Machine Learning with Domestic Energy Use Data

Sam Stern (s1134468)

February 25, 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 footprint [?] and the British public becoming increasingly frustrated by rising energy bills will little explanation to the cause in this rising price [7], the UK Government is currently executing on a plan to distribute smart meters to households accross the country by 2020. Smart meters, which measure a household's gas and electricity consumption in real-time, are expected to both help a household reduce its energy usage by displaying how much energy is being used , as well as increase transparency in the household's bills by eliminating the need for monthly meter readings and estimations by the energy companies. Instead, the energy providers are given a much more accurate description of the household's consumption and as a result, will be able give a more accurate bill.

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 profit [8]. Perhaps more interestingly though, and therefore the focus of this project, are the concerns which have been have been raised regarding the risk associated with measuring and storing energy consumption data [5] [6]. Particularly, to what extent can other information about a household be inferred from energy consumption readings?

The aim of this project is to explore whether (and to what extent) it is possible to construct features that can be used to predict detailed personal information of a household from their energy consumption readings, by taking on the role of a malicious individual (or group) who wishes to exploit this information to determine household properties that might be of interest to someone wishing to either target advertise or burgle a household. 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, classification 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 the information gathered by the HES survey, they are of logical 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

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 [12]. 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.

3.2 Creating Features

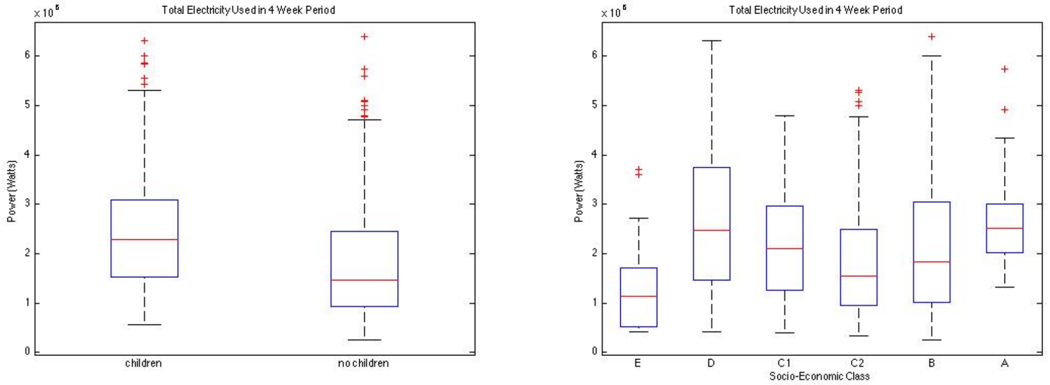
One method of extracting features would be to compute as many different types as possible, compare them all and chose those that best discriminate the classes. households could be further split into weeks, days and even hours. Consumption figures and statistical properties can then be measure for each 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.

- total energy -different classes use more energy than others
- mean weekday, saturday and sunday -different classes use more energy than others on working days or non-working days
- variance on weekday - different classes will be more active than others (i.e unemployed might be at home using appliances, however the appliances that the employed use require require more power)
- APOD
- APOD/total_energy ratio - patterns like whether cooking takes place over lunchtime of in the evenings or both
- correlation between average days of week

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 [13]



(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 3.1

Looking at **??**

Average Daily Usage

As previously alluded to, visualising the timeseries for individual households indicated that there are differences in the energy consumption depending on what part of the week is being considered. 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 of the day of the week.

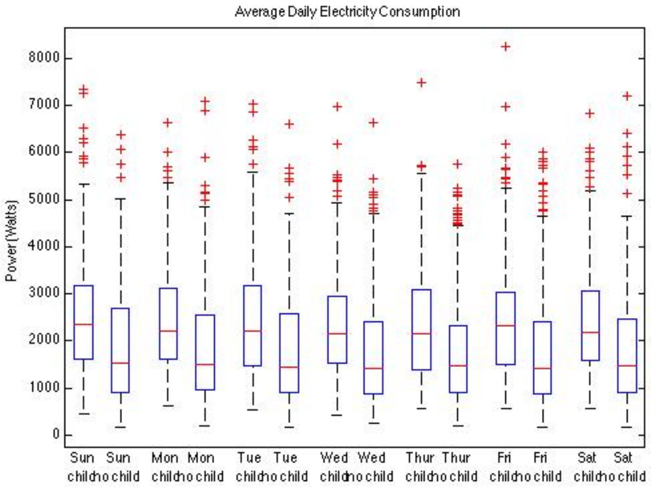
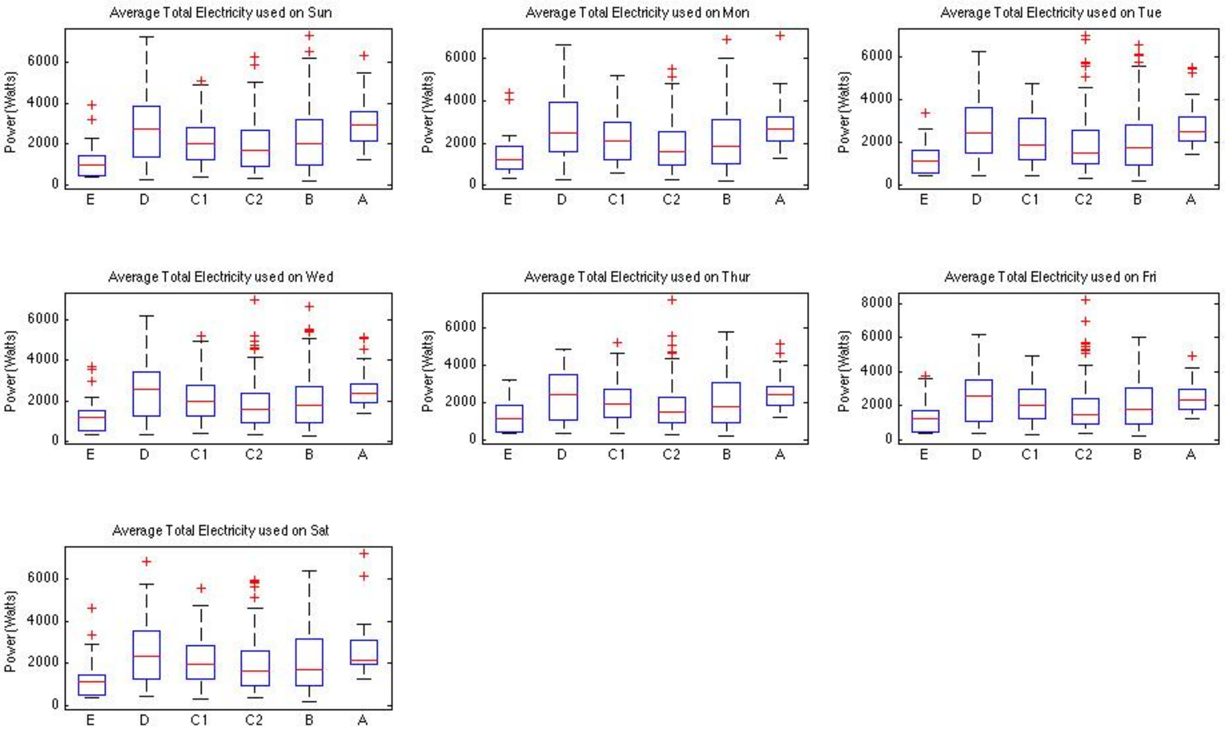


Figure 3.2: The average total energy used on each day of the week. Households are grouped by whether or not there are children present

Figure 3.3: •



Average Total Electricity used on Sun

Average Total Electricity used on Mon

Average Total Electricity used on Tue

Average Total Electricity used on Wed

Average Total Electricity used on Thur

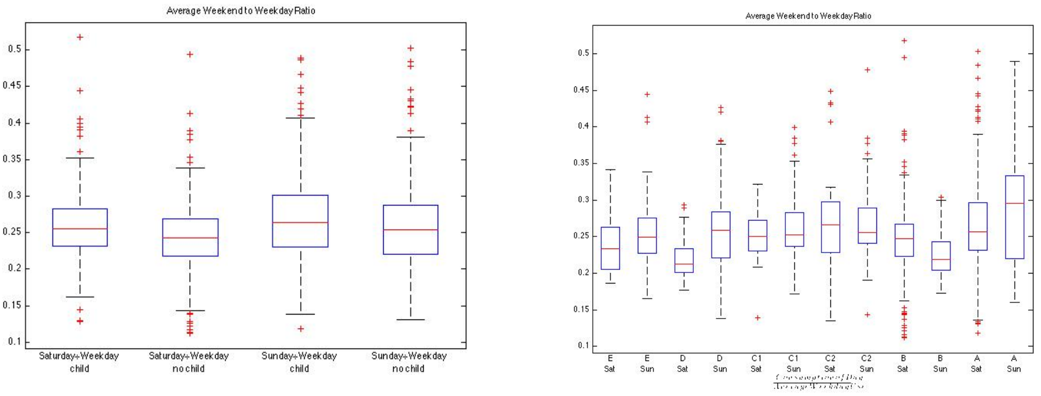
Average Total Electricity used on Fri

Average Total Electricity used on Sat

Figure 3.4: The average total energy used on each day of the week. Households are grouped by their IPSOS social grade

Mean Weekday vs. Saturday and Sunday

As well as investigating the differences in how much energy different classes use, it is also worthwhile to see when households use most of their energy, both interday and intraday.Starting with interday, it is reasonable to assume that, while some households will use roughly the same amount of energy each day, others might use proportionally more on the weekends versus weekdays to Comparing the weekday electricity consumption to that of the weekend.The justification being that manual laborers and shift workers might not use any more or less energy, while households where the cheif income earner is managerial might use more of their energy on the weekends.



(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

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.

since households vary dramatically in their energy use profiles, the plot had to remove outliers.

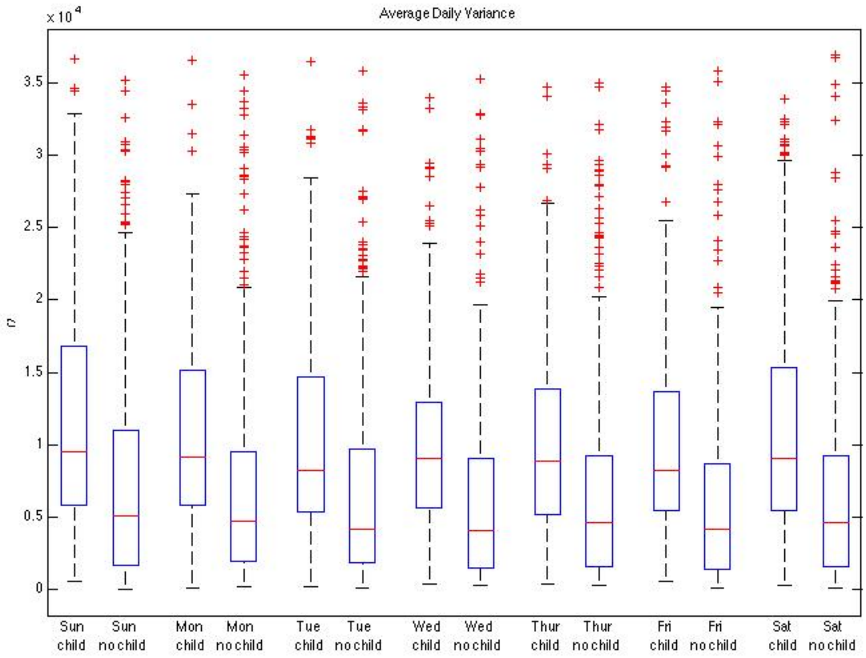


Figure 3.6: •

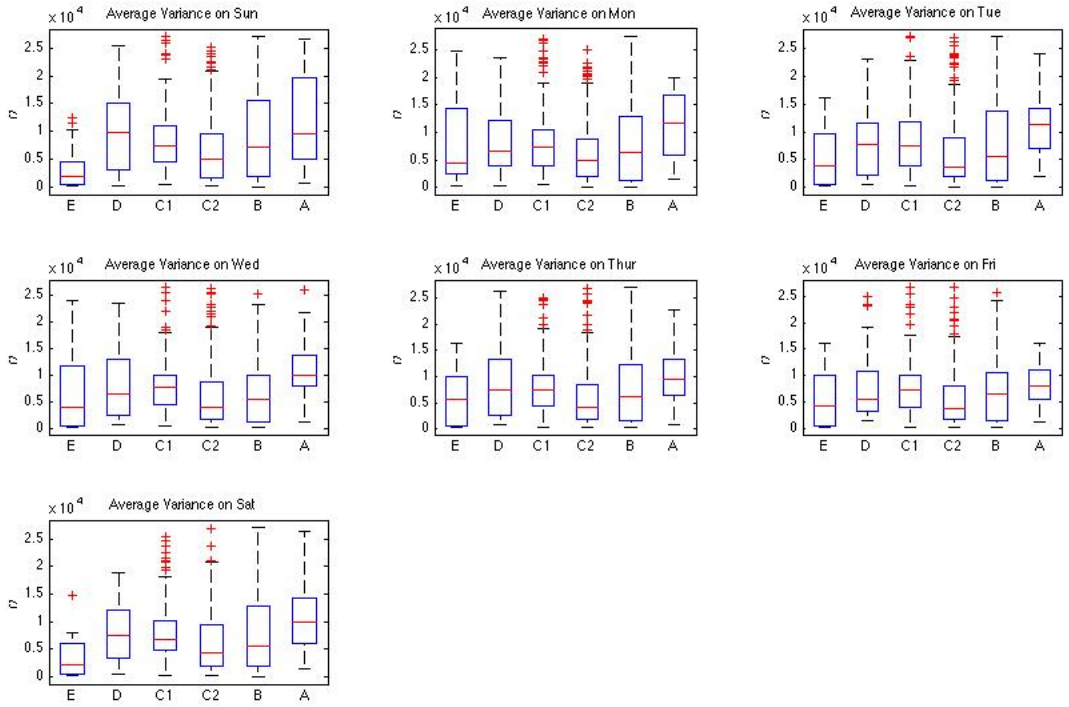

Average Variance on Sun
Average Variance on Mon
Average Variance on Tue
Average Variance on Wed
Average Variance on Thur
Average Variance on Fri
Average Variance on Sat
× 10^4
σ
0
0.5
1
1.5
2
2.5
E
D
C1
C2
B
A

Figure 3.7: •

Average Part-Of-Day (APOD)

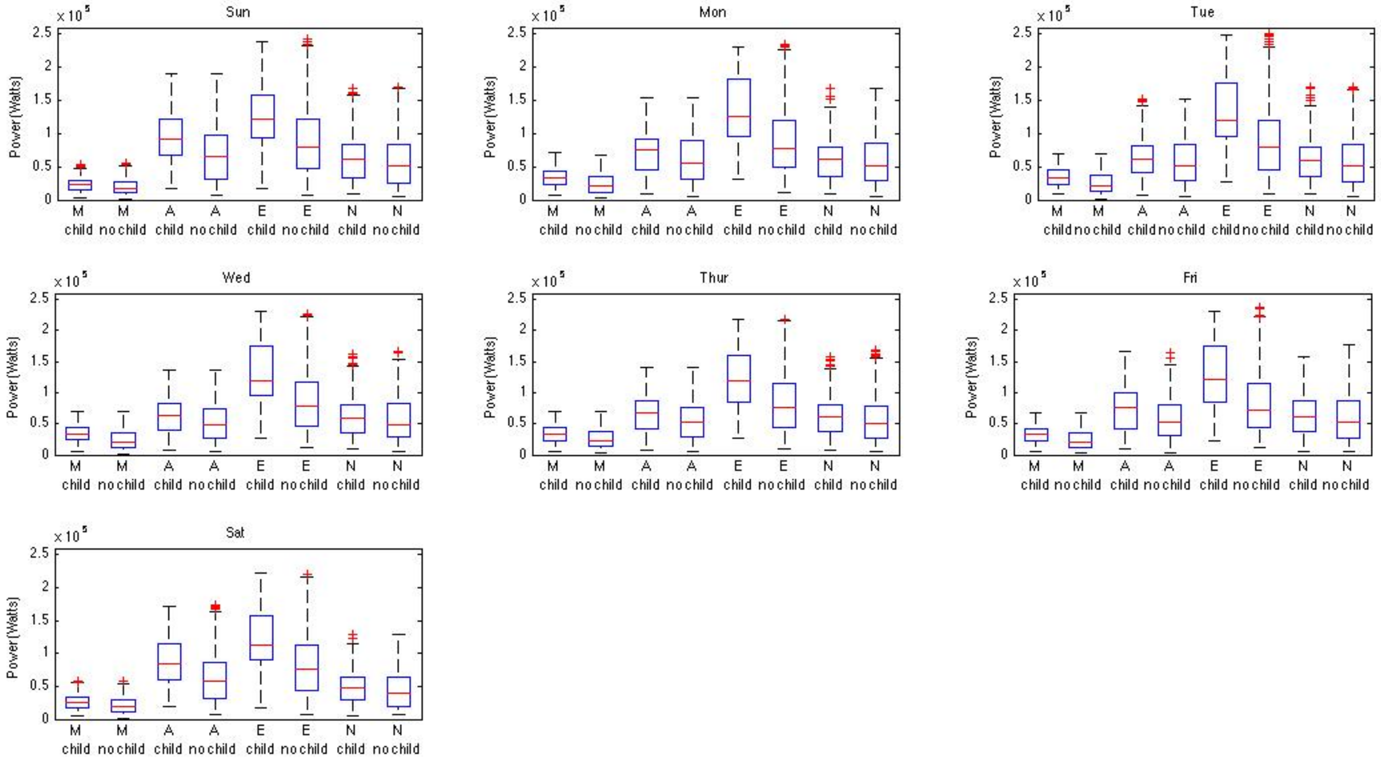


Figure 3.8: •

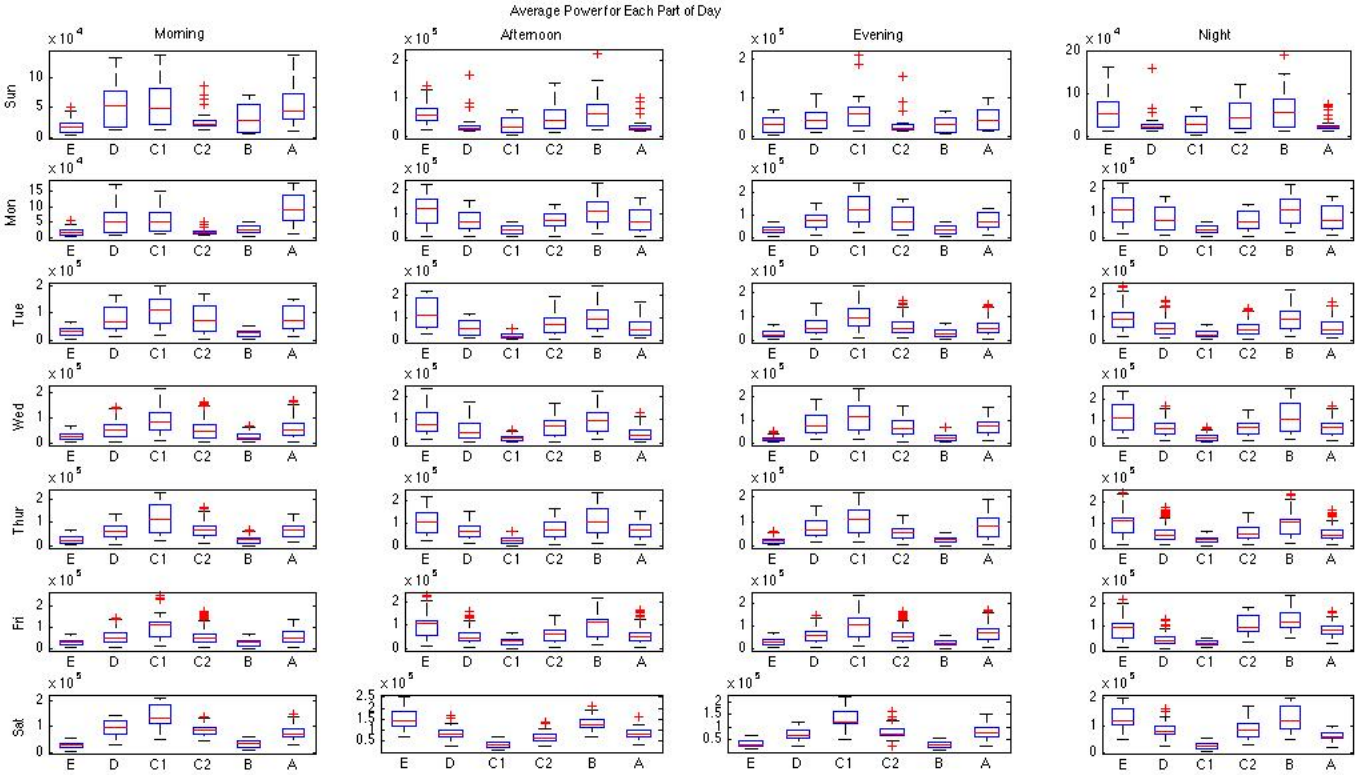


Figure 3.9: •

Correlation Between Weekdays

Examining how different days correlate with one another indicates how

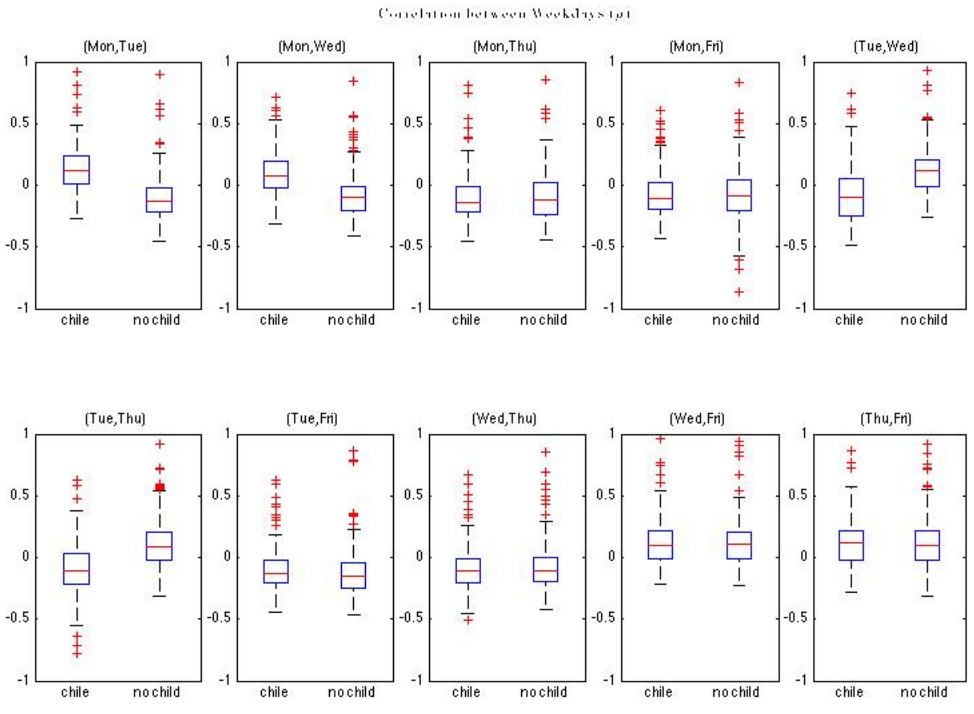


Figure 3.10: •

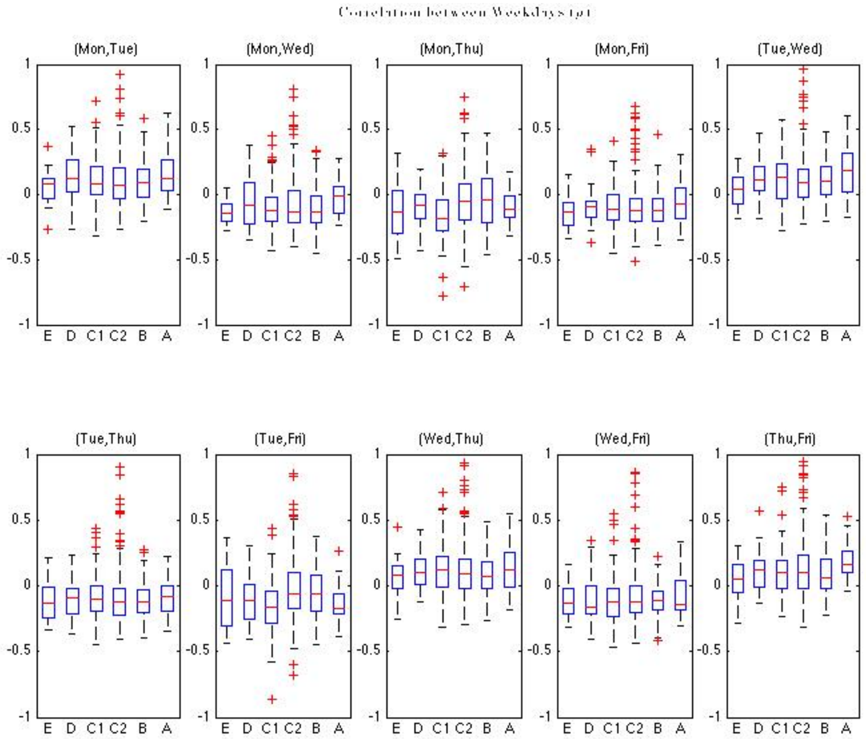


Figure 3.11: •

3.2.2 Periodicity

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

[13] 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