**Data Generation**

**Full Unit Project: Report 3  
Benedict Wilkins**

**Supervisor: Zhiyuan Lou**

**2016**

**Introduction**

Data will be generated by a Data Generation module which will be global – it will be accessible to all House Environments. When generating data we want it to represent a real scenario as much as possible. The best way to do this is to acquire some real data; the energy usage of different households, and fit a representative model to it. We can then sample from the fitted model to generate realistic data. It is necessary to perform some analysis of the data (see below section Data Analysis) to find a good representative model. The analysis of the data gives some insight into desired properties of a prediction model e.g. linearity.

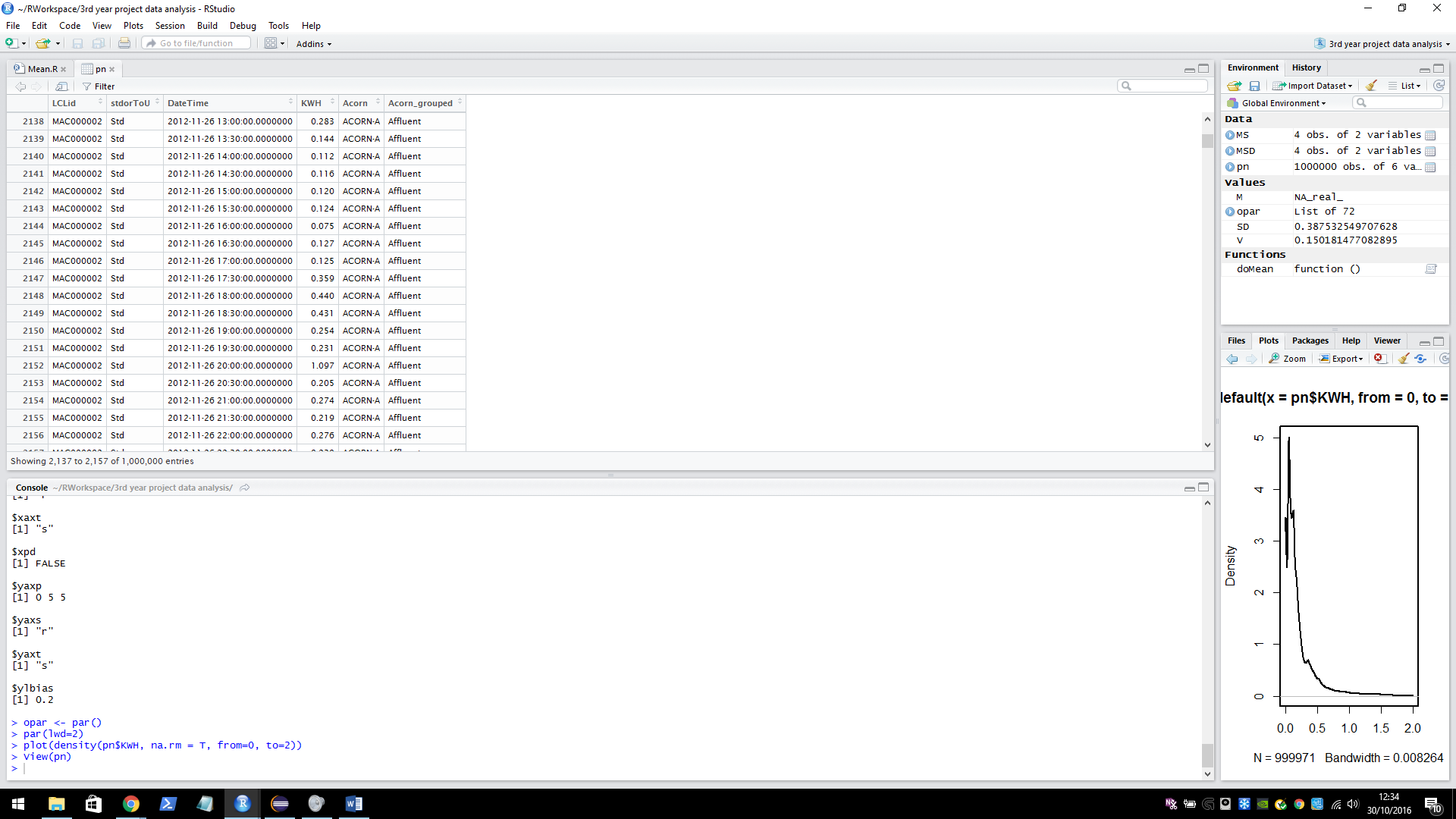
The below section contains analysis of the Low-Carbon-London dataset. This data set almost exactly the kind of data that the simulation will be using – half hourly data in different households for different financial situations.

**Data analysis**

**Low-Carbon-London-Dataset**

The data set includes KWH per half hour readings for a number of households. The data set is grouped into Affluent, Comfortable, Adversity and ACORN-U depending on the customer status.

The first section from this containing 1 million entries was used in the analysis. The complete dataset contains 167 million entries. The sample below shows 8 rows of the data set with the column names.

**LCLid**: the unique house identifier.

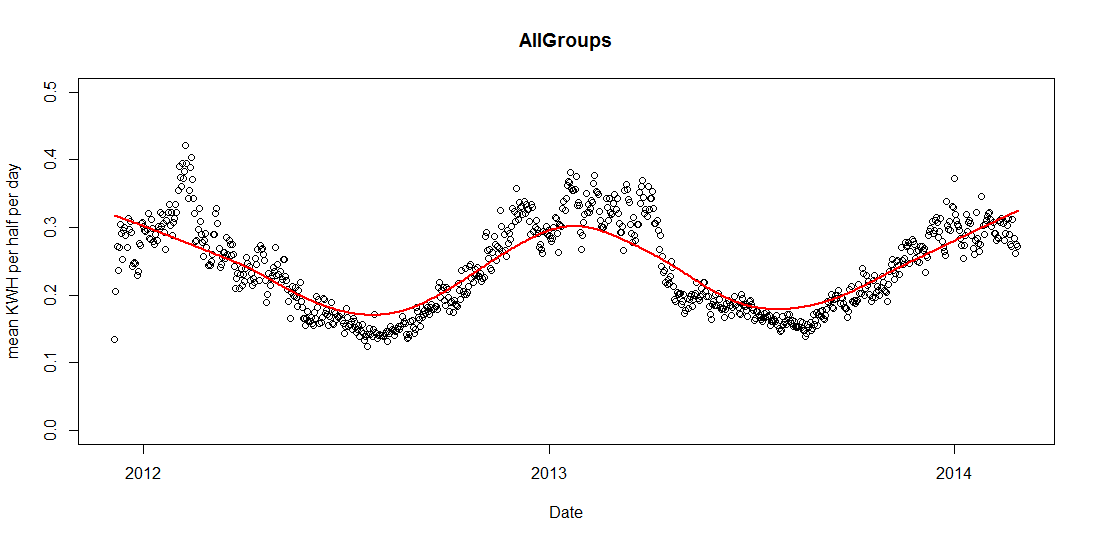
**stdorToU**: Tariff.

**KWH**: energy used KWH per half hour.  
**Acorn/Acorn grouped**: the grouping of the customer.

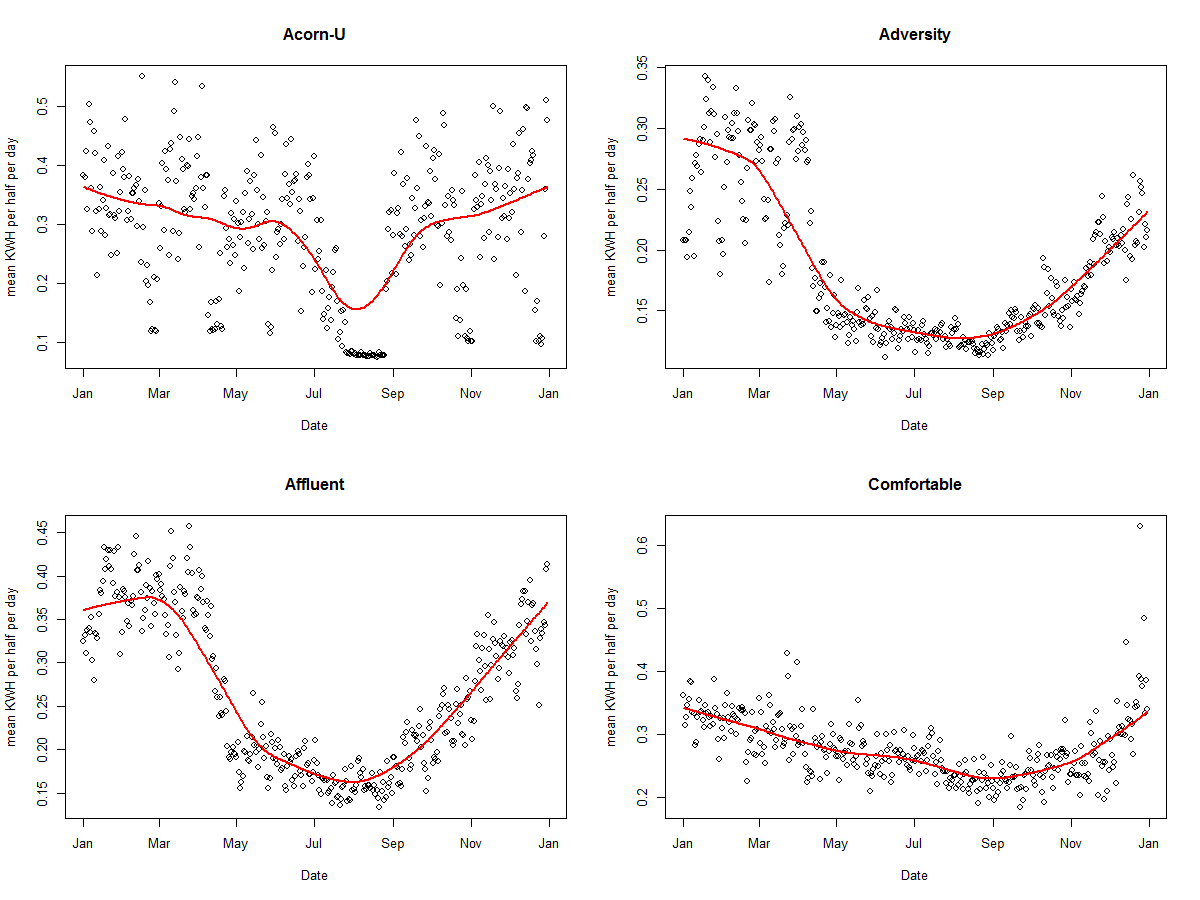
The data set can be found here <https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households>. Details on the ACORN groups can be found here <http://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf>.

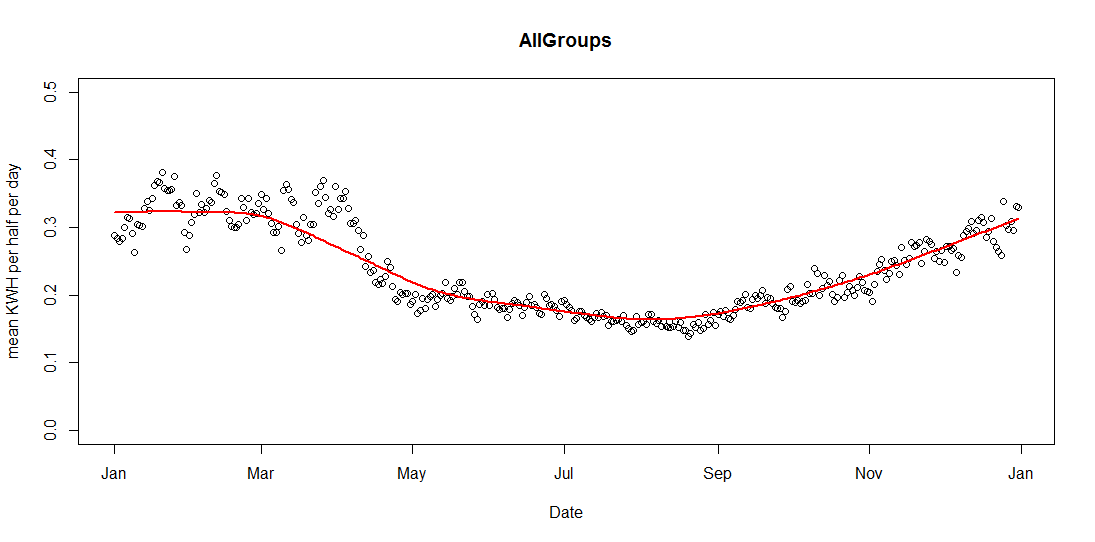
**Seasonality**

The plot below illustrates seasonality over the two years that data was collected. We can see that the average KWH usage is higher in the winter than in the summer. On the y axis: the mean KWH per half hour usage per day, on the x axis: the date (in days).



The seasonality of each group varies, however for the sampling model one seasonality functions will be developed that represents all groups. This will be similar to the above graph (the average of all groups). The graphs below show the seasonality of each group in the 2013 year.





**Building a sampling model**

When building the sampling model the equations for each group will take the form:

b­1\*normal(μ1, σ1) + b­2\*normal(μ2, σ2) + c

Manually fitting two normal curves to the mean KWH per half hour of the ACORN-U group. The blue line shows the fitter combined normal curve, the curve loops around the boundary, this is because hours are continuous – it is time series data. The red line shows a calculated mean values of the data group per half hour. The Y axis shows KWH per half hour, the X axis shows 48 half hourly time intervals which covers one day (24 hours) starting at 00:00:00 and ending at 11:30:00. (The graphs below were generated in the java class DataFitter using the JFreeChart graphing library).

|  |  |
| --- | --- |
|  | **Group: Acorn-U**  b­1 = 3.8  μ1 = 40.0  σ1 = 3.0  b­2 = 6.0  μ2 = 25.0  σ2 = 7.0  c = 0.1 |
|  | **Group: Adversity**  b­1 = 1.0  μ1 = 1.5  σ1 = 1.5  b­2 = 2.0  μ2 = 20.0  σ2 = 14.0  c = 0.12 |
|  | **Group: Affluent**  b­1 = 6.0  μ1 = 48.0  σ1 = 9.0  b­2 = 1.5  μ2 = 24.0  σ2 = 5.0  c = 0.15  discrepancy in normal curve \* |
|  | **Group: Comfortable**  b­1 = 5.0  μ1 = 40.0  σ1 = 5.0  b­2 = 4.0  μ2 = 22.0  σ2 = 6.5  c = 0.1 |

**\*** In the Affluent graph there is a small discrepancy in the normal line. This is because the equation wraps around the boundaries up to half of the range in each direction. It occurs when there is a significant different in the middle values (for each wrapped half range), in most cases the difference is negligible because the normal line tends to 0, it has no effect when added. In this case however the stand deviation of the curves is sufficient enough so that the line is not close to 0 and so the addition is noticeable. It will have a very small effect on the data sampling at the centre point but as the curve is an approximation and some error term will be added anyway it is not something to be too concerned about.

**Proceedings**

After looking at a good representative example of the data that the system will be dealing with (above). It has come to attention that an ANN prediction model may not be the most suitable. Changing it in favour of a simpler and more intuitive model will benefit the final result of the project. Having a simulation of the system including the representative model formed above will allow testing of different prediction models. The selection of models to test does not necessarily exclude an ANN model.

**Implementing and Integrating Data Generation**

Before loading the data into the agent system it was processed using R. (The R scripts can be found in the code repository). The scripts grouped and summarised the data – calculating means and stand deviations of the readings for each time period and for each ACORN group. The data was then formatted and written to file that could be read by the agent system.

**Loading in data**

Data is loaded from a text file by the *DataReader* class. The method *readFile* reads from a text file and returns a *DataFrame* that represents the read data.

**Fitting curves**

**Plotting**

**Integrating**