Machine Learning with Domestic Energy Use Data

Sam Stern (s1134468)

March 1, 2015

Abstract

As part of the UK Government's insentive to reduce the Nation's energy consumption, smart meters are being rolled out to households and small businesses accross the UK. In this project aims to assess some of the security risks associated with gathering data relating to a households energy consumption.

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 [7], 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 [8]. 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 [5] [6]. 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 from its energy consumption readings, and if so, if the results would be reliable, and 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 household electricity consumption information collected by the Household Electricity Survey (HES), a DEFRA sponsored national survey of energy use collected over a period from 2010 to 2011, classifcation models are created to predict two household properties: (1) The presence (or absence) of children in a household and (2) the IPSOS social grade of the chief income earner of the household. 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.

1.2 Smart Meters

Following the example of EU Countries such as Italy, Sweden, Finland, Switzerland and Germany [3][4],

1.3 Related Work

In recent years researchers have been putting much effort into it in focusing on N I LM (nonintrusive load of monitoring)Using aggregated electricity readings from households researchers tackle the problem of disaggregating thisBad consumption into its constituent appliances.

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), which was a survey undertaken by the 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 of the study was to identify and catalogue the range and quantity of electrically powered appliances found in a typical home, understand the household's frequency and patterns of electricity usage and to collect 'user habit' data that emerge from using a range of appliances [11].

The HES study monitored 250 households, of which 26 were monitored for one year while the remaining 224 were monitored for roughly one month. Each household had between 13 and 85 individual appliances being monitored in their homes such that when aggregated (as outlined in section 2.2), the result gives 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 information about the household such as the number of occupants, employment status, IPSOS social-grade and whether there are 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 the total energy used as was required for this project. The HES study recorded measurements for 250 possible appliances that a household could have (giving values of 0 to appliances that weren't monitored). The resulting raw data was large csv files with largely redundant entries.

The first step in pre-processing the data was to use create a MySQL database and import the the appliance readings into a table. Cambridge Architectural Research Ltd had additional files that mapped which appliances needed to be aggregated for each household in order to create 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 participated in the HES study, which is a relatively small number for a machine learning task as there might not be enough data to build models that accurately sample the entire English 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 measure 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 is important since some consumption features, would be affected by a difference 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 timeseries instances were manipulated so that they 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 [13]. It is therefore beneficial to transform and aggregate the data in such a was as to reduce the dimensionality as well as capture differences in the consumption patterns between classes.

According to [2], possible features that are interesting for classification of households based on energy consumption are: consumption figures, ratios, temporal properties, and statistical properties. Consumption figures are the average, maximum and minimum energy consumption over some time period. Ratios are features that calculate the ratio between consumption figures and can capture relevant patterns that occur through different time intervals. Temporal features capture the first (or last) time some event takes place which or at what time the daily maximum occurs or any periodicity within the household's electricity consumption. Finally, statistical properties, such as variance, give insight into the consumption curve (for example how a households energy consumption correlates with itself.

Numerous statistical methods expect the input data to follow a normal distribution. Therefore, the data was visualized and compared against a normal quantile plot in order to find the right non-linear transformations [17] [**?**]. 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. Thiese transformations are important for classifiers such as k-nearest neighbor which rely on the distance between samples based on their features

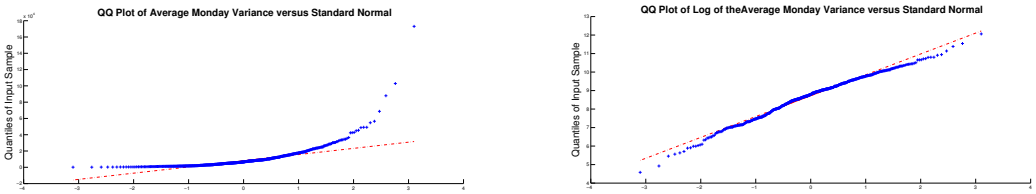


Figure 3.1

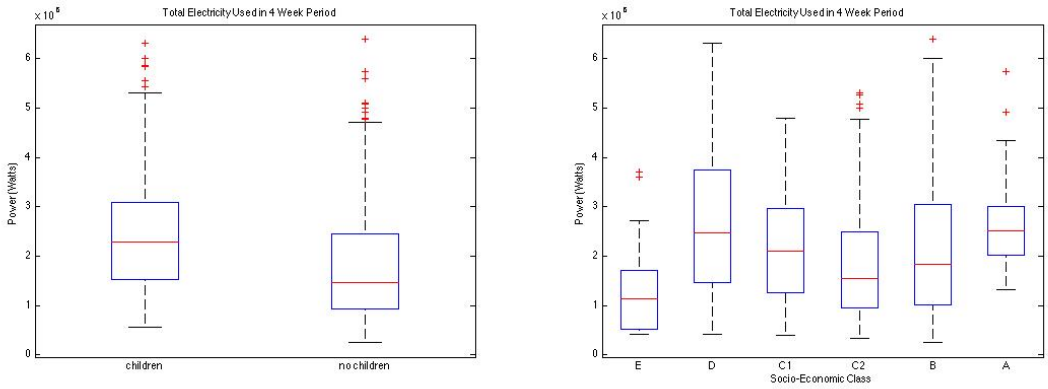
3.2 Creating Features

One method of extracting features would be to compute as many different types as possible, compare them all and chose those that best discriminate the classes. households could 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 to do the project. Instead of creating features in an ad hoc manner, feature selection was done in the following way: 1) An 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 the 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/didn't discriminate between classes while other times, more sophisticated methods were used, as described in 3.3.

The rest of this section describes features that were created from the energy reading data and justifies why they were may have been able to discriminate between classes. Both classification problems (socio-economic classification and child classification) were considered when choosing features to evaluate.

3.2.1 Total Electricity

When 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 1500 Watts per 10 minutes, others averaged as little as 65 Watts per 10 minute, and while one household consumed up to 19500 Watts in a 10 minute period, another never used more than 1190. Therefore, the first feature that was explored was the total energy consumed in a give period of time. Since it was, at this stage, not known if other factors such, as time of day and the day of the week, have an influence the consumption. Therefore 28 day timeperiods ensured independence from these. 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 used by the household [14]



(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 figure, which compares the households with children against those without, shows that those with children do indeed tend to use more energy. The right hand graph, which compares the total electricity grouped by social grade does indicate 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 use more energy than others, it is worthwhile to dig deeper and determine if there is any factor that influences 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 energy than others, but if it is dependent on the day of the week.

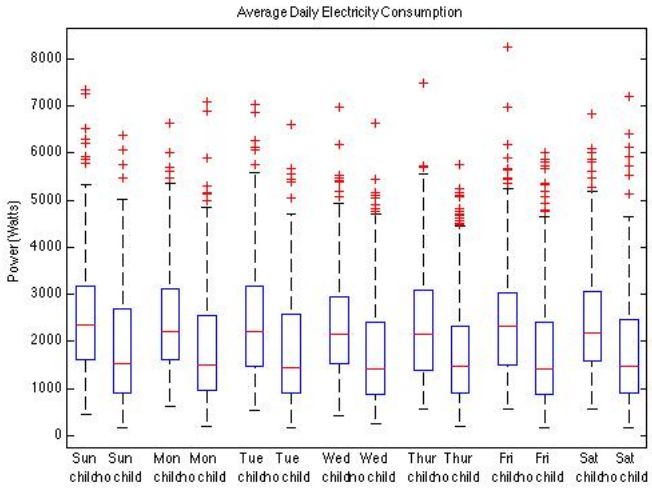


Figure 3.3: The average total energy used on each day of the week. Households are grouped by whether or not there are children present

Figure 3.4

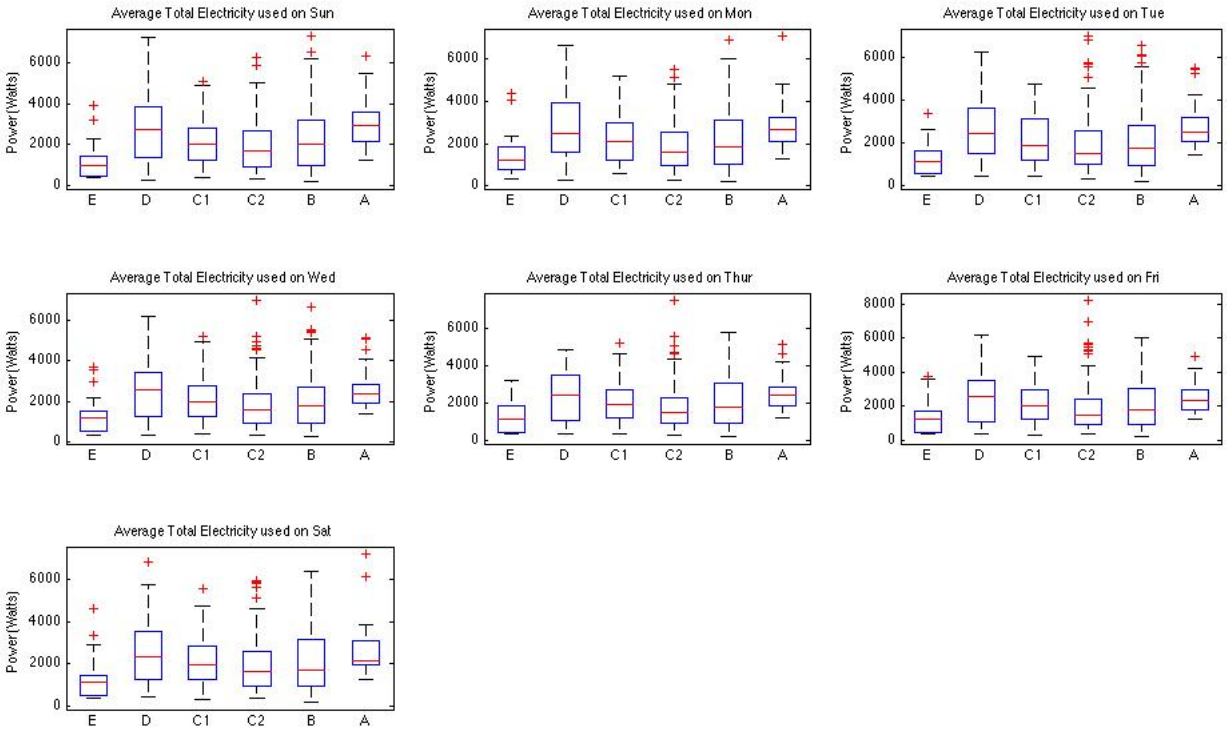


Figure 3.5: The average total energy used on each day of the week. Households are grouped by their IPSOS social grade

While figure 3.4 does further show that households with children use more than those without, it does not give any more insight into 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, looking at the average daily usage of different socio-economic groups does not give any more understanding of the differences between classes. There is no particular day where the differences in electricity consumption between classes is different 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, that 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 employed [16]. 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.

According to [15], most schools days in England begin at 9:00 and finish between 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. 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.

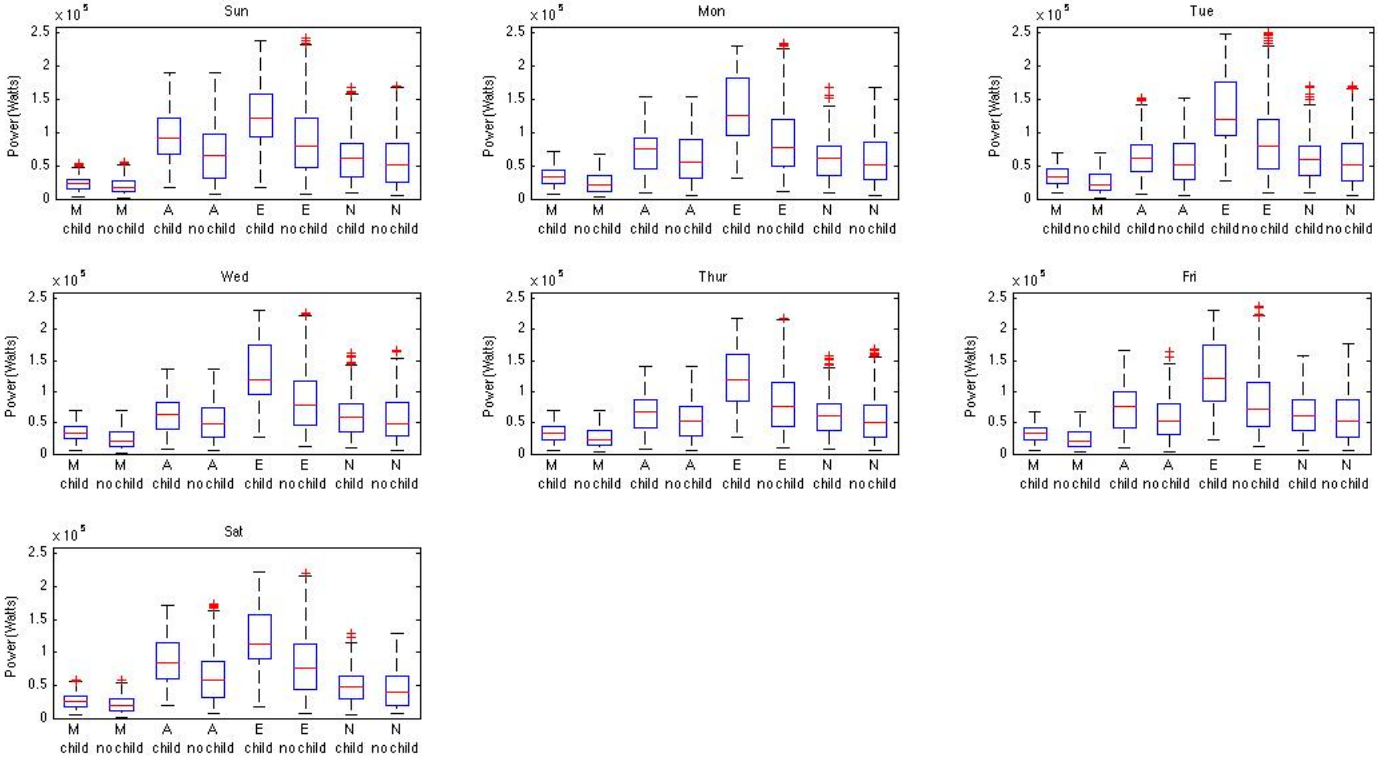


Figure 3.6

The data portrayed in figure 3.6 does indicate that the energy use patterns are indeed different for households with and without children. We see that much of the differences in household electricity consumption can be attributed that used in the evenings, 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.

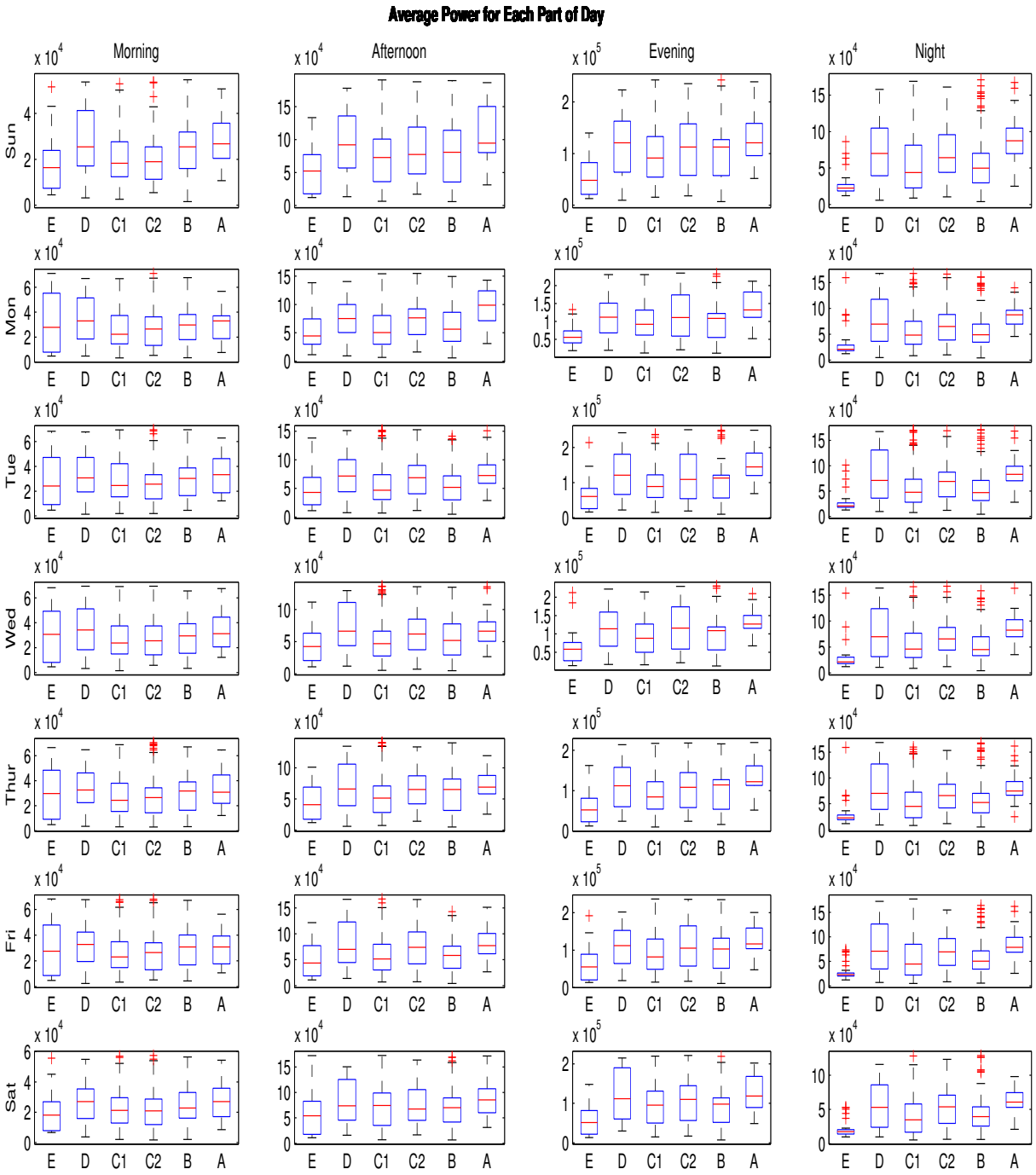


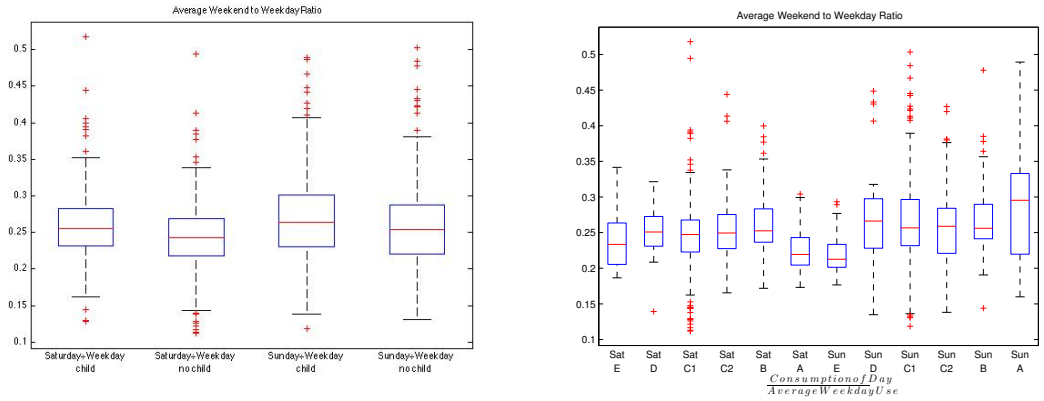
Figure 3.7

Figure 3.7 Shows again the same results as the previously computed features. Households of social grade E appear to use particularly 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 appear to have the opposite pattern, using more energy than others in the evening 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 used on an average weekend day and weekday is capable of determining if a household is using proportionally more of it's energy during the week or the weekend. The rational being that households of social grades E,D and C2, whose chef 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(b) The ratio between how much energy is used on the weekends and how much is used on weekdays. Householdsweekends and how much is used on weekdays. Households are grouped on whether or not there are children present 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 household's use more of their energy on Sundays than they do on Saturdays, this is independent of the socio-economic class and therefore is unlikely to be of use in distinguishing between classes.

Variance on Weekday

Thus far, the features that have been computed are 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.

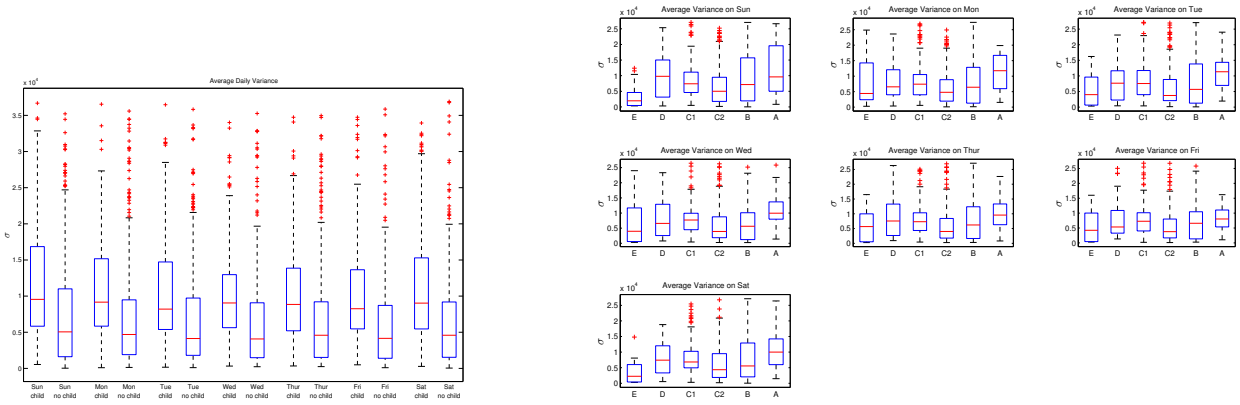


Figure 3.9

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 skewness 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 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.

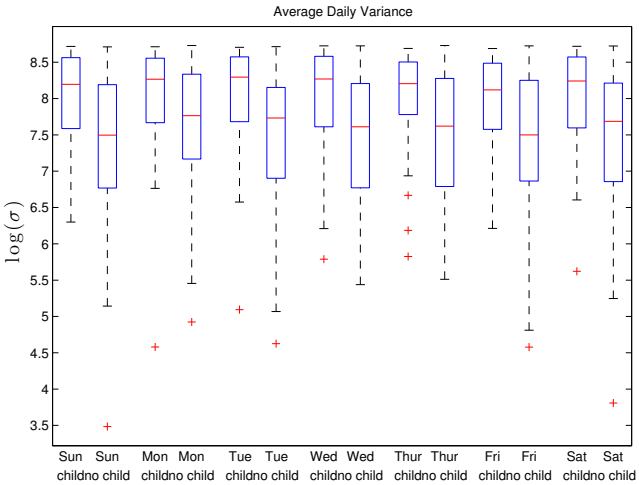


Figure 3.10

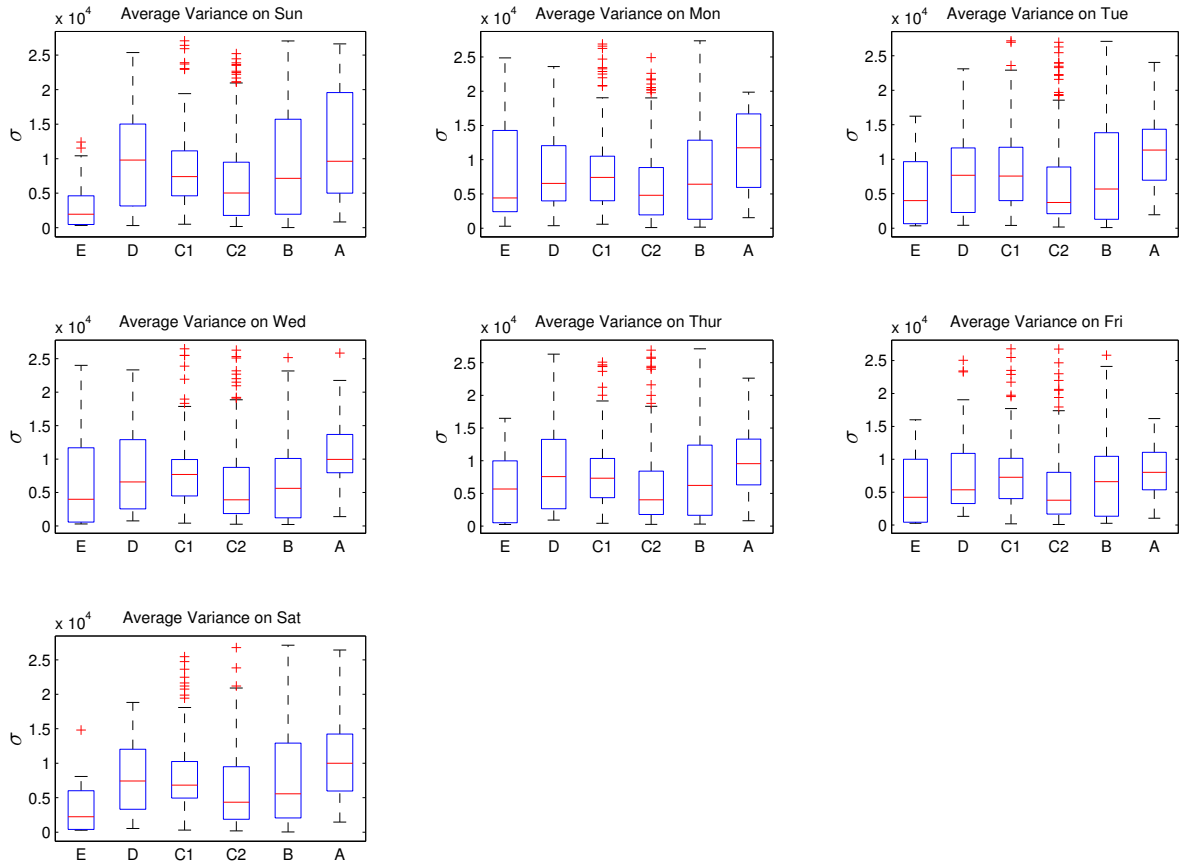


Figure 3.11

Correlation Between Weekdays

The average correlation coefficient between a 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.

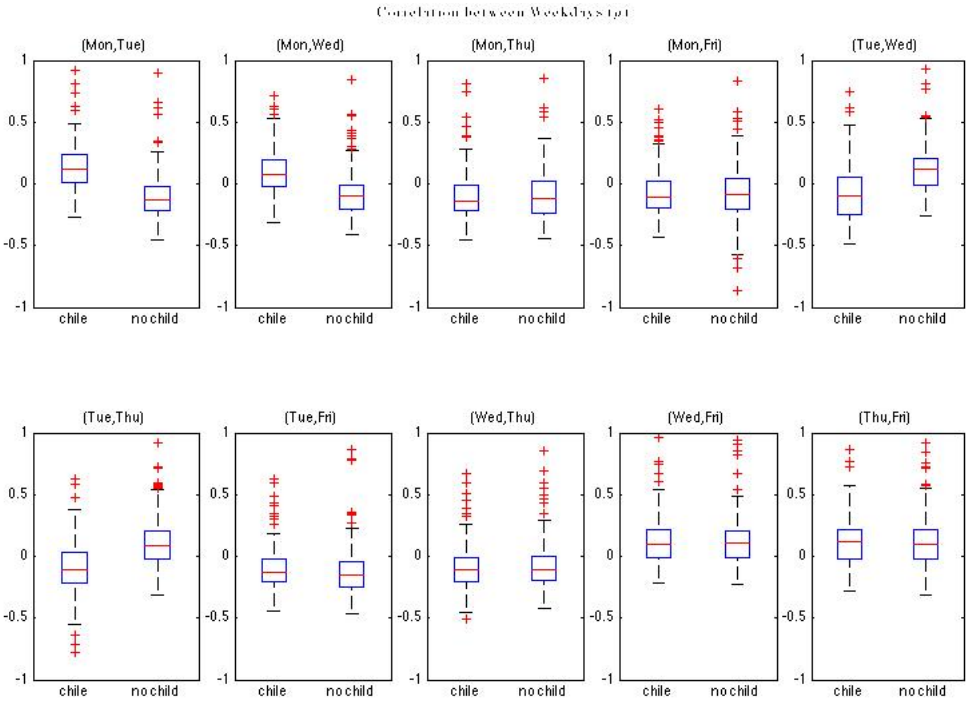


Figure 3.12

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 and that depending on which two days are being considered, the correlations of one class tend to be greater or smaller than the others. For example, it would appear that households with children have a slightly higher correlation between their Monday and Tuesday electricity use pattern than those without.

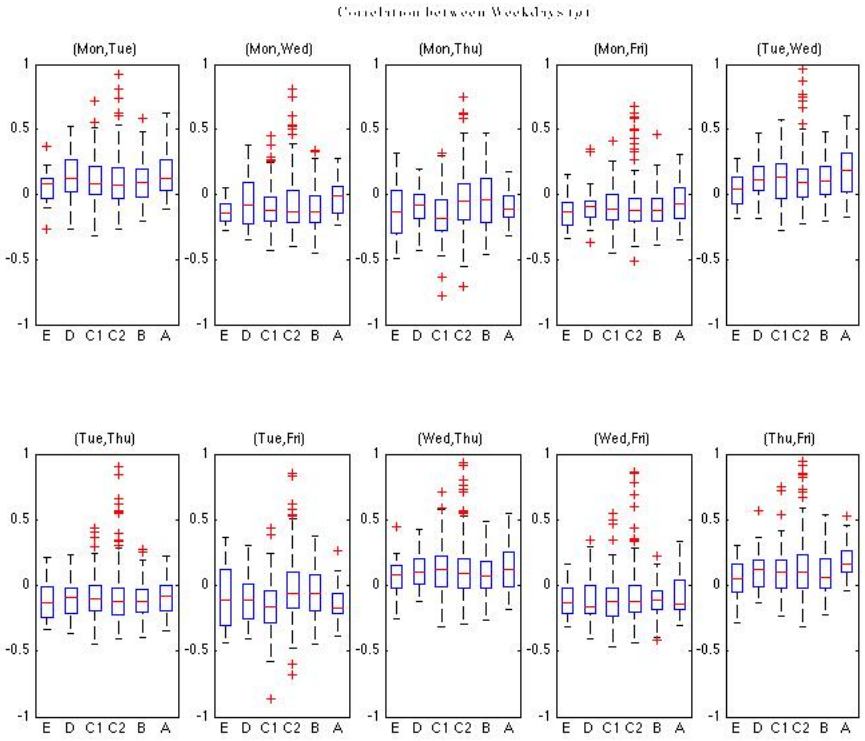


Figure 3.13

3.2.2 Periodicity

Another approach used for feature extraction was to exploit the periodic consumption patterns exhibited by many of the households in order to search for temporal structures that are present in some classes but not others. This method of feature extraction has been used particularly in studies involving forcasting and clustering. Methods outlined by Fabian Mörchen [13] for time series feature extraction are used to project the household's consumption into the frequency domain from which the most important frequencies are used as features. McLoughlin *et. al.* [20] showed in their research that temporal structure is present in household electricity consumption data and can be used to charachterise domestic energy demand.

3.2.3 Signal Smoothing

Before projecting the electricity consumption into frequency space, the Gaussian averaging operator was applied to each set of readings filter the noise whilst retaining the temporal structure of the data. Gaussian filter can improve performance compared with direct averaging as more structure is retained whilst noise is removed [12]. This is done as Fourier transforms have 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.

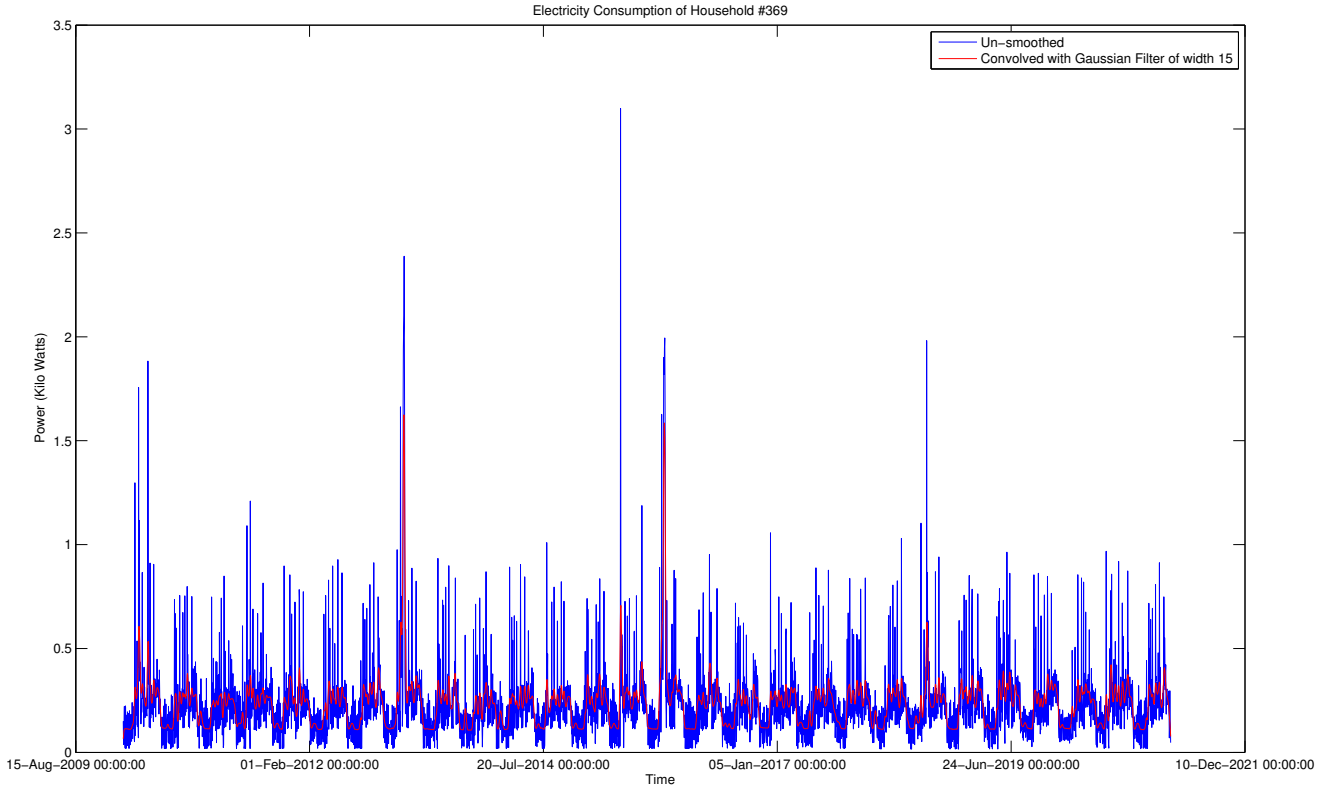


Figure 3.14: The electricity use of household No.369 shows that households may have both a daily and weekly pattern. The clusters of peaks represent individual days while the regions without peaks are the indicative of night time. Additionaly, the large spikes are observed roughly every seven days, on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

3.2.4 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 [13].

Doesn't handle discontinuities [21]. The fourier transform measure global frequencies and the signal is assumed to be periodic. This assumption can cause poor aproximations at the boarders of the time series [13].

3.3 Feature Selection

3.4 Class Cardinality

Bibliography

[1] Christian Beckel, Leyna Sadamori, and Silvia Santini. "Automatic socio-economic classification of households using electricity consumption data". In *Proceedings of the Fourth International Conference on Future Energy Systems*, e-Energy '13, pages 75–86, New York, NY, USA, 2013. ACM.

[2] Christian Beckel, Leyna Sadamori, and Silvia Santini. "Towards automatic classifi- cation of private households using electricity consumption data". In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, BuildSys '12, pages 169–176, New York, NY, USA, 2012. ACM.

[3] Office of Gas and Electricity Markets (OfGEM). *Transition to smart meters.* `https://www.ofgem.gov.uk/electricity/retail-market/metering/transition-smart-meters`.

[4] Jorge Vasconcelos. "Survey of regulatory and technological developments concern- ing smart metering in the european union electricity market". `http://hdl.handle. net/1814/9267, 2008.`

[5] Elias Leake Quinn. "Privacy and the new energy infrastructure" (february 15, 2009). `http://ssrn.com/abstract=1370731` or `http://dx.doi.org/10.2139/ssrn.1370731`.

[6] M.A. Lisovich, D.K. Mulligan, and S.B. Wicker. "Inferring personal information from demand-response systems". *Security Privacy, IEEE*, 8(1):11–20, Jan 2010.

[7] Office for National Statiscics. 2014. "Full Report: Household Energy Spending in the UK, 2002-2012". [ONLINE] Available at: `http://www.ons.gov.uk/ons/dcp171776_354637.pdf`. [Accessed 26 January 15].

[8] STOPSMARTMETERS. 2012. Stop Smart Meters. [ONLINE] Available at: `http://stopsmartmeters.org.uk/`. [Accessed 26 January 15].

[9] Smith Steven W. (1997) *The Scientist and Engineer's Guide to Digital Signal Processing* California Technical Pub

[10] Household electricity survey. `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/278809/10043_R66141HouseholdElectricitySurveyFinalReportissue4.pdf`, January 2014.

[11] Jason Palmer, Nicola Terry, Tom Kane *Early Findings: Demand side management* 17 June 2013.

[12] Mark S. Nixon and Alberto S. Aguado. *Feature Extraction and Image Processing*. Academic Press, 2008, p. 88.

[13] Fabian Moerchen (2003) *Time series feature extraction for data mining using DWT and DFT*

[14] Leticia M. Blázquez Gomez, Massimo Filippini, Fabian Heimsch (2013) *Regional impact of changes in disposable income on Spanish electricity demand: A spatial econometric analysis* Energy Economics 40:58–66

[15] Teaching Jobs — Supply Teaching Jobs - Teaching Personnel 2015 `http://www.teachingintheuk.com/go/uk-teaching-info/school-system/`

[16] Mel Bartley, Charlie Owen "Relation between socioeconomic status, employment, and health during economic change, 1973-1993" *BMJ* 1996;313:445

[17] Osborne J. *Notes on the use of data transformations*. Pract Assess Res Eval 2002;8(6):1e8.

[18] Wang MC, Bushman BJ. *Using the normal quantile plot to explore meta-analytic data sets*. Psychol Methods 1998;3(1):46e54.

[19] C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X.

[20] Fintan McLoughlin, Aidan Duffy, Michael Conlon. *Evaluation of time series techniques to characterise domestic electricity demand* Energy, 50 (2013), pp. 120–130

[21] Graps A.*An introduction to wavelets*. IEEE Computational Science and Engineering Summer 1995;2(2) [Online]. Available from: http://www.amara.com/IEEEwave/IEEEwavelet.html [accessed: 30.11.2011].