Machine Learning with Domestic Energy Use Data

Sam Stern (s1134468)

March 27, 2015

Abstract

Contents

Chapter 1

Introduction

1.1 Introduction

Amidst international pressure on countries to reduce their carbon footprints [?] and the British public's becoming increasingly frustrated by rising energy bills with little to no explanation as to the reasons behind the increases [?], 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 [?]. 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 [?] [?]. 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 [1]. Chicco, for example, gives an overview of the clustering techniques used to establish suitable client groups for analysing electricity load pattern data [**?**]. Cao et.al also grouped consumers using electricity load profiles, however focusing on finding households with the same peak usage [**?**].

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 [**?**]. 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) [**?**].

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 [1]. 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 came from The Household Electricity Survey (HES), a UK-government sponsored study of residential energy usage jointly commissioned by the Department for Environmental, Food and Rural Affairs (Defra) and the Department of Energy and Climate Change (DECC). Britain's most detailed look at energy consumption in the home environment to date [A], HES tracked the electrical power demands and sources of energy utilization in a variety of owner-occuppied homes in England over the period May 2010 to July 2011 [?]. The study sought to identify, catalogue and monitor the range, quantity and energy demands of appliances found in 'typical' British homes, with the underlying aim being to better understand households' frequency and patterns of electricity usage, and to collect any 'user habit' and/or other socio-economic data that might emerge. [?]. This information would be plied in a variety of ways and purposes, not the least of which would be as an aid in developing energy policy (both at the consumer and energy provider levels), and to help justify the £12bn smart meters roll out.

The HES study monitored 250 households, of which 26 were observed for one year, and the remaining 224 for roughly one month. Although all the participating households were located in England, they did not share the same demographic or geographic profiles. This was reflected in the wide spectrum of number and ages of appliances. Whereas one household, for example, registered just 13 appliances, another had 85. There was a forty-one year old freezer, several brand new televisions and a broad assortment of ages and types of devices in between. [B] When aggregated, (as outlined in Section ??), the result could be considered an estimate of an average mains reading.

Smart meters record the total energy being used in a given interval, whereas when discussing individual appliances, it is common to talk about the energy used per unit time (i.e power). For example, an average kettle might use 3kW of power. If the kettle is running for 2 minutes then the energy used would be 6kWm (kilo Watt minutes). A potential issue arises when we consider another appliance that uses less power but for a longer period of time. A hairdryer, for instance, uses 1kW of power. If a hairdryer was being used for 6 minutes, then the total energy used also be 6kWm. Although the hairdryer uses less power than the kettle, the smart meter reading would record the same number (6kWm). The

Smart meters used in the HES study took readings either every 2 minutes or every 10 minutes in units of deciwatt hours (dWh or 0.1Wh). This is a measure of the total energy that the home consumed since the last reading. As it is conventional to describe the energy consumed in terms of kilowatt hours (kWh), the readings are each multiplied by 10,000.

In addition to the data collected on appliance types and the meter readings, participating households also kept diaries of how they used their mains appliances, and provided supplemental information about the household and its constellation, such as: the number of occupants, employment status, Ipsos MORI social-grade and whether there were children present.

2.2 Extracting the Data and Pre-Processing

As explained in Section 2.1, electricity readings from individual appliances and sockets were taken for each household. This was in contrast to the total-energy-consumed figures needed for this project. In organizing its data, the HES study assigned values to the 250 possible appliances that the designers of the study expected a household to have. Appliances that were not present in a household were designated a 0. The resulting raw data was held in large csv files. Since no household had all 250 potential appliances, there were a significant number of redundant entries. In order to use the data to perform data mining, therefore, numerous pre-processing steps needed to be performed. This was accomplished by writing and implementing python scripts with embedded SQL.

Specifically, 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 (CAR) [**?**] , an architectural consultancy, provided additional files that mapped which appliances needed to be aggregated for each household to arrive at an estimated mains reading, as 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.

Another consideration was the number of participating households. 250 is a relatively small number for a machine learning task, since they would probably not produce enough data to build models that would 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. This generated an additional 281 household instances. While it did not create a more diverse group, it did add more instances to train and validate, as well as with which to test a classifier. To avoid overfitting the classification models to the data, all instances from the 26 (split) households were either in the training or test set, but never in both.

The inconsistency in measurement intervals alluded to at the beginning of this section also had to rationalised. While some households reported how much energy they used every 10 minutes, 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 the households had 10 minute granularity. This step was important because some of the consumption features would have been affected by differences in measurement intervals. Once all the households were represented in terms of 10-minute intervals and in units of 0.1Wh (deci Watt hours), each reading could be multiplied by 1000 to convert the data to kWh.

The last stage in pre-processing was to ensure that each instance was of the same duration, with each day of the week occurring the same number of times across households while still maintaining being ordered by their point in time. The justification for this is as follows: When visualising the data, temporal structure was observed both intraday and intraweek. It was noted that in addition to an obvious daily pattern (more energy being used during the day than at night), there was also a repetition over weekly cycles, where it was possible to distinguish some days of the week from others as a pattern emerged every 7 days. Ensuring each household had an integer number of weeks would mean that no single day of the week would effect the total energy used by the household. For example, it emerged that more energy is used on Sundays than other days of the week. If one household had three occurrences of a Sunday while being monitored while another had four occurrences of a Sunday, then when computing features such as the average daily energy would be different for the two households simply due to the extra Sunday. Because more households were recorded for roughly one month, it was decided to ensure that the data for each household was 28 days long (4 weeks). Furthermore, it was important to ensure that the days of the week remained in order (i.e that a Monday was followed by a Tuesday, which is followed by Wednesday etc.). This is because some features, such as computing the Fourier transform, expect the data to be in sequential order. Finally, the data was made to start on the same day of the week for each household as this made it far more convenient to extract specific parts of the data.

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 (this 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 day's worth of readings.

Figure **??** gives a visual example of data that has been 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 than the required number of days, days are either reused or, if possible, taken from the days that have been chopped from the top[1].

Figure 2.1

[1] It was noted that this method does create a bias in the features. For example, when computing the average energy used on a Monday, if a household only has three unique instances of a Monday and one instance has been reused, then this will affect the feature.

2.3 Household Classes

Each household that participated in the HES study completed a survey with questions about the building they occupied (such as the year the house was built), the household itself (such as the number of occupants) as well as their 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 **??** and **??** show the sample sizes for each class of the two classification problems being considered 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	% Population
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 **??** and **??** 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 ??, 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 ?? and ?? 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.

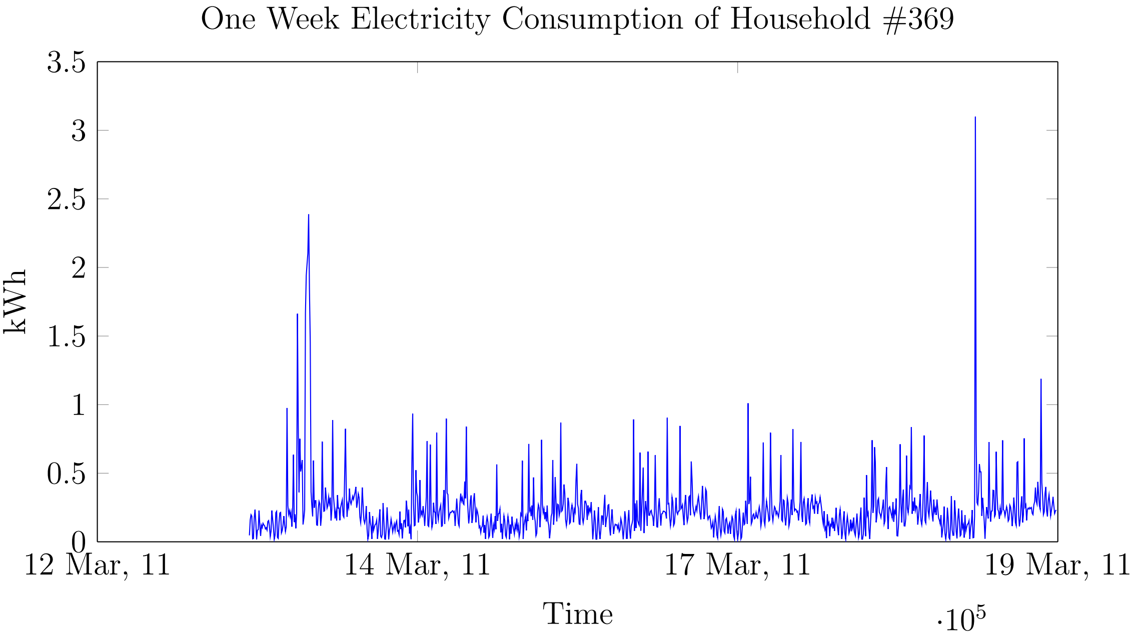


Figure 2.2: The electricity use of household No.369 over a week after being preprocessed

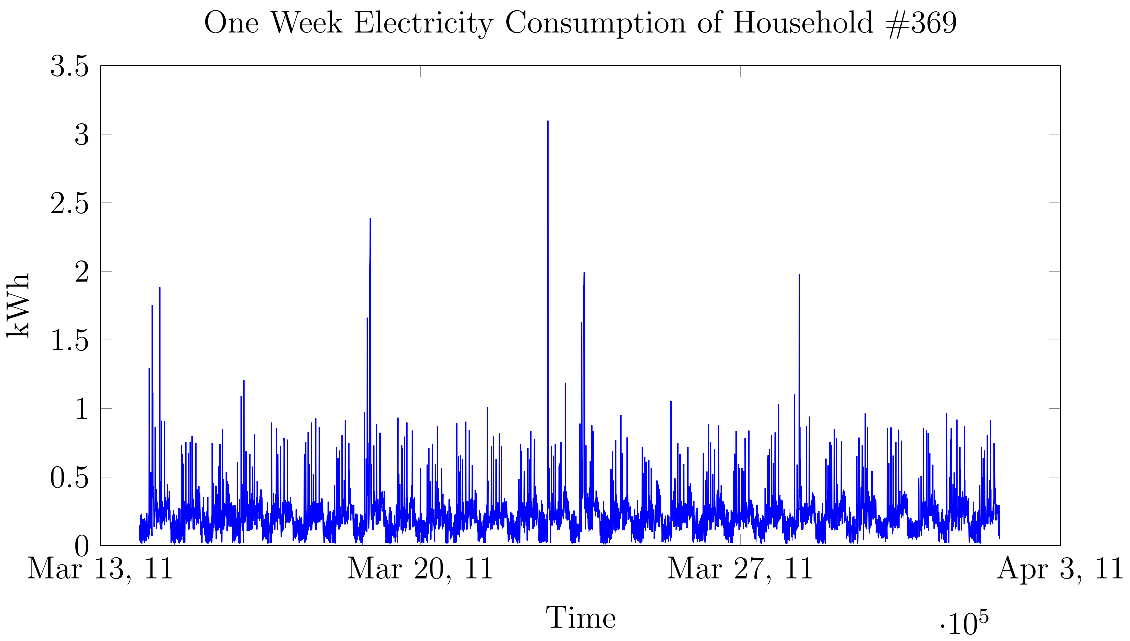


Figure 2.3: The electricity use of household No.369 over a four week period after being pre-processed

2.5 Issues

As with any study, several issues arose which needed to be recognised and dealt with. These were invariably related to the data. Some were simply the result of environmental influences, while others reflected the methodolgy used in conducting the study. Those most relevant to this project are outlined below; more can be found in the CAR report [?].

The first problem with the data concerned the number of households that participated in the HES study. Comparing the UK's 250 households to, for example, the 4,232 that took part in Ireland's CER (Commission for Energy Regulation) study of household electricity consumption [2] (used by Beckel et. al. and McLoughlin [?, 10, 1, ?]), it is less likely that the UK results generalised as well as the Irish ones, particularly for the multi-class classification problem (where there are as few as 32 households per class).

Moreover, only English homes were included in UK study (i.e, Scotland, Wales and Northern Ireland were not represented), and all of the houses were owner-occupied. While 84% of the British population does live in England, only 64% of homes in England are owner-occupied [?]. As such, the subset of participants considered in the HES study was not fully representative of the UK as a whole. It is important to remember, however, that the aim of this project was to determine *whether it was possible* to infer a household's properties from its electrical power consumption, not to build a classifier that could be used to infer British household properties from smart meter data. The distinction being that this is a proof-of-concept project that looks at whether information about a household is contained in the energy use patterns, rather than an attempt to build a commercial product. Therefore, the quality of the sample population households (or lack thereof) is

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

not detrimental to the aim of this project.

A more bothersome issue involved the quality of the data that was gathered during the HES study. Looking at the household in Figure **??** it credibly shows the characteristics of a typical home's consumption; it is one of the 'better' households in the sample. There are many others, however, that do not follow the same kind of trend, such as the two presented in Figure **??**. Either they do not have the same well-defined periodicity, or they may use significantly more (or less) energy than the average household. In these cases, the task then became to find a means of determining whether these discrepancies were reasonable differences that could be attributed to variations between households, or whether they were 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 could not be given with certainty (as it was not definitively known whether all appliances and sockets in a household were recorded. This complication is made evident in Figure **??** where Household #75 is always using at least some energy while Household #121 sees its consumption drop to 0. Household #75 serves as a better estimate as it is reasonable to assume that there will always be a small amount of electricity used by a home since appliances are not 100% efficient and usually leak electricity.

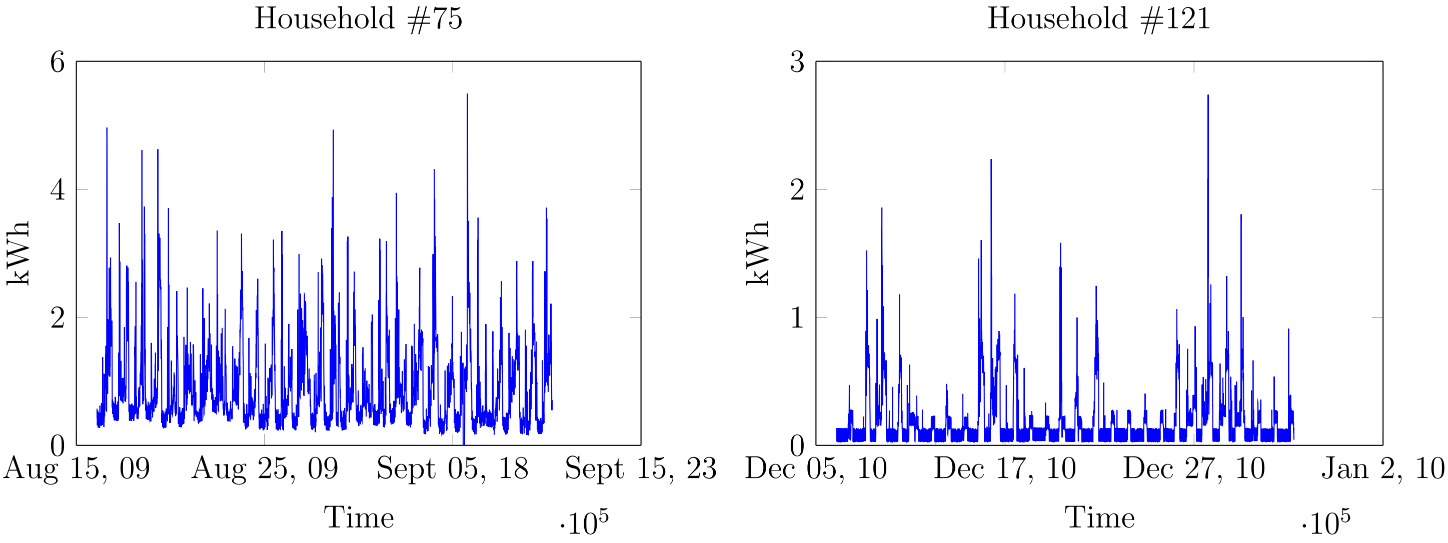


Figure 2.4: The electricity consumption of two households that do not show the same pattern of consumption as other households

Table **??**, is taken from a Center for Sustainable Energy [**?**] report and shows how much power various appliances use. Comparing these values to the data from the meter readings, it is reasonable to see a household use anywhere from 50W to upwards of 15,000W. However, the chart also indicates that some of the most expensive appliances (which have large effects on a households' consumption) are not always present in all households. These include electric cookers, electric showers, electric heaters and tumble driers. In he HES study, only 38 of the 250 households used electric water heating. While it could be expected that these factors would impact a household's consumption, and knowing this would aid in classifying the households, it was left to be work out independently, as the so called *disagregation problem* (see section **??**) is a popular topic of research in and of itself.

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

In examining the data closely, a small number of instances could be noted where the electricity usage for selected households showed consumption levels either flat-lining or completely disappearing for one or more days. To compensate for this anomoly in the data, the usage patterns for each affected household before and after the aberration were analysed and compared to the event. In most cases, it was possible to visually determine whether the incongruity was most likely related to a residence being unoccupied (i.e., the family were on holiday), or if it was due to some technical issue(s) linked to the energy readings themselves (i.e., the meters presenting erroneous data). A decision was taken to discard households with a consumption reading of 0kWh when that reading represented a statistically significant proportion of the total time for which the household was being observed. If a reading for a single day appeared completely out of keeping with all the other readings for that household, then that day's data was discarded and replaced by an equivalent day from another trial week.

The next factor that needed to be considered was the effects of weather, the time of year in particular. Colder temperatures and shorter periods of sunlight during colder months have been shown to precipitate higher electricity usage. [?]. Although CAR was able to provide a document outlining which appliances needed 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 in the study. 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, seasonal 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 [?]. 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[10], 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 [?] [?]. Figure **??** 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.

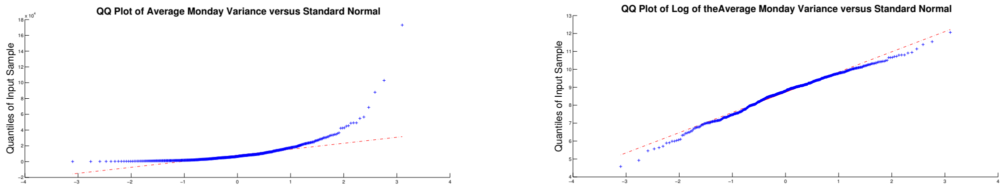


Figure 3.1

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.

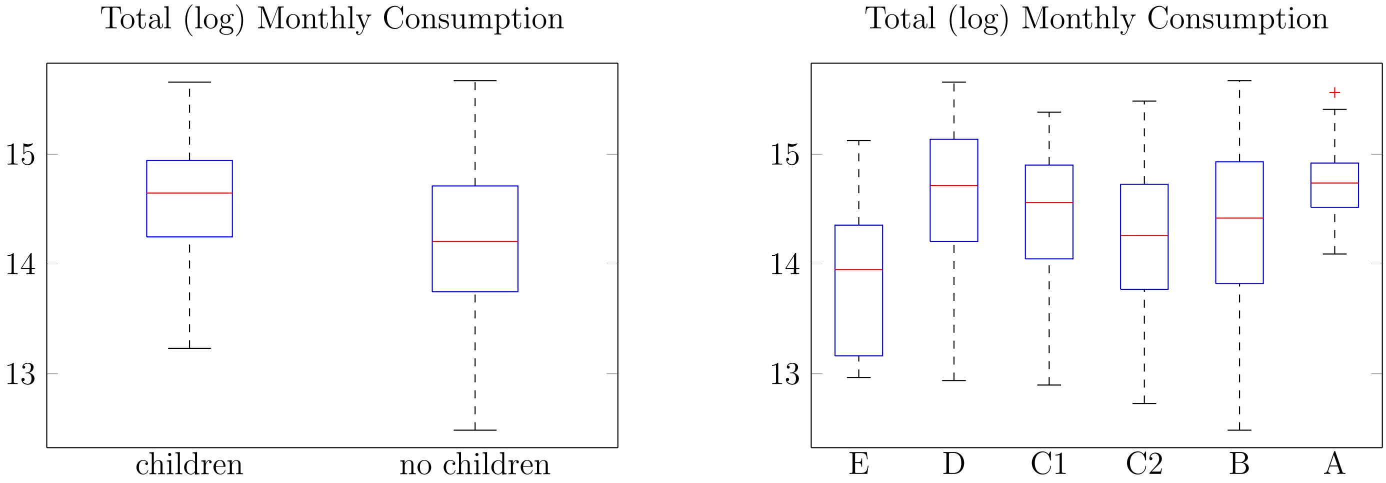
Instead of creating features in an ad hoc manner, a more cost efficient approach was taken. Feature selection was done in the following way: 1) Assumptions were made regarding the distinction between classes (e.g., households with children use more energy overall). 2) Features were created to capture 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 9500 Watts in the space of 10 minutes, others averaged as little as 390 Watts for 10 minutes; while one household was recorded to consumed up to 19500 Wh in 10 minutes, another never used more than 1900 Wh over 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 [?].



(a) Total electricity used by households in a 28 day period, 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 ??, 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.

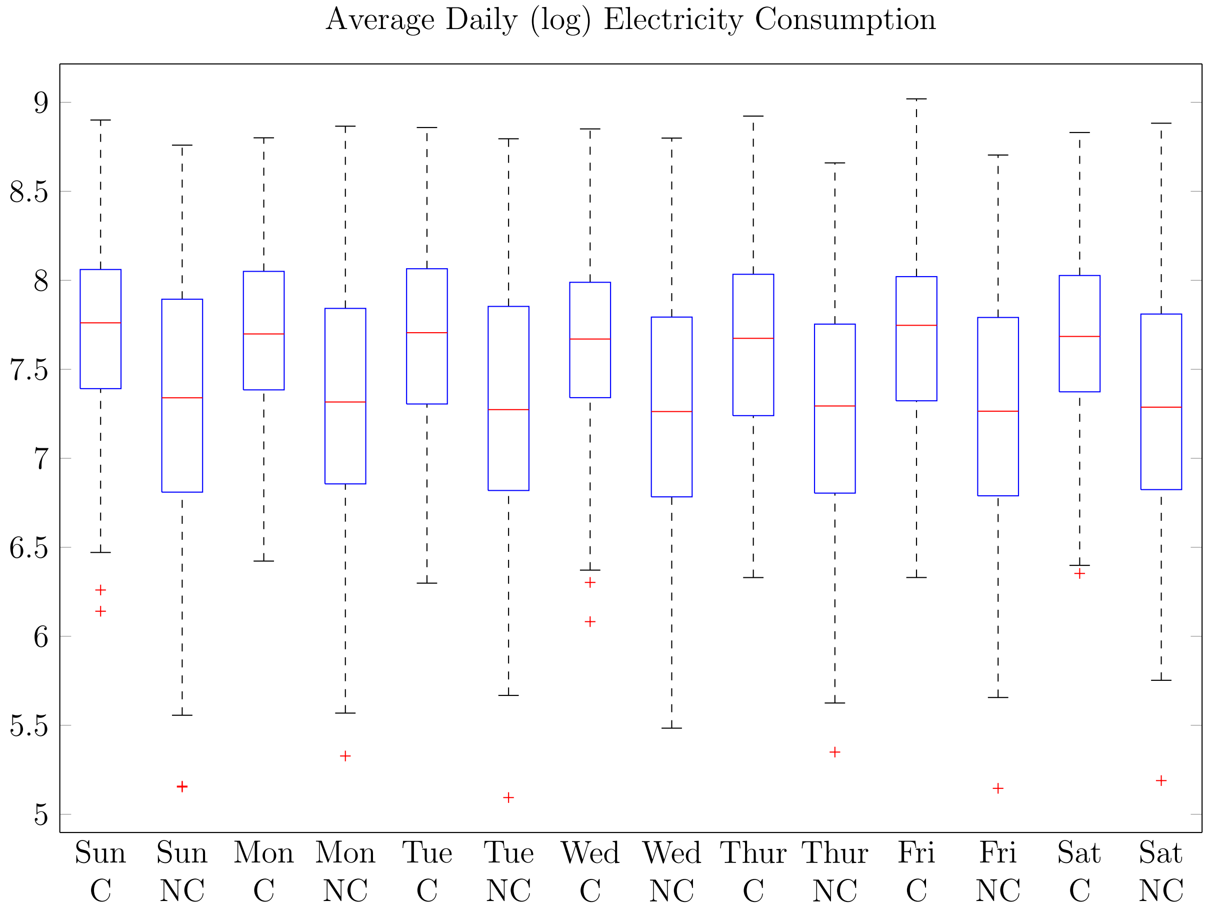


Figure 3.3: The average total energy used on each day of the week. Households are grouped by whether or not there are children present

Figure 3.4

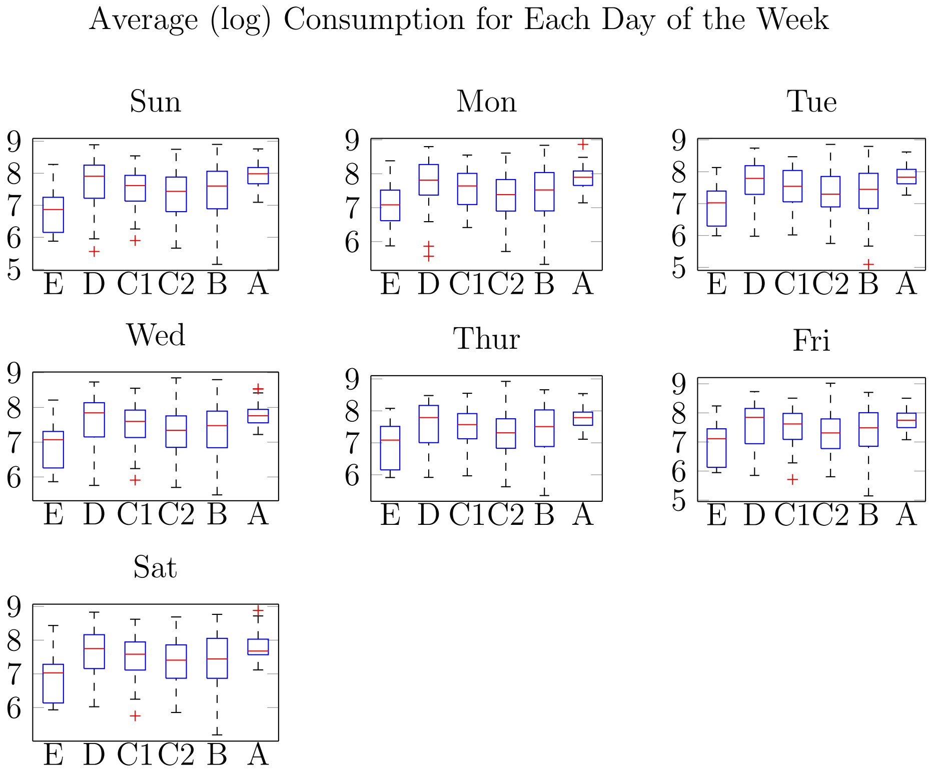


Figure 3.5: The average total energy used on each day of the week. Households are grouped by their IPSOS social grade

While Figure ?? 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 ??, 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 [?]. 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 [?]. 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.

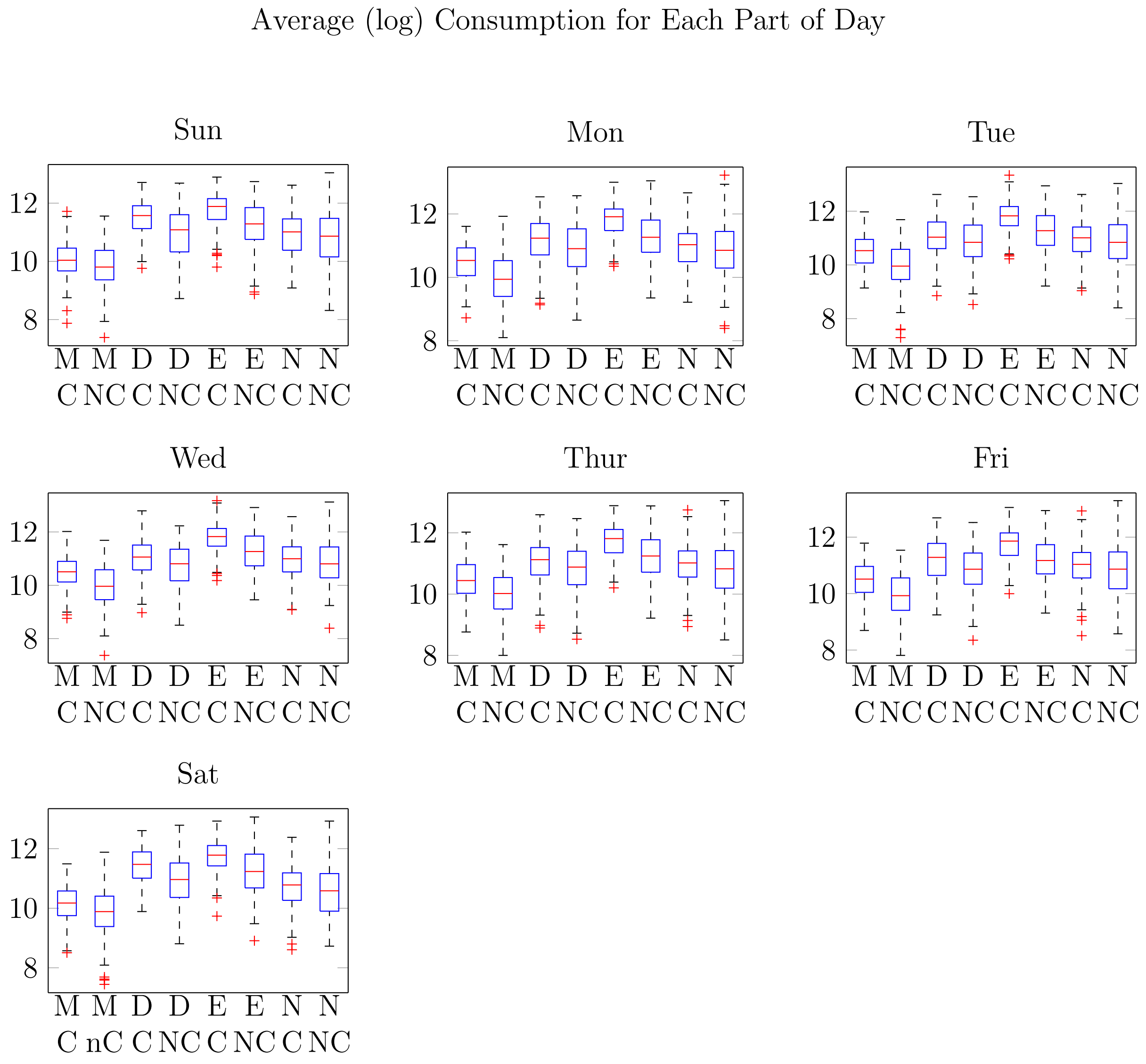


Figure 3.6

The data portrayed in Figure ?? 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.

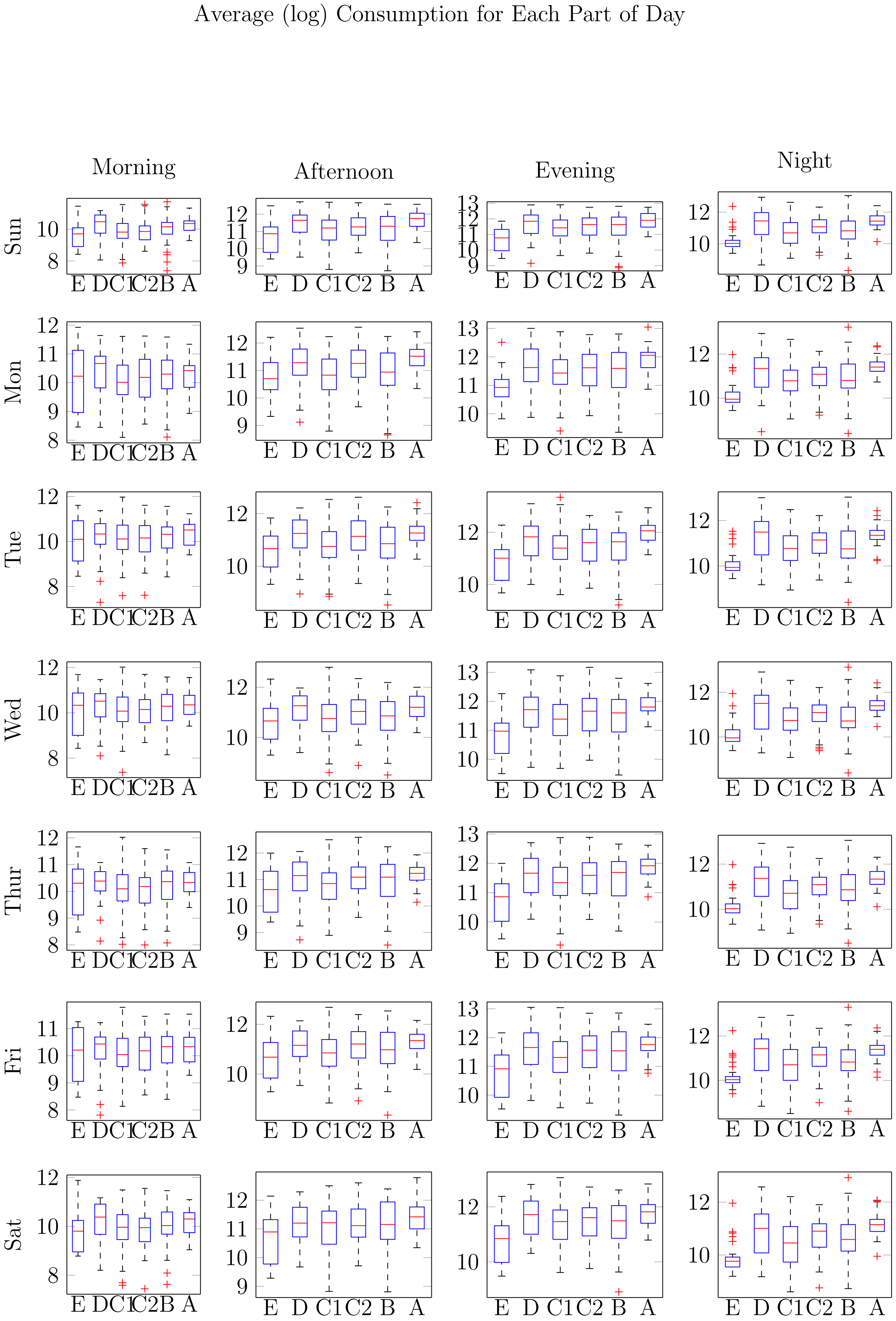
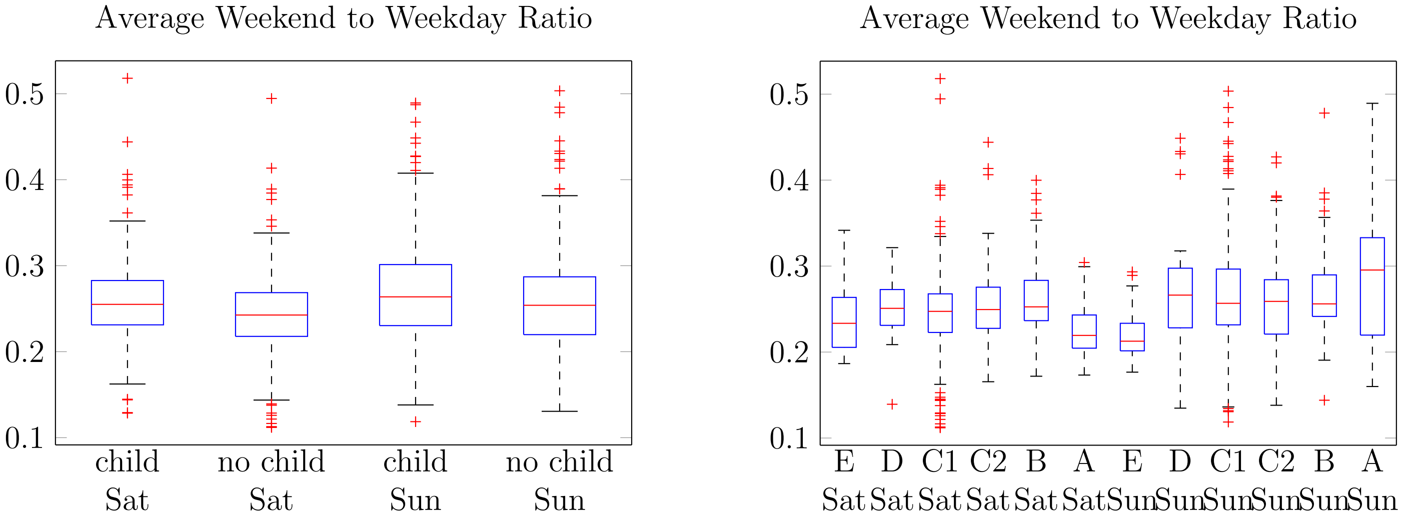

Average (log) Consumption for Each Part of Day
Morning
Afternoon
Evening
Night
Sun
Mon
Tue
Wed
Thur
Fri
Sat
E D C1 C2 B A
8
9
10
11
12
13

Figure 3.7

Figure ?? 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. Households are grouped on whether or not there are children present

(b) The ratio between how much energy is used on the weekends and how much is used on weekdays. Households are grouped on their IPSOS social grade

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 ?? 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.

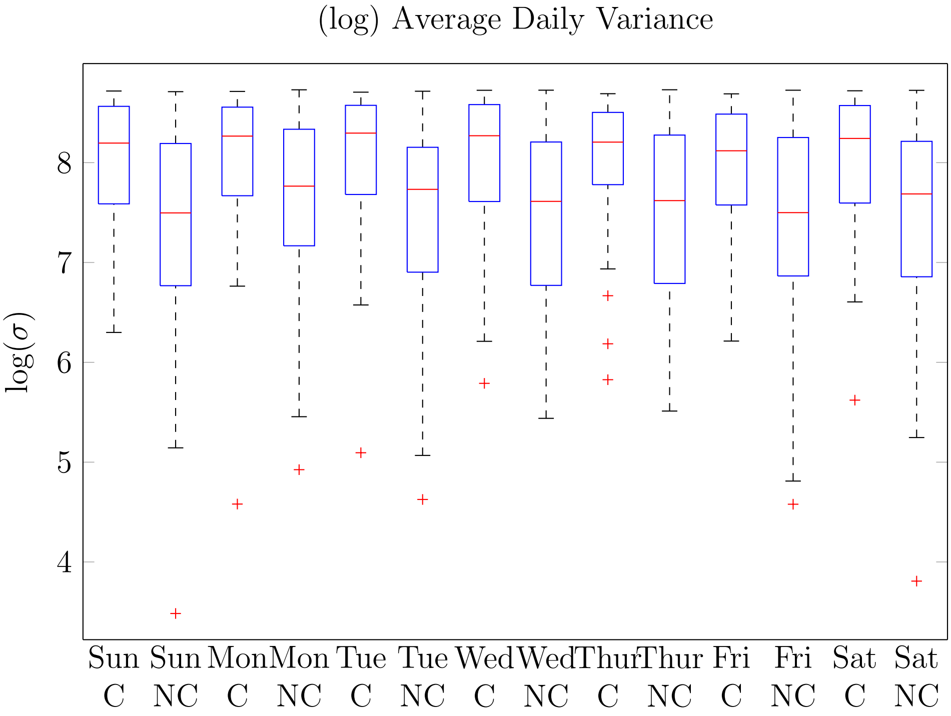


Figure 3.9: The variance of a household's daily consumption grouped by the day of the week and whether the household has children (C) or not (NC)

Although the average daily variance of households is volatile in and of itself, the results shown in Figure **??** 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.

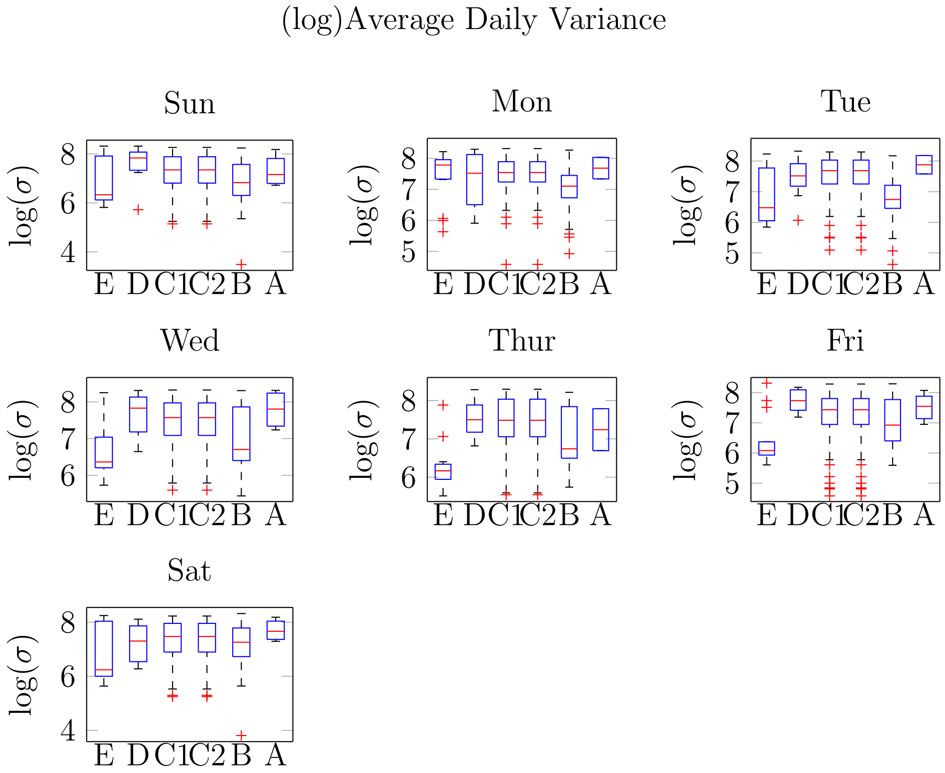


Figure 3.10: The variance of a household's daily consumption grouped by the day of the week and the Ipsos social grade

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.

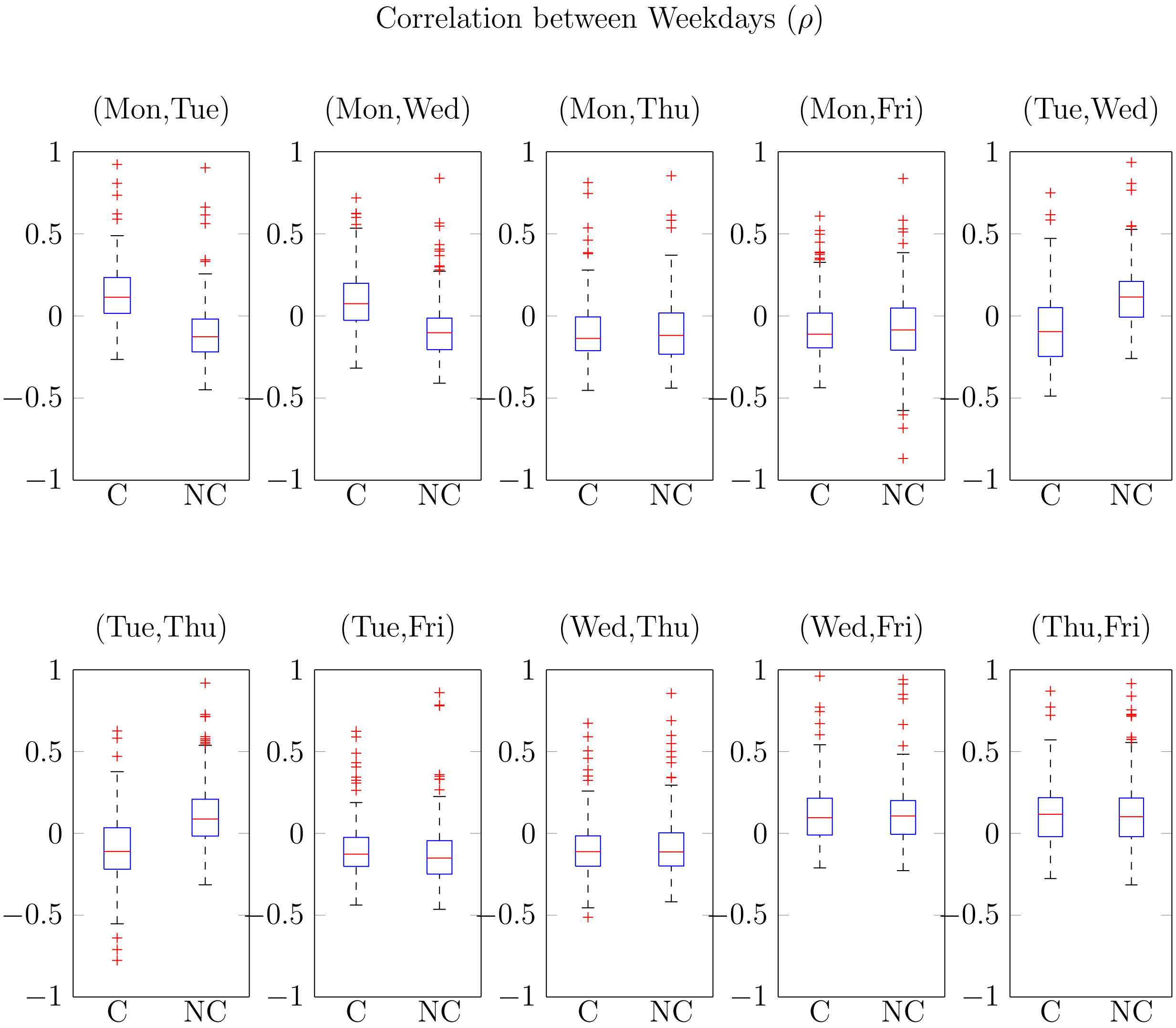


Figure 3.11: Average correlation coefficient between weekdays grouped by whether a household has children (C) or not (NC)

Looking at Figure **??**, 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 **??**, the correlation between days does not result in features that separate classes.

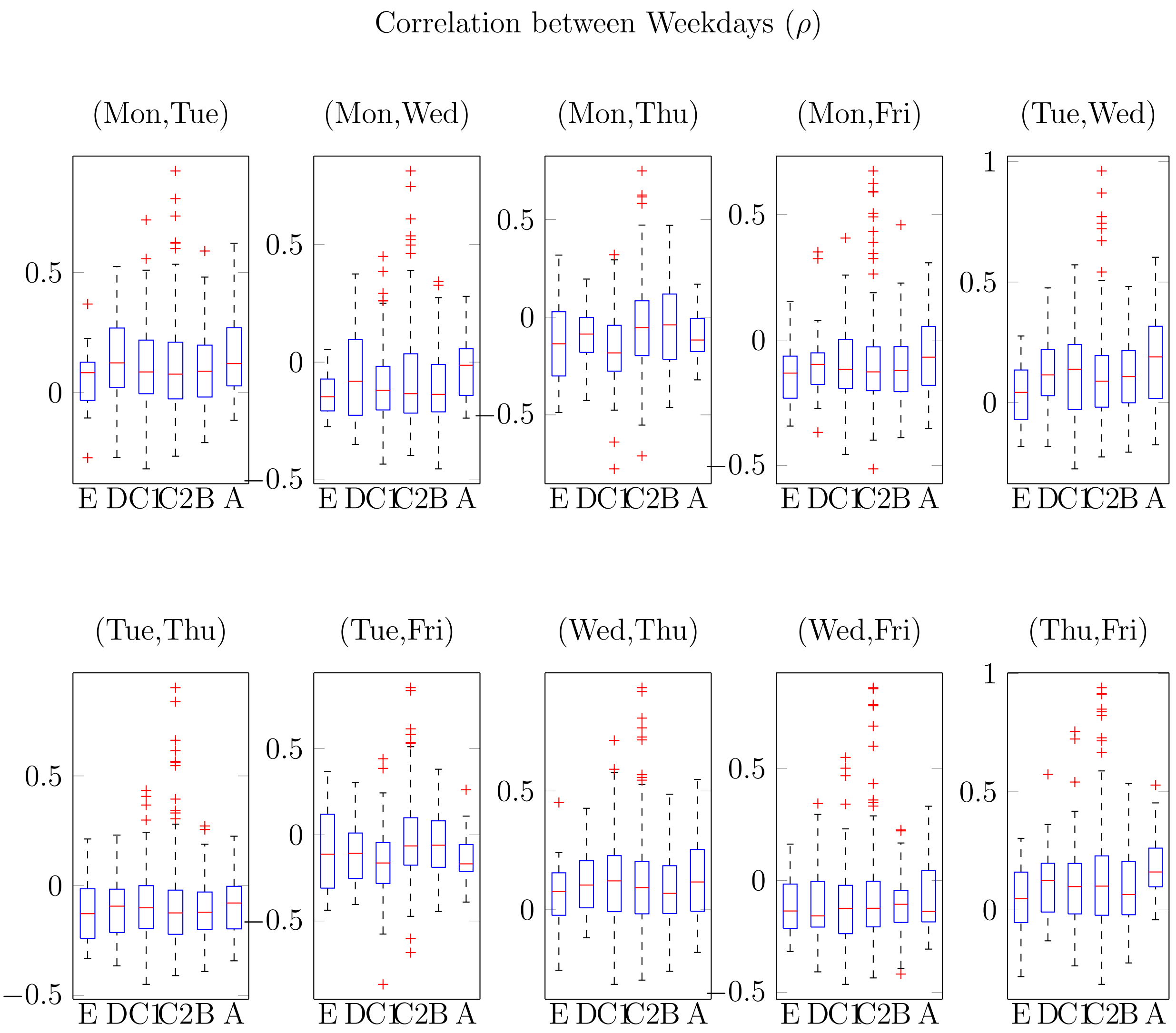


Figure 3.12: Average correlation coefficient between weekdays grouped by Ipsos social grade

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 [?] 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. [?] 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 [?]. This is done because the time-frequency transformation used (the discrete Fourier transform method) has difficulty characterising small intervals of large electricity demand [?].

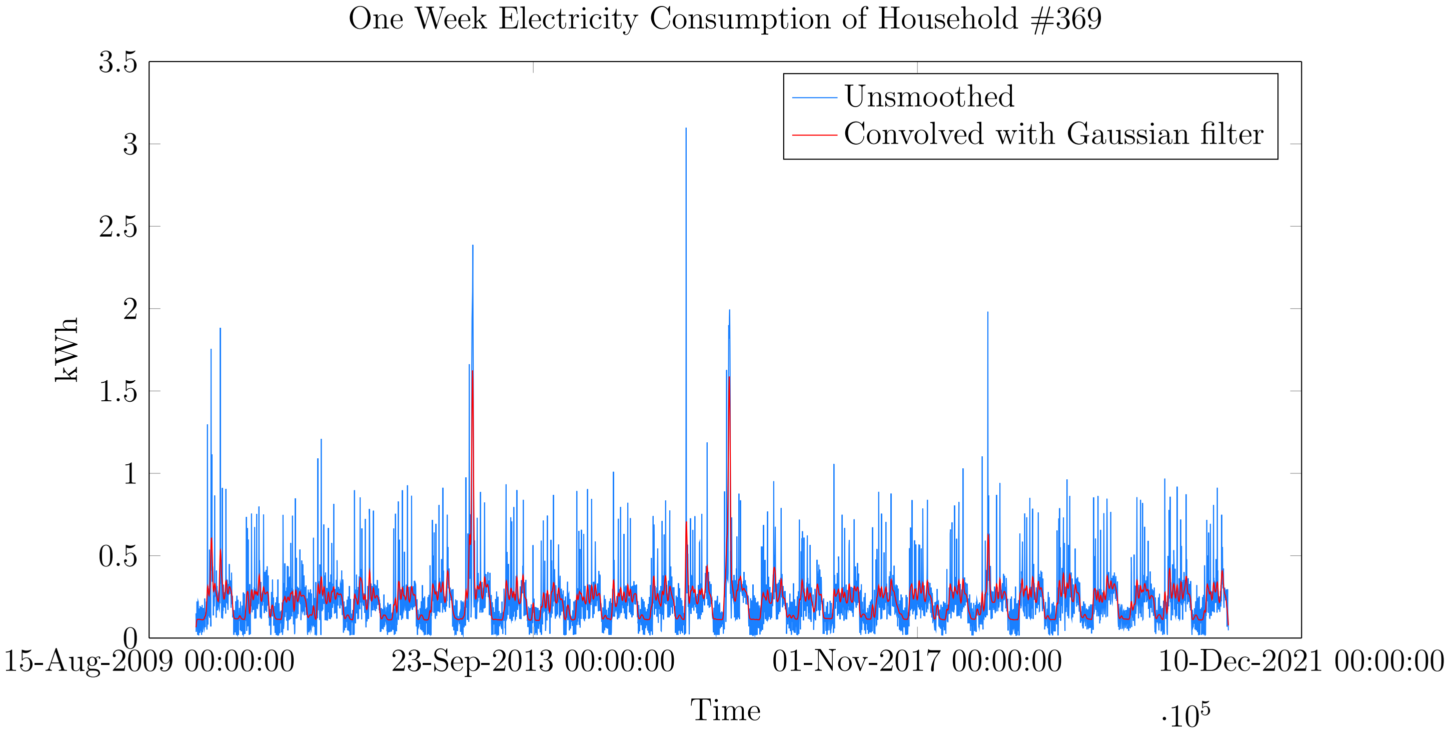


Figure 3.13: The electricity use of household No.369 shows that households may have both a daily and weekly pattern. The clusters of peaks represent individual days while the regions without peaks are the indicative of night time. Additionaly, the large spikes are observed roughly every seven days, on 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 [?].

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 [?].

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 [?], 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 [?] 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 [?].

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 [**?**].

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 [**?**], as well as adaptive boosting [**?**].

We use *sequential floating selection* (SFS) [**?**] 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 [**?**]. A superior method, *sequential forward floating selection*, has been proven to perform better [**?**], 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 [1]. 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[2].

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 missclassify 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.

McCullagh's proportional odds model [3] is a variation of a generalised linear model (glm) where the dependant variable is thought of as being continuous, but is recorded ordinally. It can be thought of as asking asking a linear model to tell you what range a dependent variable is in (as opposed to an exact value). The regression model is then given by:

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

This model relies on the *proportional odds assumption*, which is that the **w**s are independent of the classes (hence they have no subscripts). This translates to the assumption that the weights are the same for each cutoff, but rather the classes have different intercepts b, in contrast to multinomial logistic regression (where the dependent variables are assumed to be nominal) which learns a new set of weight parameters for each cutoff point. A further description of regression models for ordinal data in the context of machine learning is given by Herbrich et. al. [4].

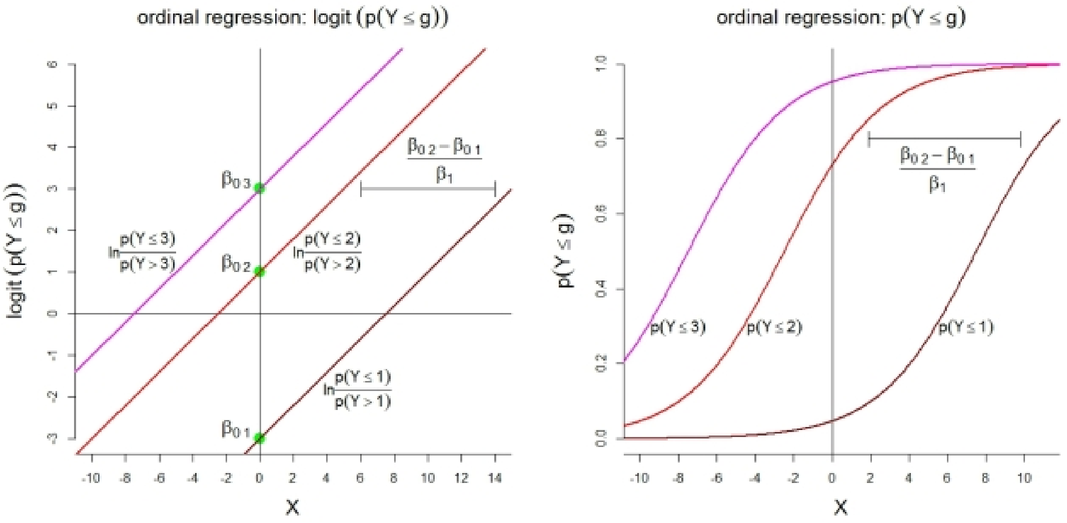


Figure 4.1: The proportional odds model assumes that the regression function for different response categories are parallel on the logit scale. Figure obtained from: http://www.datavis.ca/courses/grcat

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 **??**). 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 [5].

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 [6].

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. [7]

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[8].

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.

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

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 [9]. 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'. 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[10, 1]. 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.

The final metric that is presented is the Matthews correlation coefficient (MCC) which is a value between -1 and +1 representing the correlation between the predicted and true values of a binary classifier. A MCC of -1 indicates that that there is no correlation between the predicted a true class while a value of +1 would indicate a perfect classifier while a value of 0 means it is no better or worse than a random guess. MCC is a worthwhile metric to report as it gives a value to the performance without inflating the imbalances in class sizes [11].

Let X, Y each be an $S \times N$ matrix where S is the number households and N is the number classes. $X_s, n = 1$ if a sample s is predicted to be the n^{th} class and 0 otherwise, $Y_s, n = 1$ if a sample s belongs to the n^{th} class and 0 otherwise. The covariance of X and Y can then be written as

$$cov(X, Y) = \frac{1}{N} \sum_{s=1}^{S} \sum_{n=1}^{N} (X_{s,n} - \bar{X_n})((Y_{s,n} - \bar{Y_n})$$

where $\bar{X_n}$ and $\bar{Y_n}$ are the means of the n^th columns of X and Y respectively. The MCC is then defined as,

$$MCC = \frac{cov(X,Y)}{\sqrt{cov(X,X) \cdot cov(Y,Y)}}$$

For binary classification this can be interpreted as,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

While the MCC is mot commonly used when assessing binary classifiers, based on the definition of the MCC above, it can be extended to multi-class problems as described by Gorodkin[12]. If C the confusion matrix of a given classifier, then the MCC is given as,

$$MCC = \frac{\sum\limits_{k,l,m=1}^{N} C_{k,k}C_{m,l} - C_{l,m}C_{k,m}}{\sqrt{\sum\limits_{k=1}^{N}\left[\left(\sum\limits_{l=1}^{N} C_{l,k}\right)\left(\sum\limits_{f\neq k,g=1}^{N} C_{g,f}\right)\right]} \cdot \sqrt{\sum\limits_{k=1}^{N}\left[\left(\sum\limits_{l=1}^{N} C_{k,l}\right)\left(\sum\limits_{f\neq k,g=1}^{N} C_{f,g}\right)\right]}}$$

5.2 Feature Selection

As explained in section **??**, SFS was used to determine which features are of greatest value for each classifier. What was noticed, however, is that when running the SFS algorithm multiple times, the selected features were not always the same for each of the classifiers (even with cross validation). This was seen particularly in random forest (as might be expected for an algorithm that uses bootstrap aggregation). Therefore, the feature selection algorithm was run multiple times and the features that appeared most often were used. To evaluate the feature selection method, two additional sets of features were made (one for each classification problem) by choosing features based on how they appeared to seperate classes in **??**. All classification models are evaluated using the features found by SFS as well as the features found manually (labled as MAN). The lists of features used for each classifier are outlined in tables 2.1 and 2.2.

Children

SFS			**Manual**
Log Reg	**KNN**	**Rand Forest**	
Mon Daytime	Mon Evening	Thur Total	Month Total
Tue Evening	Mon Night	Sun Night	Sun Daytime
Wedy Night	Wed Evening	Mon Morning	Sat Evening
Thur Daytime	Fri Morning	Mon Daytime	Thurs Variance
Fri Morning	Sat/Weekday Ratio	Mon Evening	Mon Morning
Tue Variance	Sun Variance	Tue Daytime	Fri Evening
Thury Variance	Mon Variance	Fri Night	Sat Variance
Saty Variance	ρ(Mon Thur)	Thur Variance	Mon Total
ρ(Mon Tue)	ρ(Mon Friy)	Fri Variance	Sat Total
ρ(Wed Thur)	ρ(Tue Wed)	ρ(Mon Wed)	ρ(Mon Tue)

Table 5.1

Social Grade

SFS				**Manual**
Ord Log Reg	**Nom Log Reg**	**KNN**	**Rand Forest**	
Mon Morning	Fri Total	Fri Total	Sun Total	Mon Night
Tue Morning	Mon Morning	Tue Morning	Wed Total	Tue Variance
Tue Daytime	Tue Morning	Wed Night	Mon Morning	Mon Total
Fri Night	Tue Daytime	Thur Morning	Mon Night	Tue day
Fri Variance	Wed Morning	Fri Morning	Wed Morning	Sun Night
Sat Variance	Wed Daytime	Sat Evening	Wed Evening	Thur Night
ρ(Mon Tue)	Wed Evening	Sun Total	Fri Night	ρ(Tue,Fri)
ρ(Mon Fri)	Friday Morning	Thur Variance	Thur Variance	Thur Total
ρ(Wed Fri)	Fri Night	Fri Variance	ρ(Mon Fri)	Sat Night
First Fourier Feature	Sat Daytime	ρ(Wed, Thur)	First Fourier Feature	Fri Evening

Table 5.2

5.3 Classification Results

Here the results of the classifiers are presented of running each model on unseen data as outlined in 2.1. First the results of discriminating between households with and without children is presented, then the results of the socio-economic classifiers are shows. In the next section, we will the results are discussed to determine which classifier, and set of features, performs best on a given task and compared to the results obtained by similar studies..

The same training and test sets are used for each classifier to ensure the that the results are fair. While the training set is the same that had been used in cross-validation to optimize the classifiers, the test data is completely unused until this point.

5.3.1 Children vs No Children

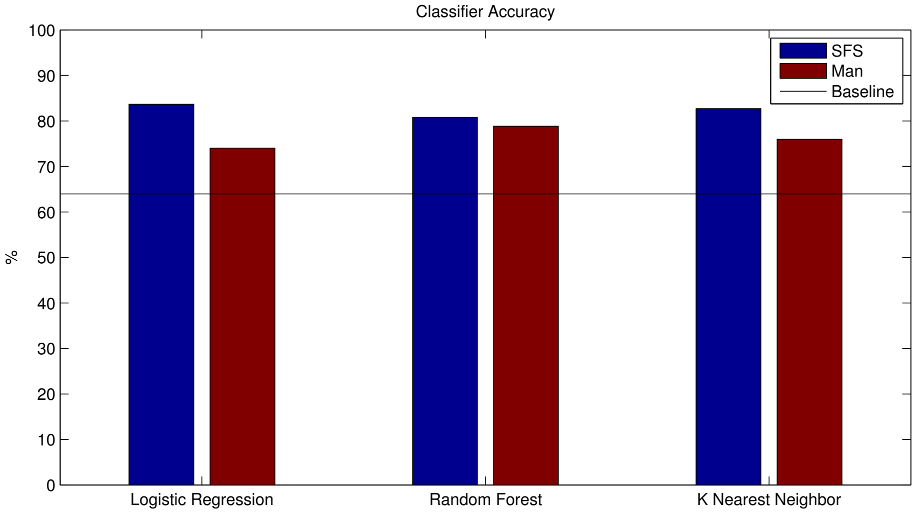


Figure 5.1: The accuracy of each classifier when used to classify unseen data defined as the number households classified correctly as a percentage of the total households in the test set. SFS classifiers are those built using features determined to be best by the sequential forward feature selection method while MAN are the features found by visually determining which features appear to best discriminate between the data

The results from Figure 2.1 show the accuracy of each classifier built to discriminate between households with and without children. All classifiers perform better than a biased random guess with logistic regression with features selected using SFS predicted the greatest percentage of households correctly. Furthermore Figure 2.2 gives the Matthews correlation coefficients of each of the classifier which all take on positive values, meaning that the classifiers are less inclined to missclassify a household regardless of which class has a higher prior probability.

While the logistic regression classifier using the SFS features classifier gave both the highest accuracy and MCC, the performance is not significantly better than random forest and Knn. What is noticeable, is that all classifiers performed better when trained with features found by SFS features compared with manually feature selection, with the accuracy of all three classifiers that used SFS having an accuracy of above 80% and all those with manyally selected features being below 80%. The distinction is even more pronounced when considering the MCCs of each of the classifiers, where all classifiers using SFS features have an MCC of over 0.5, while none of the ones using manually selected features do. In the case of logistic regression, there is the greatest difference in performance with the SFS classifier having an MCC of 0.64 and the MAN classifier only having an MCC of 0.42.

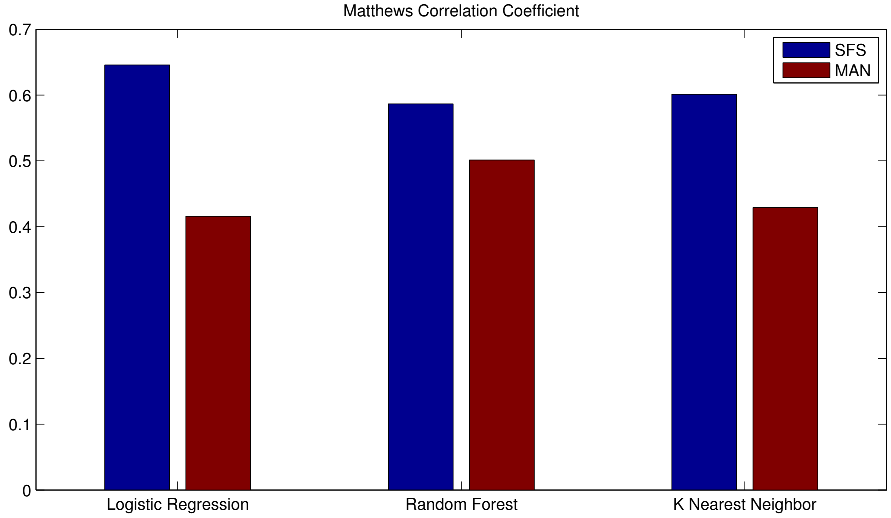


Figure 5.2: The Matthews correlation coefficient of each classifier when used to classify unseen data. SFS classifiers are those build using features determined to be best by the sequential forward feature selection method while MAN are the features found by visually determining which features appear to best discriminate between the data

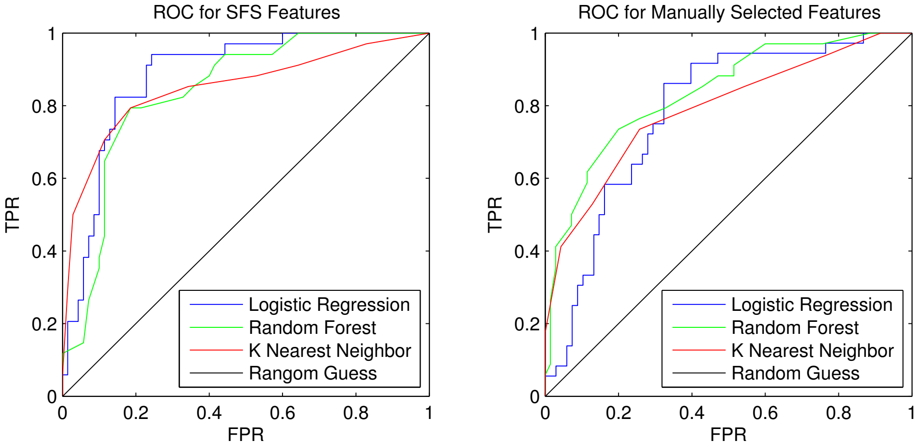


Figure 5.3: Reciever operating curve curves that show the trade-off between true positive rate and false positive rate.

The ROC curves shown in figure 2.3 shows how the TPR changes when varying the FPR. Again, it can be seen that SFS features generate better results than MAN features and that, again, the logistic regression classifier gives the best performance (indicated by the area under the curve). Especially, for Knn, the TPR quickly increases with little increase in FPR, however once the specificity (1-FPR) decreases, the performance of the random forest and logistic regression classifier become more accurate than knn.

Finally, the table below shows the TPR and TNR for each of the classifiers along with the other statistics discussed previously. While the logistic repression classifier did have the greatest accuracy and MCC, it does not have the highest TPR, meaning that it was not best at identifying positive instances (households without children), as opposed to the Knn classifier which was able to identify

Children vs No Children

	Accuracy (%)		**MCC**		**ROC Area**		**TPR**		**TNR**	
Classifier	SFS	MAN	SFS	MAN	SFS	MAN	SFS	MAN	SFS	MAN
Log Reg	83.6538	74.0385	0.6457	0.4159	0.884	0.7831	0.8429	0.8235	0.8235	0.5833
Random Forest	80.7692	78.8462	0.5866	0.5012	0.8397	0.8357	0.8143	0.8857	0.7941	0.5882
KNN	82.6923	75.9615	0.6013	0.4289	0.8502	0.7905	0.8857	0.8714	0.8857	0.5294

88.6% of households without children. Looking at the TNR of each of the classifiers, again, Knn gives the best results, identifying 88.6% of households that have children when using the SFS features, compared to only 79% using a random forest classifier. This indicates, that although the logistic regression classifier had the greatest accuracy, Knn is better at correctly identifying the households from those without. From looking at the TNR, it can also be seen that the models built using MAN features rely much more on the imbalances in the sample sizes whereas models trained using SFS features are able to perform better than just a random guess.

5.3.2 Socio-Economic Group

While the results from the precious section showed that it is possible to discriminate between households with and without children, the classifiers built to determine a household's socio-economic status are not as promising. By looking at the confusion matrix generated by each classifier, as shown in Table 2.3, insight can be gained into how each of the classifiers predicts unseen data. The first thing that is noticed is that the ordinal logistic regression models are little better than a biased random guess. They predict almost all test instances as either class C1 or C2, which are the two classes with the highest prior probability based on the sample population ($p(C1) = 0.38$,$p(C2) = 0.25$). While the nominal logstic regression classifiers are not significantly more accurate, they are not as heavily biased towards the most probable classes. This indicates that the proportional odds assumption is not satisfied.

Confusion Matrix

	SFS	**MAN**
Ord Log Red	$\begin{pmatrix} 1 & 0 & 4 & 1 & 0 & 0 \\ 0 & 0 & 3 & 5 & 0 & 0 \\ 1 & 0 & 7 & 18 & 3 & 0 \\ 0 & 0 & 4 & 32 & 4 & 0 \\ 0 & 0 & 3 & 11 & 1 & 0 \\ 0 & 0 & 0 & 6 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 2 & 4 & 0 & 0 \\ 0 & 0 & 1 & 7 & 0 & 0 \\ 0 & 0 & 3 & 26 & 0 & 0 \\ 0 & 0 & 2 & 38 & 0 & 0 \\ 0 & 0 & 1 & 14 & 0 & 0 \\ 0 & 0 & 0 & 6 & 0 & 0 \end{pmatrix}$
Nom Log Reg	$\begin{pmatrix} 4 & 0 & 0 & 2 & 0 & 0 \\ 1 & 0 & 1 & 2 & 3 & 1 \\ 0 & 0 & 11 & 16 & 1 & 1 \\ 2 & 1 & 4 & 29 & 4 & 0 \\ 2 & 0 & 1 & 9 & 2 & 1 \\ 0 & 0 & 1 & 3 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 & 0 & 4 & 0 & 0 \\ 1 & 0 & 0 & 3 & 4 & 0 \\ 1 & 0 & 9 & 16 & 2 & 1 \\ 3 & 1 & 4 & 31 & 1 & 0 \\ 0 & 0 & 0 & 14 & 1 & 0 \\ 0 & 0 & 1 & 4 & 0 & 1 \end{pmatrix}$
Random Forest	$\begin{pmatrix} 4 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 5 & 0 \\ 0 & 1 & 18 & 9 & 1 & 0 \\ 1 & 1 & 5 & 26 & 5 & 2 \\ 0 & 0 & 1 & 7 & 6 & 1 \\ 0 & 0 & 1 & 2 & 2 & 1 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 4 & 2 & 0 \\ 0 & 1 & 20 & 7 & 2 & 0 \\ 2 & 1 & 6 & 25 & 4 & 4 \\ 1 & 0 & 3 & 6 & 4 & 1 \\ 0 & 0 & 3 & 1 & 1 & 1 \end{pmatrix}$
Knn	$\begin{pmatrix} 4 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 4 & 2 & 0 \\ 0 & 1 & 16 & 9 & 1 & 2 \\ 2 & 1 & 8 & 24 & 5 & 0 \\ 1 & 0 & 2 & 9 & 3 & 0 \\ 0 & 0 & 1 & 3 & 2 & 0 \end{pmatrix}$	$\begin{pmatrix} 3 & 0 & 20 & 1 & 0 & 0 \\ 1 & 1 & 0 & 2 & 4 & 0 \\ 1 & 0 & 20 & 6 & 2 & 0 \\ 1 & 2 & 6 & 28 & 2 & 1 \\ 1 & 0 & 0 & 7 & 7 & 0 \\ 0 & 0 & 3 & 0 & 2 & 1 \end{pmatrix}$

Table 5.3: Each confusion matrix shows the number of households that were classified as a given class, compared to their true class. Each row represents the true label of an instance while the columns represent the predicted label, i.e element (i, j) represents the number of households that were classified as j but who's true class is i. The rows and columns go in order of increasing social grade, so the 1^{st} row/column represents class E, the second represents class D ect.

From the confusion matrices it can also be seen that all of the classifiers give a very low probability of a household being in class D. This is not uprising as the figures outlined in Chapter **??** suggested that it is difficult to distinguish between households of socio-ecomic classes D,C1 and C2, and since D has a much lower prior probability that the other two, a probabilistic classifier is likely to assign an unseen instance to either class C1 or C2. The random forest classifier does, however,

Looking at Figure **??**, it can be seen that, although the accuracy of the classifiers starts to quickly increase as more of its neighbors are considered positive, the benchmark accuracy also increases rapidly. This is because the most probably classes are in the middle of the odering and are therefore more likely to be a neighboring class neighbor.

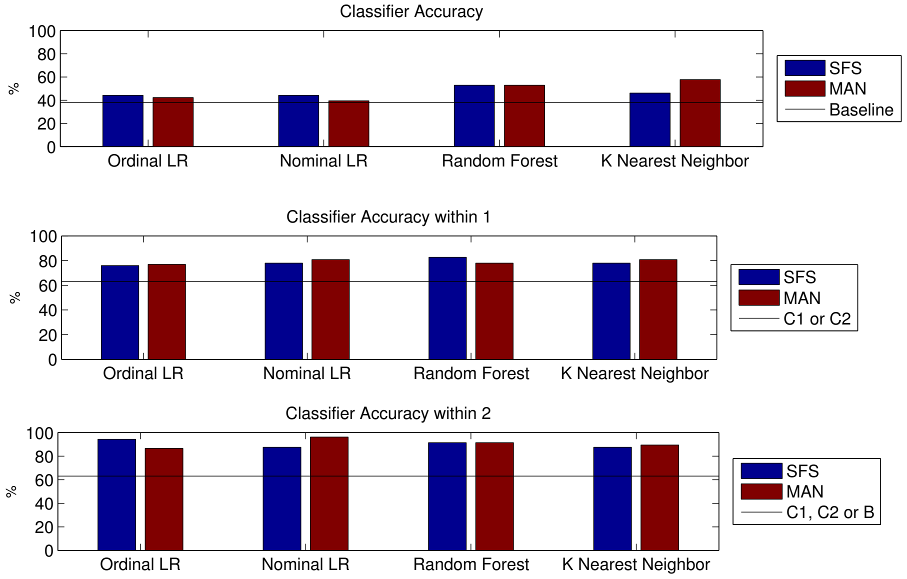


Figure 5.4: The accuracy of each of a classifier is given by the sum of the diagonal elements of its confusion matrix. The accuracy with n is the sum of the diagonal elements in addition to the elements up to n columns to the left and right of the diagonal.

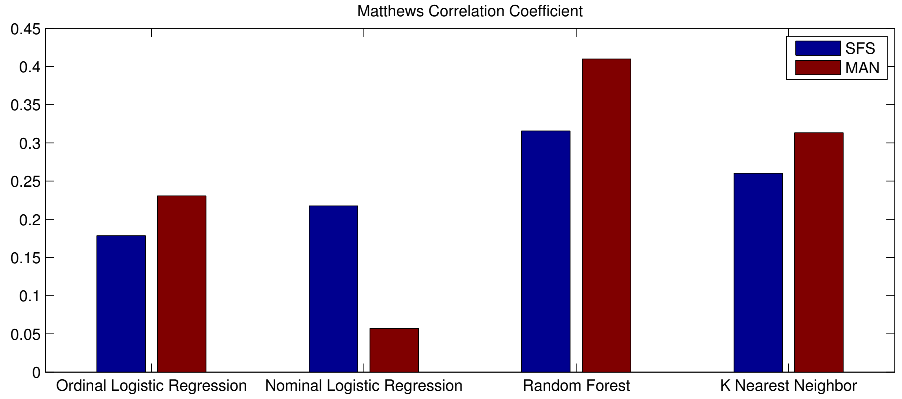


Figure 5.5: The Matthews correlation coefficient of each classifier when used to classify unseen data. SFS classifiers are those build using features determined to be best by the sequential forward feature selection method while MAN are the features found by visually determining which features appear to best discriminate between the data

Figure 2.5 shows the MMCs of each classifier. logistic regression performs poorly MAN features tend to do better than SFS. This could likely be because SFS is a greedy algorithm and, particularly for multi-class problems with limited amounts of data to perform validation on, a backwards step such as that used in SFFS could render a more optimal set of features. Random forests are less effected by this as they, in a sense perform as they determine which feature is best at each stage of being built. The ROC curves generated by the classifiers is given in Apendix A, however and confirm that the classifiers' performance is only useful when trying to extract households of the classes C1 and C2.

Bibliography

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

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

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

[4] 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.

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

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

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

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

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

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

[11] David M W Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation. Technical report, University of South Australia, 2007.

[12] J. Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational Biology and Chemistry*, 28(5-6):367–374, 2004.