side management campaigns*, pages 4733 – 4738. IEEE, 2013.

[8] J. Z. Kolter and Tommi Jaakkola. Approximate inference in additive factorial hmms with application to energy disaggregation. In Neil D. Lawrence and Mark A. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, volume 22, pages 1472–1482, 2012.

[9] Intertek. *Household Electricity Survey A study of domestic electrical product usage*. 2012.

[10] Jason Palmer, Nicola Terry, and Tom Kane. *Early Findings: Demand side management*. 2013.

[11] Household electricity survey: Cleaning the data.

[12] Christian Beckel, Leyna Sadamori, and Silvia Santini. Automatic socio-economic classification of households using electricity consumption data. *Proceedings of the fourth international conference on Future energy systems*, pages 75–86, 2013.

[13] Christian Beckel, Leyna Sadamori, and Silvia Santini. *Towards automatic classifi-cation of private households using electricity consumption data*, pages 75–86. ACM, 2013.

[14] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. Evaluation of time series techniques to characterise domestic electricity demand. *Energy*, 50:120–130, 2013.

[15]

[16]

[17] Department of Energy and Climate Change. *Domestic energy use study: to understand why comparable households use different amounts of energy.* 2012.

[18] Fabian Moerchen. *Time series feature extraction for data mining using DWT and DFT.* 2003.

[19] Jason Osborne. Notes on the use of data transformations. *Practical Assessment, Research and Evaluation*, 2002.

[20] Morgan C. Wang and Brad J. Bushman. Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*, 3(1):46–54, 1998.

[21] Leticia M. Blazquez Gomez, Massimo Filippini, and Fabian Heimsch. Regional impact of changes in disposable income on spanish electricity demand: A spatial econometric analysis. *Energy Economics*, 40:S58–S66, 2013.

[22] M. Bartley and C. Owen. Relation between socioeconomic status, employment, and health during economic change, 1973-93. *BMJ*, 313(7055):445–449, 1996.

[23] Teachingintheuk.com. Teaching jobs — supply teaching jobs - teaching personnel, 2015.

[24] Mark S Nixon and Alberto S Aguado. *Feature extraction and image processing for computer vision.* Academic Press, 2012.

[25] Amara Graps. An instroduction to wavlents. *IEEE Computational Science and Engineering*, 2(2):50–61, 1995.

[26] Mark A. Hall and Lloyd A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper, 1999.

[27] Andre Elisseeff Isabelle Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 2003.

[28] Ruihu Wang. Adaboost for feature selection, classification and its relation with svm, a review. *Physics Procedia*, 25:800–807, 2012.

[29] Somol P., P. Pundil, J. Novicova, and P Pacli'k. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20(11-13):1157–1163, 1999.

[30] Juha Reunanen. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 2003.

[31] Carl Edward Rasmussen and Christopher K. I Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[32] Peter McCullagh. Regressuib models for ordinal data. *Journal of the Royal Statistical Society*, 1980.

[33] Klaus Ob ermayer Ralf Herbrich, Thore Graep el. Regression mo dels for ordinal data: A machine learning approach. Technical report, Technical University of Berlin, 1999.

[34] leo Breiman. Random forests. *Machine learning*, 2001.

[35] Leif E. Peterson. K-nearest neighbor, 2009.

[36] JERZY STEFANOWSKI. Data mining - evaluation of classifiers. Poznan University of Technology.

[37] Lisa Gaudette and Nathalie Japkowicz. title = Evaluation Methods for Ordinal Classification,. In *Advances in Artificial Intelligence*.

[38] Willem Waegeman, Bernard De Baets, and Luc Boullart. Roc analysis in ordinal regression learning. *Pattern Recognition Letters*, 29(1):1–9, 2008.