Machine Learning with Domestic Energy Use Data

Sam Stern (s1134468)

March 10, 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 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 and (2) the

[1] Department for Environment, Food and Rural Affairs

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 the house sets and relevant household and energy-use information;

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 researchers 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 specifit client groups.

- Chicco gives an overview of the clustering techniques used to establish suitable client groups for analysing electricity load pattern data [5]. Cao et.al also grouped consumers using electricity load profiles, however focusing on finding households with the same peak usage [6].

- Others, such as Zoha et. al and Carrie et.al have performed research on NILM (*non intrusive load monitoring*). Taking aggregated energy consumption data from households and disaggregating the consumption of the constituent appliances.[7][8].

- 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 cheif 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 [9]. While much of the of the work presented in the report overlaps with that done the authors, we concider a different set of classifiers (random forrest and logistic regression) as well as a nother class of features taken from the time-frequency transorm of the data.

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 [10]. 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 [11].

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 aggregate, (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 kilowatt hours (kWh).

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.

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 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.

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.

The last stage in pre-processing was to ensure that each instance was of the same length. As explained in TBD, 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.

2.3 Issues

1. Homes were not perfectly representative of the population
    - only homeowners were included
    - only concidered homes in England, not the entire UK
    - class size ratios not representative of population
2.
3. The purpose of the project was to determine whether it is *possible* to distinguish between households, and to show how this might be achieved.
4. Several households have periods where their energy consumption pattern vanishes and very little or no energy is used. It is likely that these are periods where the members of the household are away or on holiday.
5. The 'total' electricity is not always well estimated.
6. Initially, data from the IDEAL study was going to be used however as this was not available, data from the HES study was used. This resulted in a delay to the project.

2.4 Comparison to Previous Work

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 [12]. 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 [14] [15]. 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.

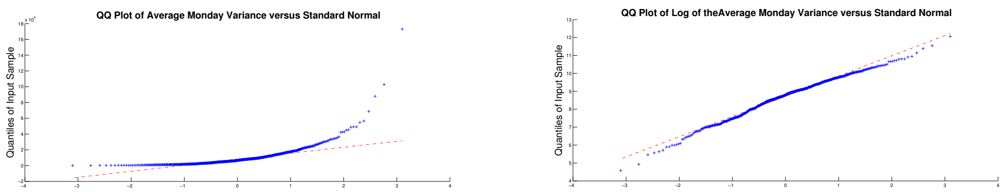


Figure 3.1

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 is potentially wasteful of the limited resources available to do the project.

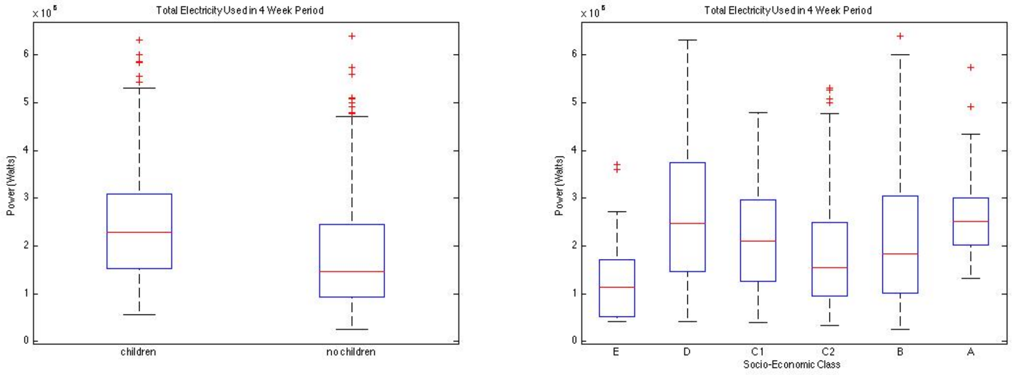
Instead of creating features in an ad hoc manner, a more cost efficient approach was taken.Feature selection was done in the following way: 1) Assumptions were made regarding the distinction between classes (e.g., households with children use more energy overall). 2) Features were created to capture 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 [16].



(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 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.

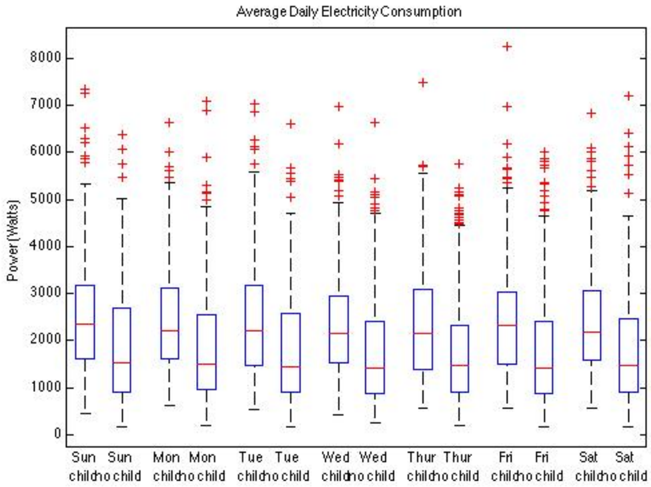


Figure 3.3: The average total energy used on each day of the week. Households are grouped by whether or not there are children present

Figure 3.4

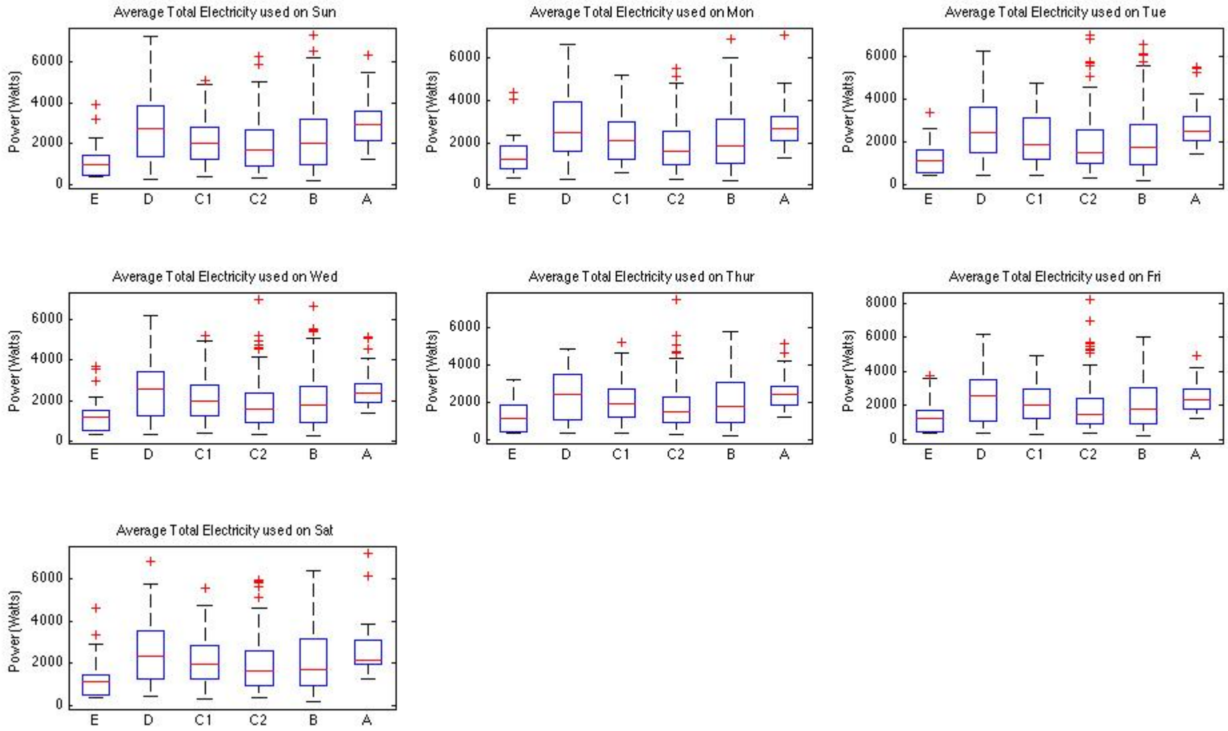


Figure 3.5: The average total energy used on each day of the week. Households are grouped by their IPSOS social grade

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 [17]. 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 [18]. 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. Afternoon (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.

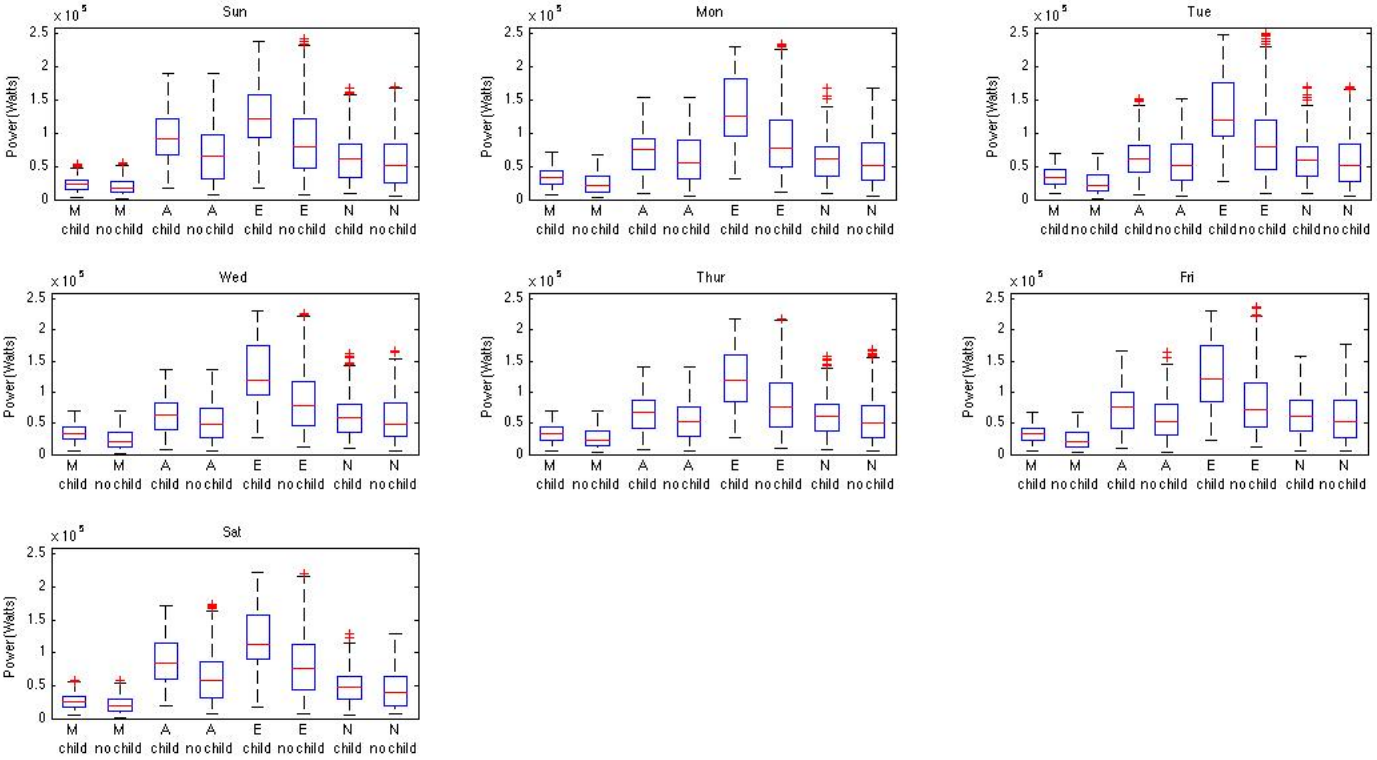


Figure 3.6

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 afternoons (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.

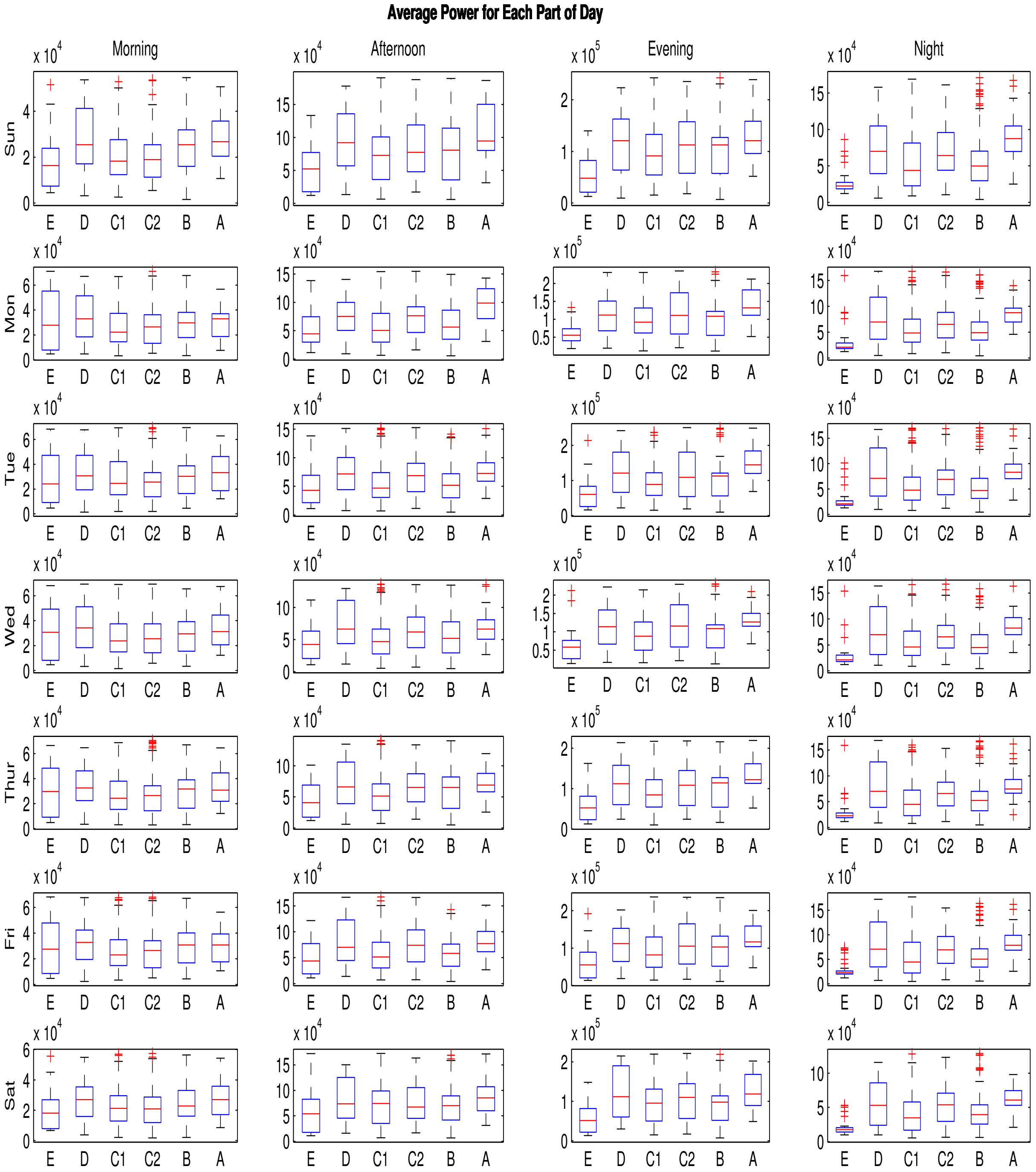


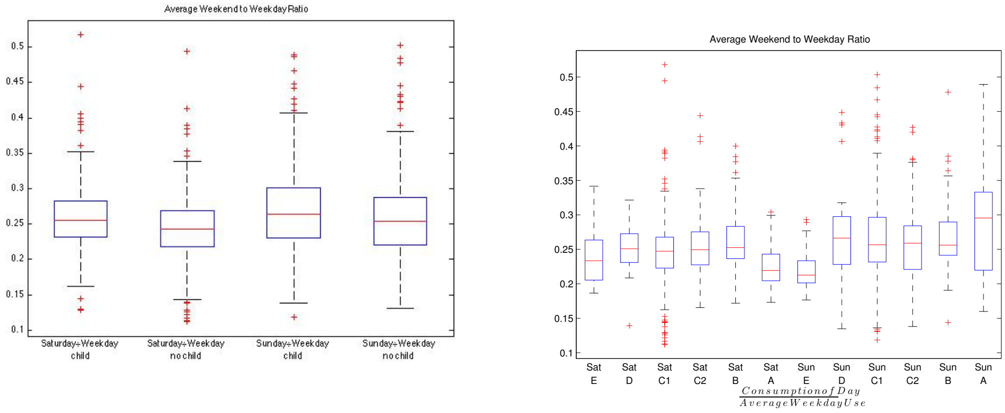
Figure 3.7

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. 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 3.8 suggests that households use more of their energy on Sundays than they do on Saturdays, this is independent of the socio-economic their class and therefore is unlikely to be of use in distinguishing between classes.

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.

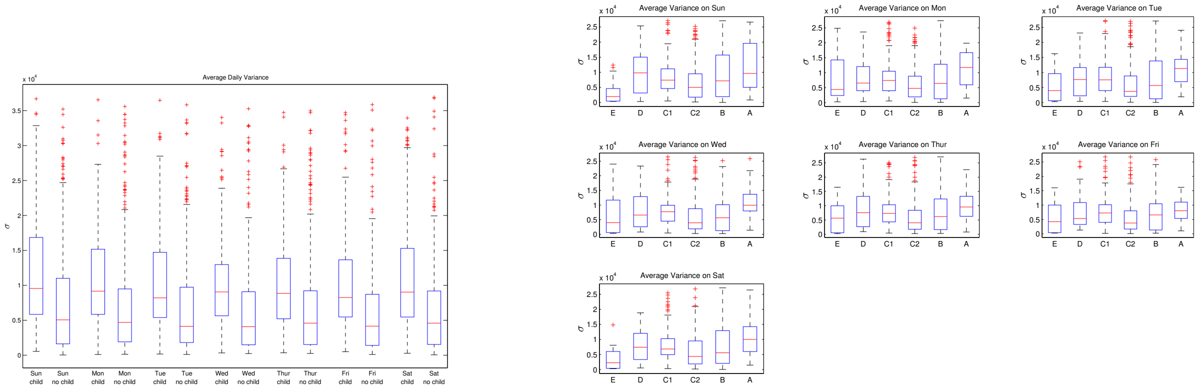


Figure 3.9

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. Furthermore, the skew indicates that it might be beneficial to take a transformation of the feature, such as the logarithm, the results of which are plotted in Figures 3.10 and 3.11. Here it can be seen that households in socio-economic group C2 tend to have lower volatility in their daily consumption than some of the other classes. This is of interest because the consumption features failed to distinguish between the middle socio-economic classes.

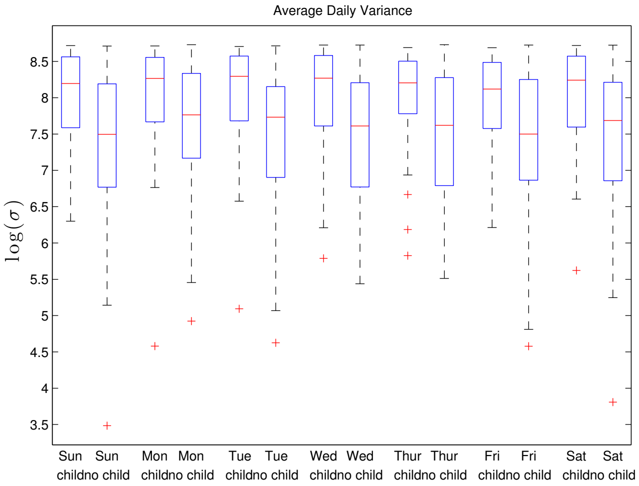


Figure 3.10

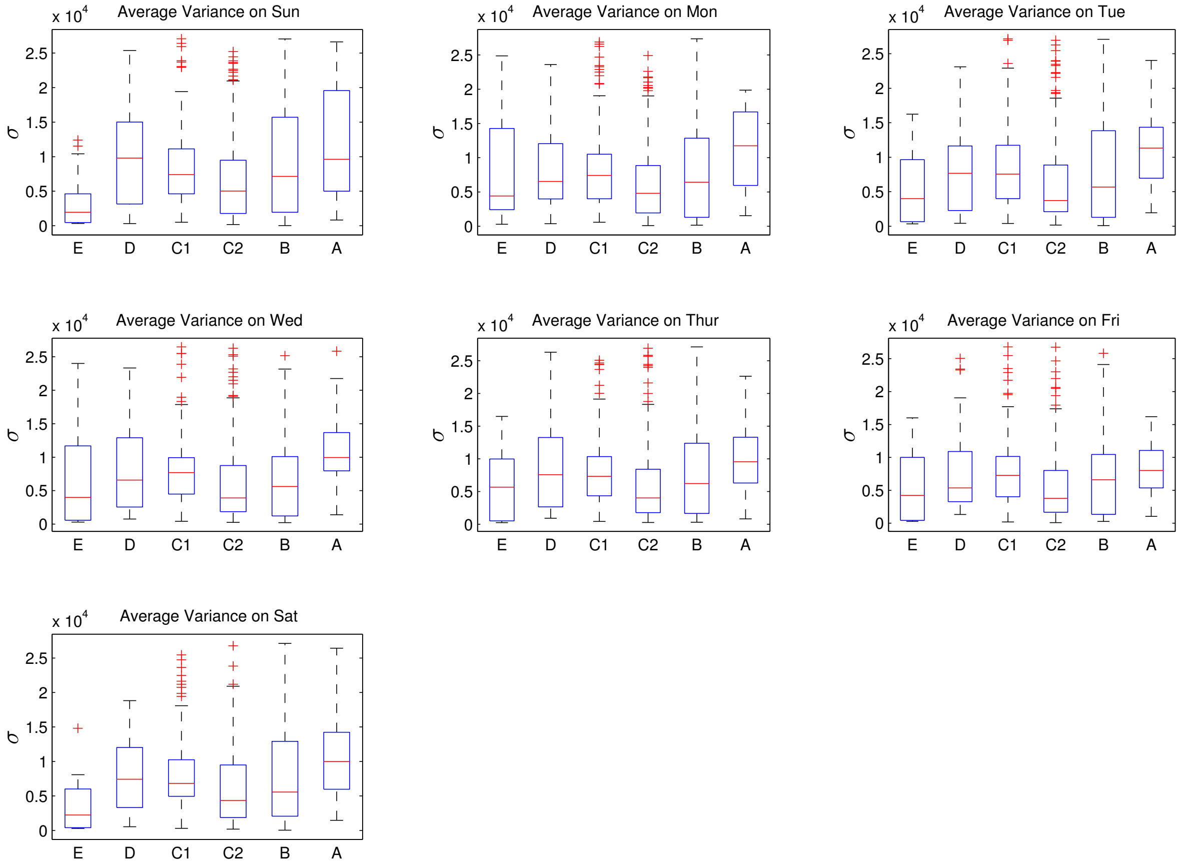


Figure 3.11

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.

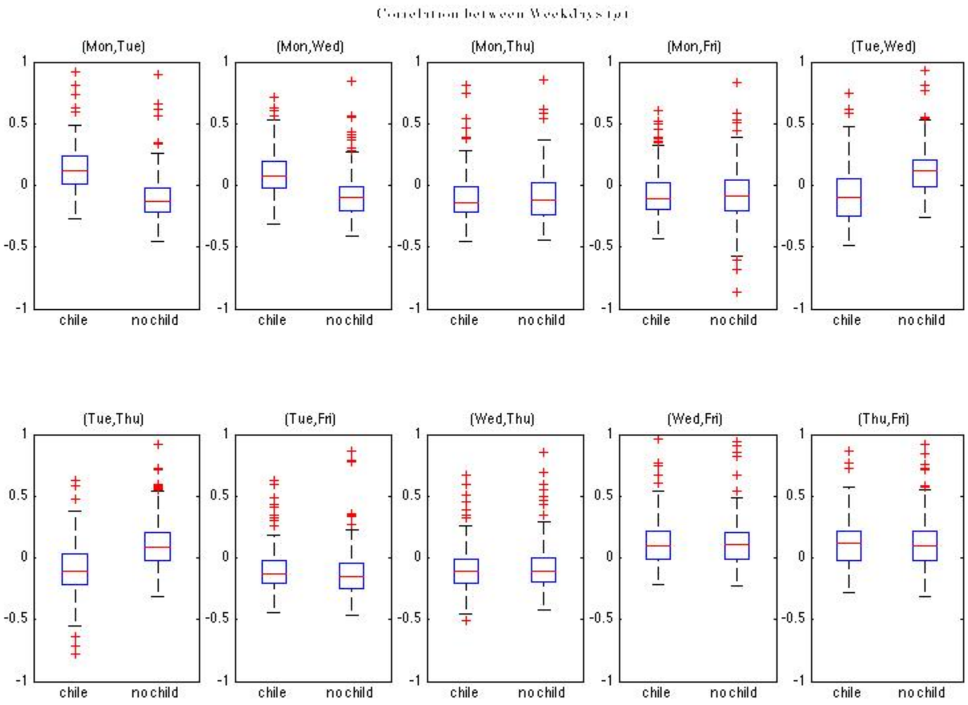


Figure 3.12

Looking at Figure 3.12, 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.

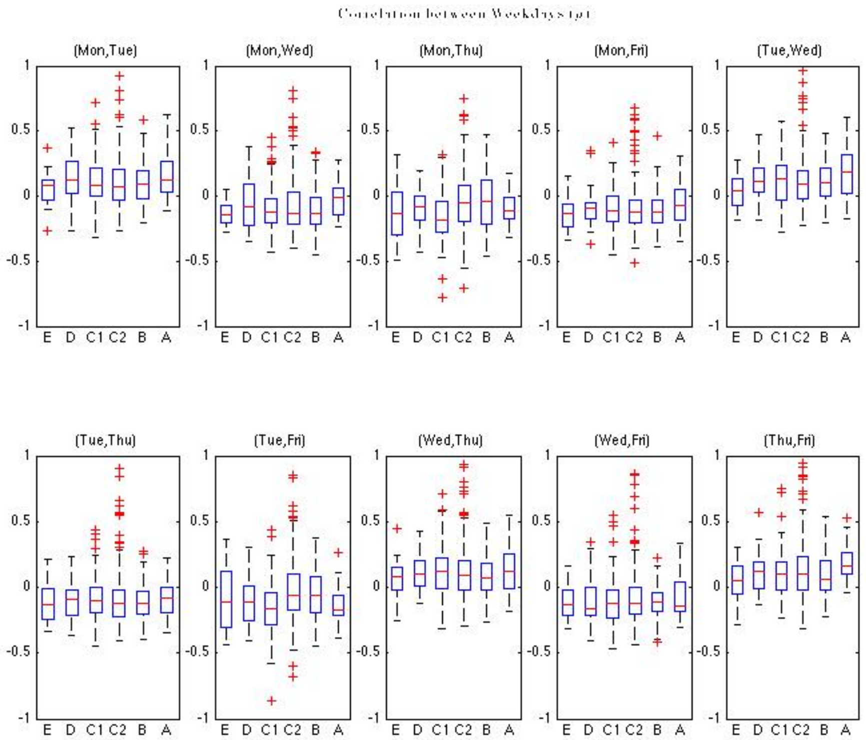


Figure 3.13

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 [12] for time series feature extraction are used to project each household's consumption into the frequency domain from which the most important frequencies are used as features. McLoughlin *et. al.* [19] showed in their research that temporal structure is present in household electricity consumption data and can be used to characterise 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 can improve performance compared with direct averaging, as more structure is retained whilst noise is removed [20]. This is done as the time-frequency transformation used (the discrete Fourier transform method) has difficulty characterising small intervals of large electricity demand [21]

Gaussian filtering (or Gaussian smoothing) is performed by convolving the time series with the Gaussian function.

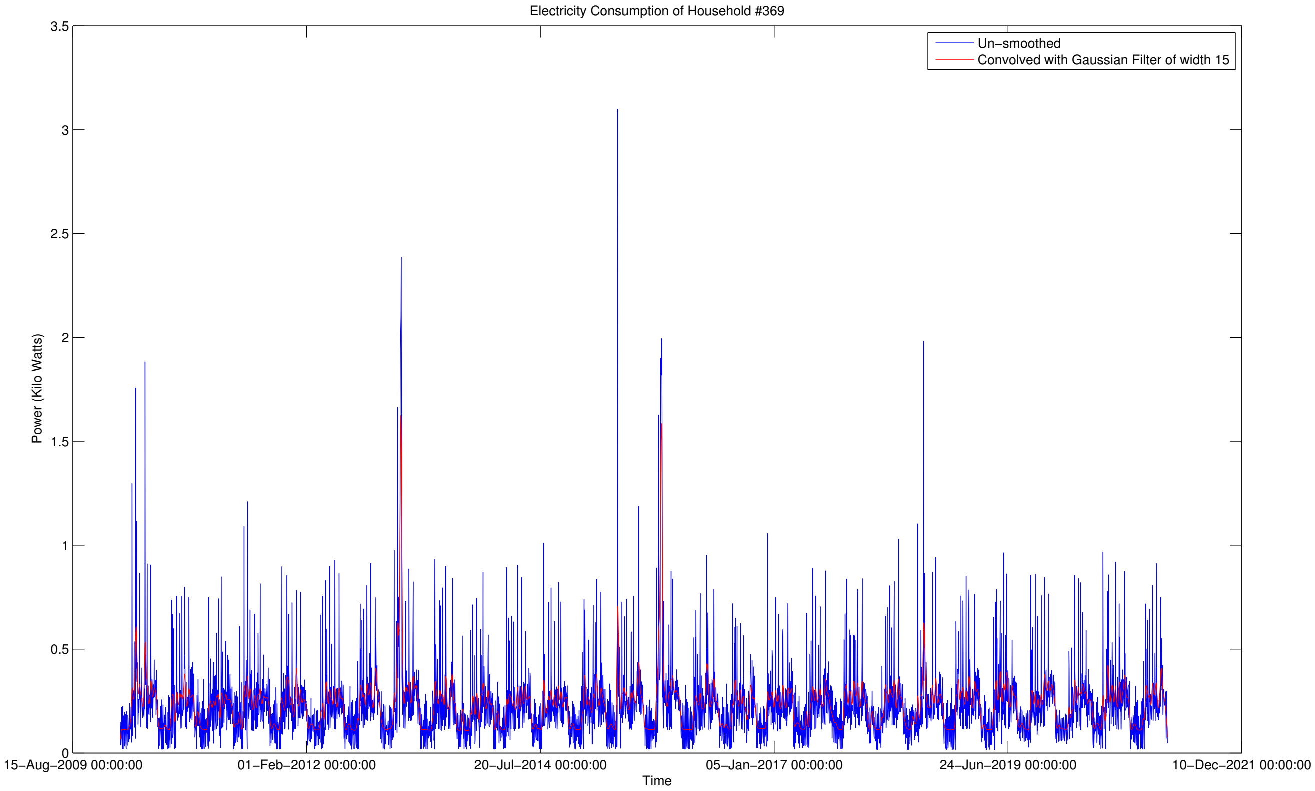


Figure 3.14: The electricity use of household No.369 shows that households may have both a daily and weekly pattern. The clusters of peaks represent individual days while the regions without peaks are the indicative of night time. Additionaly, the large spikes are observed roughly every seven days, on 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 [12].

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 aproximations at the boarders of the time series [12].

Energy Preservation

For l time series of length m, the DFT produces an $l \times m$ matrix C of coefficients of l rows and m columns, 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 [12], 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 correspondent to signal $f(t)$.

Let I be a function measuring the importance of coefficient j on all l values, 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 [12] 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, and 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, they are all dependent on the input data. In order to achieve accurate results with the least amount of computational time, it is necessary to insure that as little noise and redundancy is present in the input. Feature selection is the process of identifying and filtering out as much irrelevant and redundant information as possible [22].

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

Like k-nearest neighbour, the need for dimensional reduction is less to do with removing redundancy, but more to reduce noise and computational cost. logistic regression will account for highly correlated features by lowering their weight, however uninformative features would cause learn weights that do not improve the performance of the classifier.

Random forrests are not as succeptible to the problem of overparameteristaion 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 (therefore avoiding overfitting). An issue would only start to arise when the number of redundant or noise 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 subet where no good features are present becomes significant.

Dimensionality reduction can usually be charachterised as one of two tasks: *Feature selection* and *feature transformation*. Feature transformation methods involve performing a transformation the data (such as a rotation of projection) to create a new set of features (of smaller size) that has ore descriptive power than the origional set. An commonly used example of this is *principal component analysis* (PCA) which finds a set of orthogonal unit vectors which point in the directions of greats variance of the data. The features are then just the result of 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 [23].

As is might be of interest to see which features are more responsible for differences between classes. Therefore, instead of using feature transformation methods, instead feature selection is used to find. There exist numerous methods of performing feature selection such as nested subset methods, filters or direct objective optimisation [23], as well as adaptive boosting [24].

We use *sequential floating selection* (SFS) [25] to find the optimal set for features. Starting with an empty list, SFS sequentially consideres each feature not present in the list for selection and assesses it's impact on a given evaluation score, choosing the feature that scores best and adding it to the list. This is repeated until the list is full [26]. A superior mechod *sequential forward floating selection* has been proven superior [25], which backtracks after a new feature is included to solve the *nesting* problem, it proved inefficient to implement for the multi class.

3.5 Class Cardinality

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 [?]. 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 is directly modelling what we want, $p(y|\mathbf{x})$. 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[27]. We are also fortunate in that there is no missing data.

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 pros 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.

- interpretable -can look at the weights
- gives a confidence in our predictions -probability

4.2.1 Multi-class Logistic Regression

4.2.2 Implementation

4.3 Random Forrest

-

4.4 K-Nearest Neighbour

- number of instances is relatively small so parametric classifier won't be too expensive.

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 proble, 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. [28]

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 knowlege 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[29].

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 porportion 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 preffered over the accuracy, particularly when considering unbalanced classes as the impact of skewness can be analysed [30].

This method of evaluation can be easily applied to the binary classification task of discriminating between households with and without children. 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, 9]. 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] Gianfranco Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68–80, 2012.

[6] 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.

[7] Ahmed Zoha, Alexander Gluhak, Muhammad Imran, and Sutharshan Rajasegarar. Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. *Sensors*, 12(12):16838–16866, 2012.

[8] K. Carrie Armel, Abhay Gupta, Gireesh Shrimali, and Adrian Albert. Is disaggregation the holy grail of energy efficiency? the case of electricity. *Energy Policy*, 52:213–234, 2013.

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

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

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

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

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

[15] 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.

[16] 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.

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

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

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

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

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

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

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

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

[25] 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.

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

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

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

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

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