Using Energy Metering Data to Support Official Statistics: A Feasibility Study

Final Report to the Office for National Statistics

Dr Ben Anderson
Dr Andy Newing

Sustainable Energy Research Group

University of Southampton

Contact:
b.anderson@soton.ac.uk

www.energy.soton.ac.uk

Published: July 2015

Foreword by ONS

The Office for National Statistics (ONS) is the UK's largest independent producer of official statistics and is the recognised national statistical institute for the UK. It is responsible for collecting and publishing statistics related to the economy, population and society at national, regional and local levels. It also conducts the census in England and Wales every ten years.

ONS also plays a leading role in national and international good practice in the production of official statistics. To maintain and further its expertise, the ONS conducts and commissions research covering key topics relating to official statistics and encompassing key emerging conditions.

One emerging change relates to the new data sources becoming available through the growth of technologies such as the Internet. These data sources might have a role in official statistics in a number of ways such as helping to validate or improve official estimates, providing more timely information on trends or reducing costs and response burden through the diminishing need to collect data through normal survey processes.

One new data source of interest to statistical organisations around the world is the high frequency electricity data recorded by domestic smart meters. Such data may help with understanding energy use and expenditure as well as various features such as occupancy status or household size which may be inferred from the profile of energy use over time. All constituent countries of the UK have programs to roll out smart meters to domestic dwellings by 2020, so that information on an almost universal coverage of dwellings may be available from this date.

Energy trials using smart-type meter devices have led to the availability of data on smaller numbers of dwellings for current research and ONS has commissioned the University of Southampton to use some of these trial datasets to test the feasibility of using this data to identify features of households which may have relevance for official statistics. Specifically, this research focuses on the potential of using smart-type meter data to identify household characteristics such as the presence of retired occupants. A second objective is the development of a method to determine occupancy status.

It must be emphasised that the principal interest for ONS is the development of methods to derive estimates for groups of households so as to monitor broad trends whilst ensuring no disclosure of personal information. As a first step towards this aim, it is necessary to conduct research at the individual household level as within this paper.

ONS recognises that smart meter data poses major questions around ethics, privacy and the safeguarding of personal information. ONS has already sought advice from privacy groups on this research and been given approval so as to demonstrate more fully the benefits of using this data. Future use of this data in a production setting will involve extensive engagement with all stakeholders to ensure that the appropriate levels of security are in place to satisfy the strict controls demanded under the code of practice for official statistics (UK Statistics Authority 2009).

The University of Southampton is continuing this research under an ESRC funded project (http://www.energy.soton.ac.uk/category/research/energy-behaviour/census-2022/).

Additionally, ONS is conducting internal research using smart-type meter data through its Big Data project and regular updates are published at http://www.ons.gov.uk/ons/guide-method/development-programmes/the-ons-big-data-project/index.html

Table of Contents

1 Background

The ongoing evolution of the decennial UK Census[1] presents social, policy and commercial researchers with both a challenge and an opportunity. The challenge is to transform 'census-taking' by finding robust alternative methods for creating small area socio-economic indicators over time. The opportunity is to transform the very nature of the socio-economic indicators themselves using new analytic methods applied to new geo-coded datasets and to radically accelerate the temporal cycle from decennial to annual or sub-annual production.

Currently considered approaches include retaining decennial census-taking, more frequent social surveys, commercial or administrative data linkage or aggregation and model-based imputation. In contrast, this project explored the possibility of deriving small area estimates of traditional socio-economic indicators from 'digital trace' or transactional data collected by utility (or other) services as part of normal service provision. The work builds on previous studies highlighting the potential value of the analysis of telecommunications data for the imputation of household characteristics (Anderson, Vernitski, & Hunter, 2012) and the production of socio-economic indicators (Claxton, Reades, & Anderson, 2012).

In contrast, the work reported here explored the feasibility of developing methods for estimating 'census-like' indicators from samples of household electricity power demand data. Compared to a number of other forms of potentially useful ‘big data’, grid-connected electricity is almost universally available in the United Kingdom (unlike mains gas), connection to available supply is similarly almost universal and metering of power demand is mandatory (unlike water). Furthermore the planned universal rollout of electricity smart meters collecting at least half-hourly power demand data (DECC, 2013) means that consideration of the value of suitably anonymised and aggregated smart meter-like data in the production of official statistics is now timely. The use of this kind of data for market segmentation and other electricity related services has been noted in the literature (McKenna, Richardson, & Thomson, 2012) and was noted by Dugmore et al in the context of future census data collection (Dugmore, Furness, Leventhal, & Moy, 2011). However, as far as we are aware only one published study has investigated its potential in the development of official and/or small area statistics (Caroll, Dunne, Hanley, & Murphy, 2013).

Caroll et al’s study made use of six months of data from the benchmark period of the Irish Smart Meter trial in 2009 to 2010 and comprised half-hourly electricity usage for just over 5,000 homes. As they describe, the size of data produced (over 150 million records or 2.5Gb) meant that advanced data management and analysis processes were required to produce a range of explanatory variables and that about 65% of the project person hours was absorbed by data preparation alone (p8). The summary variables included various aspects of overall power demand (mean, maximum, standard deviation, morning maximum, load factor) and were used to try to predict membership of a family type classification using a multinomial neural network model. Overall they report some success with binomial classifications but note that this would have less value than the ability to distinguish between multiple types. Significantly, despite reporting graphs of time of day profiles of power demand for different family types, their work only used overall power demand parameters. Whilst overall power demand *may* be indicative of household composition, it seems more likely that different kinds of household will have different temporal habits. If this is the case then it may be feasible to use these different temporal power demand profiles to distinguish between different household types as has been previously demonstrated using time-use diary data (Lesnard, 2004). Indeed this approach is already used to underpin models of energy power demand (Richardson, Thomson, Infield, & Clifford, 2010).

[1] http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/beyond-2011-report-on-autumn-2013-consultation--and-recommendations/index.html

In the work reported below, we therefore concentrate on temporal patterns of power demand that may be able to more robustly distinguish between household types due to the potentially different timings and intensities of their everyday habits and routines.

2 Objectives

The ultimate aim for this research is to develop methods which could, in the future, be applied to smart meter-derived power demand data to provide either sample based or whole population based (all customers) small area estimates household characteristics.

As a necessary first step, the project explored the feasibility of predicting household level characteristics using two transactional data sources from trials of energy usage where households had expressly given their 'opt-in' permission. More specifically, following a preliminary review of available data and discussion with the Office for National Statistics, the project aimed to:

1. Assess the feasibility of predicting at the household level:
    1. the *number* of occupants;
    2. the presence of *children*;
    3. the presence of *single persons or couples aged 65+.*
2. Assess the feasibility of predicting whether occupants at a given address will be 'at home and awake' (*active occupancy*) at given times of the day and days of the week to support census (and other survey) fieldwork processes.

In both cases the work started from the hypothesis that different kinds of household occupants will have different temporal habits and whilst it might be assumed that high levels of power demand of electricity might coincide with active occupancy there are a number of potential confounding factors:

- Electric heating through storage heaters may produce high power demand values over several sequential half hour periods even if an occupant is not present. The likelihood that such power demand will be at night (Zimmerman et al., 2012) may aid interpretation in this instance but installations of electrically powered heat pumps are likely to consume power in currently unknown patterns depending on the way they are used;
- Hot water heating through an immersion heater on a timer switch would generate high power demand assuming a cold start although the sharp increase at predictable times may aid interpretation;
- Photovoltaic cells, although still relatively rare, are likely to produce maximum output near mid-day. Whilst future smart meters will be able to distinguish exported from imported power, most contemporary instrumentation cannot and thus would record high levels of production as (relatively) high levels of power demand during the middle of the day;
- Appliance use (i.e. switching), which could be assumed to be a more reliable indicator of active occupancy, may be difficult to detect in power demand data that has been aggregated to half hours (Armel, Gupta, Shrimali, & Albert, 2013). For example whilst a kettle may use 2-3Kw when in use, which is not dissimilar to an immersion heater or an electric fire, it does so for a minute at most. Thus this power demand is 'spread out' across the 30 minutes of the aggregation period.

A number of commercial companies claim to be able to disaggregate different appliances from sub one minute level power demand data[2] but it is accepted that data aggregated to more than 15 minute intervals can only distinguish between base and variable load and some large and sustained peaks (Armel et al., 2013). Unless data is available at the one minute level it may therefore not be possible to identify the use of appliances which can unambiguously indicate active occupancy. As this level of smart meter data is unlikely to be available at the population or large sample level in the future, proxy indicators based on half-hour level power demand were developed and tested in the work that follows.

[2] Katie Russell, Onzo Ltd, personal communication and see also (Hamouz, 2012)

3 Data

The data to be used, both having consumer 'opt-in' status, were:

- A dataset collected by the University of Loughborough and archived by the UK Data Service for future research use which links aggregated one minute power demand readings to a basic household occupancy and appliance ownership survey (Richardson & Thomson, 2010). The data derives from 22 dwellings observed over two years (2008-2009) and due to it's small size proved to be of value only for exploratory and experimental analysis.
- A similar energy power demand monitoring dataset held by the University of Southampton which derives from around 180 households from two case study areas (wards) of the Solent region. Instantaneous power demand data is collected every second and can be linked to repeated six-monthly survey data on household occupancy and other variables. The data does not provide complete coverage of all households in either of the case study areas and so cannot be used to produce 'whole population' small area estimates.

In both cases 30 minute summaries of this data were used to replicate the level of granularity that will initially be available from the proposed national electricity smart meter roll-out.

As will be discussed below, the second dataset comprises extremely large data files and has proved more time (and processor) consuming to work with than was originally anticipated. In common with other studies using this form of data (Caroll et al., 2013), it is important to consider the impact of missing data (for example where broadband internet connection with the recording device has been temporarily lost or where the monitoring instruments failed) on the accuracy of aggregated results. Whilst aggregating results to generate 30 minute summaries was straightforward, considerable time and computing intensive pre-processing is required in order to understand the pattern of missing data and exclude those households from subsequent analysis during any time period where sufficient data is not available. This step is likely to be an issue in any future studies using data collected remotely from domestic smart meters especially where the chosen data upload channel may be unreliable.

The remainder of the report describes analysis of both datasets, starting with the smaller 22 household 1 minute Loughborough data and then moving on to the larger 180 household Southampton sample. In each case the data processing and cleaning required, initial descriptive analysis and, where the data enables it, the results of modelling the relationship between half-hourly electricity power demand and household composition and active occupancy are described.

4 One-Minute Resolution Domestic Electricity Use Data, 2008-2009

4.1 Data Background

The dataset[3] comprises 22 households with data collected from 1/1/2008 to the end of December 2009 although not all households were monitored for the whole period of the study. Some complete days (and weeks) are missing as are a very few of the per-minute readings on days that were otherwise complete.

This is confirmed by the user guide which notes the following information for the variables in the data:

1. DATETIME_GMT - The time stamp of the meter reading as Greenwich Mean Time (GMT).
2. DATETIME_LOCAL - The time stamp of the reading taking British Summer Time (BST) into account.
3. IMPORT_KW - The mean instantaneous power demand (in kW) during the one minute period starting at the time stamp.
4. The date time fields are formatted as <YEAR>/<MONTH>/<DAY> <HOUR>:<MINUTE>.

It also notes that:

5. Where data is not available for a given minute, no row exists in the file.
6. No data is available for two of the meters in 2009, and hence two of the files are empty.

The electricity data itself comprises 1 minute resolution average kw imported from the grid producing 17,772,709 records.

The survey contains two variables of direct interest to this feasibility study:

1. number of people (only have for 17)
   a) 1 = 2
   b) 2 = 2
   c) 3 = 3
   d) 4 = 5
   e) 5+ = 5
2. household accommodation (have for 22)
   a) Detached = 11
   b) Semi = 7
   c) Terraced = 4

The survey also contains a range of indirectly interesting questions to do with which electricity using appliances people own. As noted above this data is of little value when demand is aggregated to half-hours as it is not possible to discern the useage of specific appliances and so has not been used in the work reported here. In addition:

- No household had electric heating nor did any generate their own electricity through solar panels or wind turbines and none used electrically powered heat pumps.
- 12 had an Economy 7 tariff
- 10 reported the use timers to run appliances at night

In the analysis that follows neither tariff schemes nor the reported use of timers or any other appliance was taken into account. However it should be noted that following Census 2011, the presence of different forms of heating may be a key future census variable and it seems likely that given the projected growth in electrically powered heat pumps in particular, assessment of their uptake via electricity use profiles may be both desirable and feasible.

[3] Available from http://discover.ukdataservice.ac.uk/catalogue/?sn=6583

4.2 Data Processing

The 17,772,709 monitoring records were aggregated into a single comma separated file, loaded into STATA 12 and converted to a STATA format file. This file, comprising 22 households measured at 1 minute intervals for two years was 360Mb in size. STATA's 'tsfill' command was then used to 'fill in' any missing 1 minute observations for the 22 households in one of two ways:

1. To fill in all observations missing between the first and last observation for a given household producing a dataset of 20,749,711 records (c 500 Mb). Power demand was not imputed so the record has no value other than a timestamp.
2. To fill in all observations between the first and last observations producing a dataset of 23,158,080 records (c 550 Mb). Again power demand was not imputed so the record has no value other than a timestamp.

Table 1 shows the mean observed instantaneous power demand and the percentage of observations that were zero as well as the missing characteristics under each method. As can be seen the total number of observations for each household over the period varies substantially. Of these, few report zero demand with the exception of household 3 which reports just over 4% zero values. These zero values are likely to correspond to very low demand levels that fall below the instrumentation 'detectable' threshold.

Table 1: Data characteristics (1 minute level data)

	Observed					Fill Method 1		Fill method 2	
Household	Mean power import (kW)	Min	Max	N observed	% zero import	N (filled)	% missing	N (filled)	% missing
1	0.43	0.00	8.27	894,850	0.00%	1,052,640	14.99%	1,052,640	14.99%
2	0.60	0.00	9.14	261,781	0.03%	272,160	3.81%	1,052,640	75.13%
3	0.25	0.00	14.25	588,528	4.40%	774,720	24.03%	1,052,640	44.09%
4	0.62	0.01	14.69	887,584	0.00%	1,052,640	15.68%	1,052,640	15.68%
5	0.14	0.00	5.78	931,627	0.00%	1,052,640	11.50%	1,052,640	11.50%
6	0.25	0.00	7.86	889,854	0.00%	1,052,640	15.46%	1,052,640	15.46%
7	0.45	0.00	9.11	917,229	0.31%	1,052,640	12.86%	1,052,640	12.86%
8	0.52	0.00	10.87	558,671	0.02%	881,280	36.61%	1,052,640	46.93%
9	0.33	0.01	6.43	568,770	0.00%	675,360	15.78%	1,052,640	45.97%
10	0.38	0.00	10.15	734,029	0.00%	1,052,640	30.27%	1,052,640	30.27%
11	0.38	0.00	6.86	800,599	0.57%	1,052,640	23.94%	1,052,640	23.94%
12	0.25	0.00	14.41	966,202	0.00%	1,052,640	8.21%	1,052,640	8.21%
13	0.51	0.00	15.45	990,627	0.00%	1,052,640	5.89%	1,052,640	5.89%
14	0.55	0.00	19.68	993,598	0.00%	1,052,640	5.61%	1,052,640	5.61%
15	0.37	0.00	8.26	948,902	0.00%	1,052,640	9.86%	1,052,640	9.86%
16	1.19	0.00	11.59	264,271	0.05%	264,271	0.00%	1,052,640	74.89%
17	0.47	0.00	13.01	940,173	0.00%	1,052,640	10.68%	1,052,640	10.68%
18	0.33	0.00	12.06	939,125	0.57%	1,048,320	10.42%	1,052,640	10.78%
19	0.62	0.00	9.88	958,263	0.02%	1,052,640	8.97%	1,052,640	8.97%
20	0.57	0.00	11.40	930,193	0.00%	1,044,000	10.90%	1,052,640	11.63%
21	0.73	0.01	14.58	933,834	0.00%	1,052,640	11.29%	1,052,640	11.29%
22	0.47	0.00	8.30	873,999	0.00%	1,052,640	16.97%	1,052,640	16.97%
Total	**0.45**	**0.00**	**19.68**	**17,772,709**	**0.00%**	**20,749,711**	**14.35%**	**23,158,080**	**23.25%**

Fill method 1, which 'fills in' missing observations for all households between their first and last observation, shows that on average 14.35% of observations were missing between the first and last observation with some households (8 & 10) having more than 30% missing. Method 2, which fills in all

missing records between the first and last recorded reading in the dataset, shows that records for 2 households (2 & 16) were substantially absent for most of the two years with four others (3, 8, 9 and 10) showing a high (> 30%) level of missing observations over the study period.

This is visually confirmed by Figure 1 which shows that there were some households for whom there were no data after July 2008 (households 2 & 16) whilst households 3, 8 and 9 lasted only until early 2009. There were very few partially missing days as indicated by the % missing value either being 0 or 100% in all but a very few records.

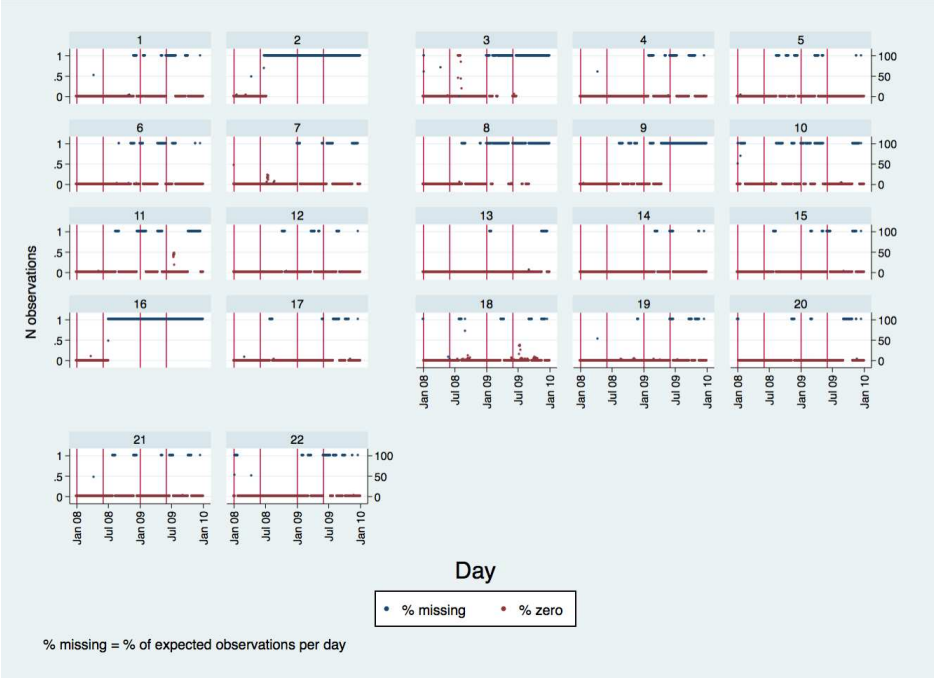

Figure 1: Distribution of missing and zero power values by day over the full time period 2008-2009

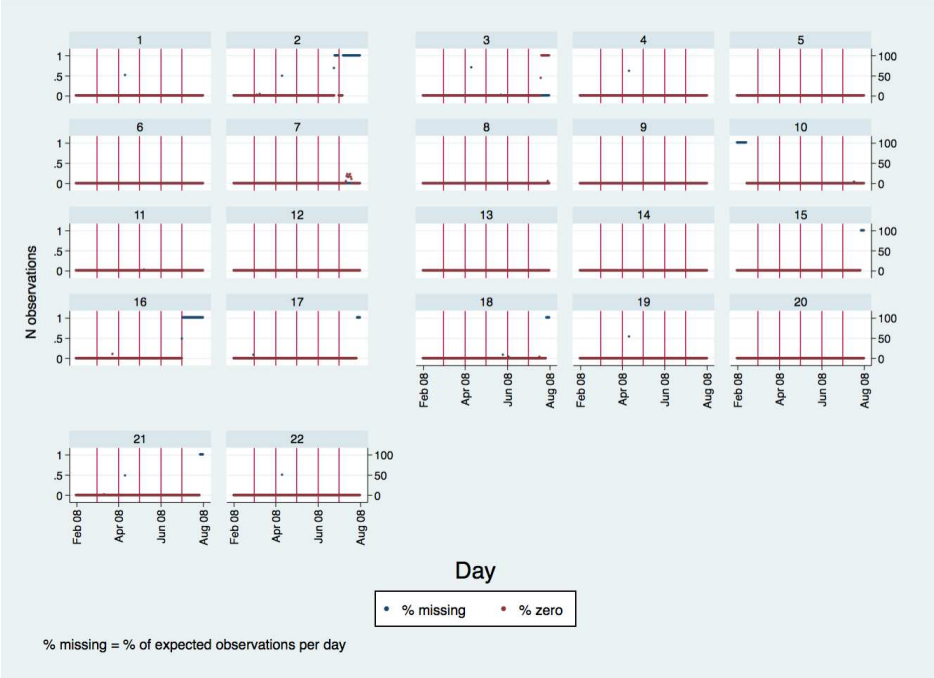

Figure 2: Distribution of missing and zero power values by day in Spring/Summer 2008

In order to provide a relatively clean dataset for analysis, all observations between February and August 2008 were selected which, as Figure 3 shows, produced a dataset with very low rates of missing and zero observations.

Further analysis (c.f. Figure 3) suggested that selecting March to June 2008 would provide a dataset with virtually no missing observations and no zero recorded demand.

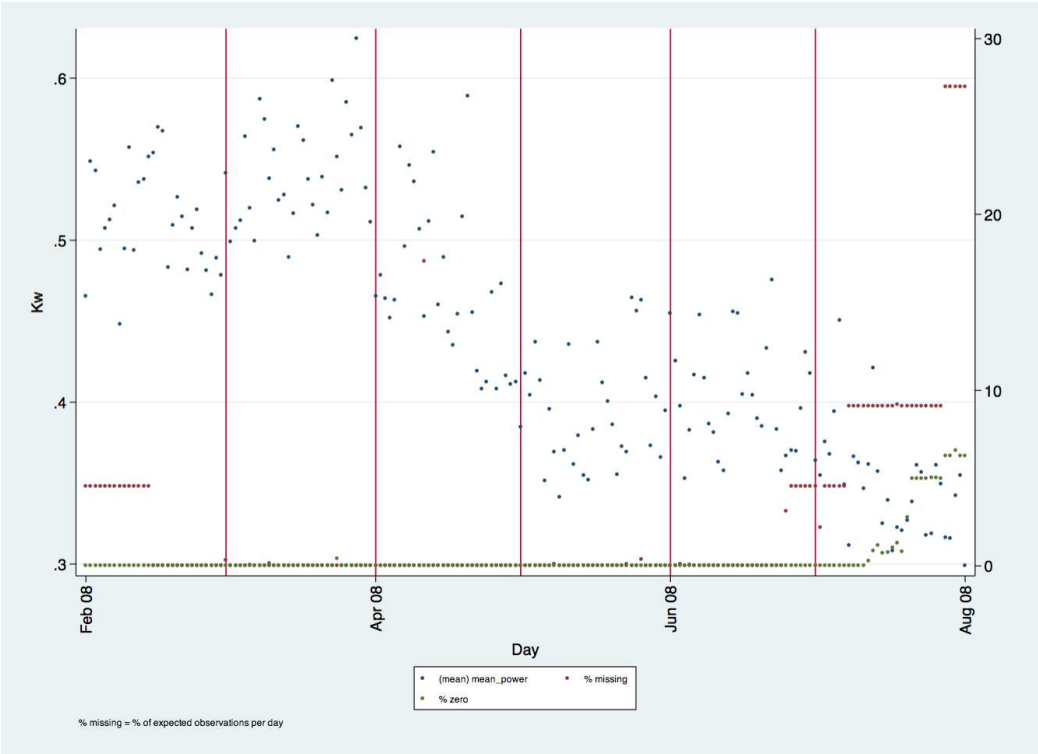

Figure 3: Mean imported power (1 minute averages), % non-missing and % zeros per day from February – August 2008

In order to avoid potential seasonal effects and to align with the time of year that the Census is traditionally conducted (late March) Saturday 1st – Friday 28th March 2008 was then selected as the final analytic datatset. This balanced sample provided 4 weeks of data with equal numbers of days of the week and also avoided the complication of including Sunday 30th March which was the start of British Summer Time when clocks were put back one hour.

The one minute data for this four week period was then aggregated to half-hourly intervals producing a file of 29,568 records of summary statistics for each half hour per day per household (see Table 2).

Table 2: Aggregation to half-hours: summary statistics

Statistic	Specification: For each half hour for each day for each household:
Mean power demand	Mean of power imported per minute
Standard deviation of power demand	s.d of power imported per minute
Median power demand	median power imported per minute
Percentiles of power demand	5%, 10%, 90% and 95% percentiles of power imported per minute

Only ten out of the 29,568 half hour intervals contained less than the expected 30 observations with five containing 28, one containing 24, one containing 17 and three containing no observations. All of the last four (1 * 17 & 3 * zeros) were for one household (17) between 07:30 and 09:00 on the 1st March and have been excluded from the following analysis.

4.3 Electricity demand: Descriptive analysis

Mean instantaneous power (electricity) demand per minute at the half hour level was 0.541 kW across all 22 households. As Caroll et al (2013) found, the distribution of this value is highly positively skewed with a median of 0.284 kW and a skewness of 2.58 (see also Figure 5).

Figure 4 shows the mean power demand for all households by time of day and day of the week and reveals the expected pattern of morning and evening peaks with particularly noticeable weekday troughs in demand in between more evenly distributed demand at weekends. The small increase in variability around March 23rd corresponds with the Easter weekend.

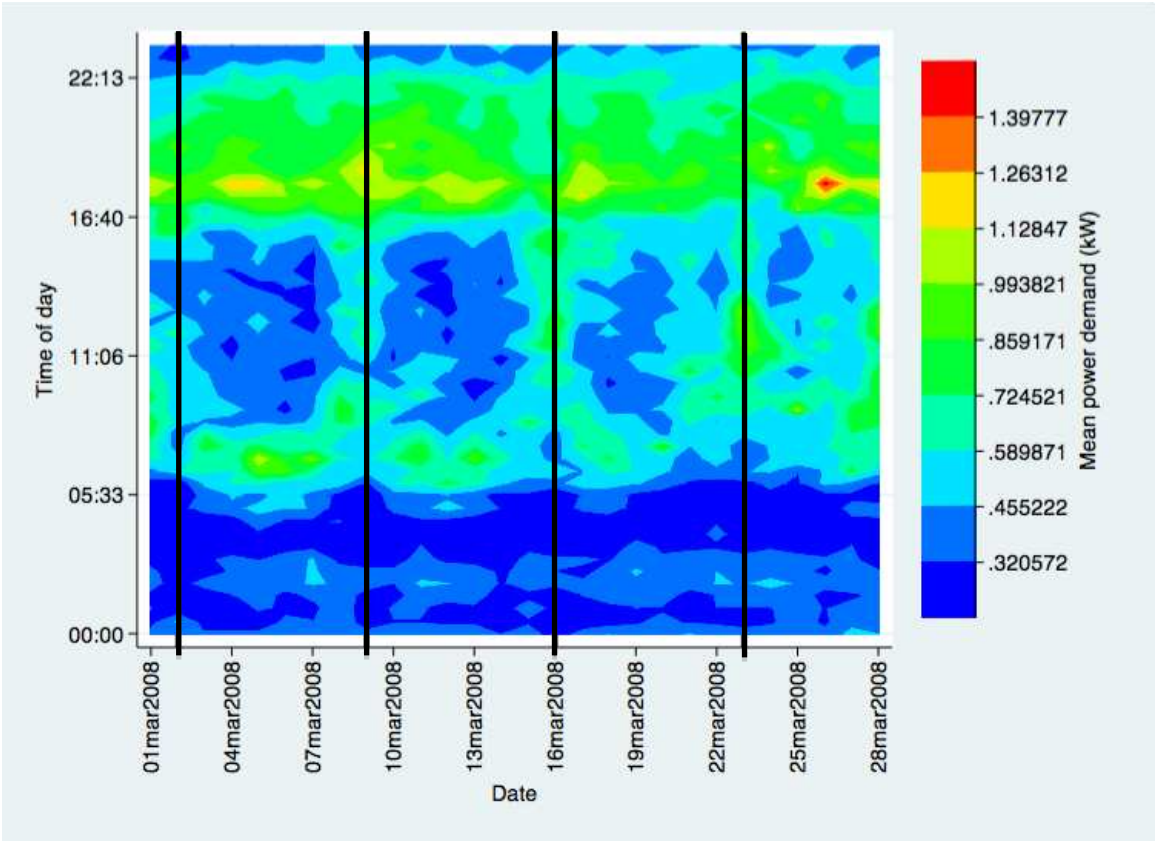

Figure 4: Mean instantaneous power (electricity) demand by time and day for March 2008, vertical lines are Sundays

Figure 5 to Figure 8 show the distribution of mean half hourly power demand across accommodation type, household size and time of day for all days combined. Figure 5 shows the positively skewed nature of the distribution with many periods of low power demand and a characteristic long tail of few periods with much higher power demand.

In the case of accommodation type and number of occupants the error bars indicate the between-household standard deviation of the means within each group. As might be expected there seems to be little relationship between power demand and accommodation type. However mean power demand appears to increase with household size although it should be remembered that there were only two homes with one person and two with two people. In the case of accommodation type, there is more variation within the detached homes group than the semi-detached and terraced. In the case of occupancy, most within-group variation occurred in the largest households.

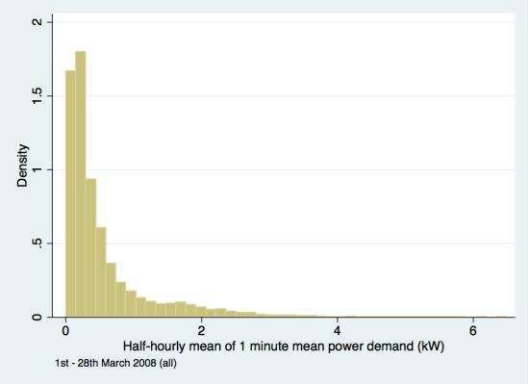

Figure 5: Histogram of mean power demand at half hour level

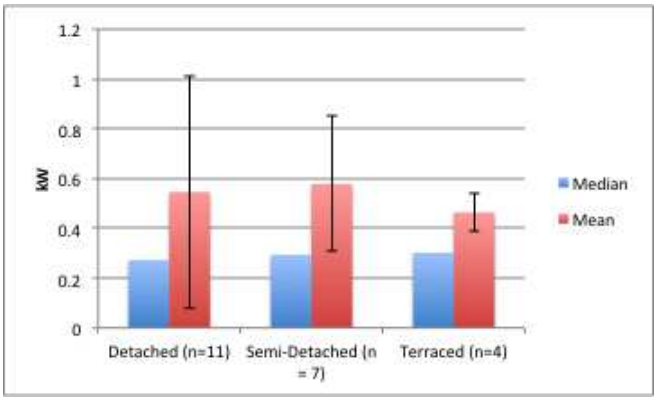

Figure 6: Mean and median power demand by accommodation type, error bars indicate the between-household standard deviation of the means for each group

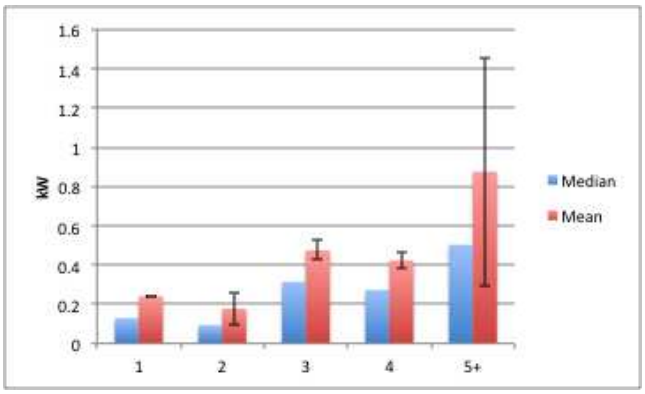

Figure 7: Mean and median power demand by number of occupants, error bars indicate the between-household standard deviation of the means for each group (NB low numbers: only 2 one person households and 2 two person households)

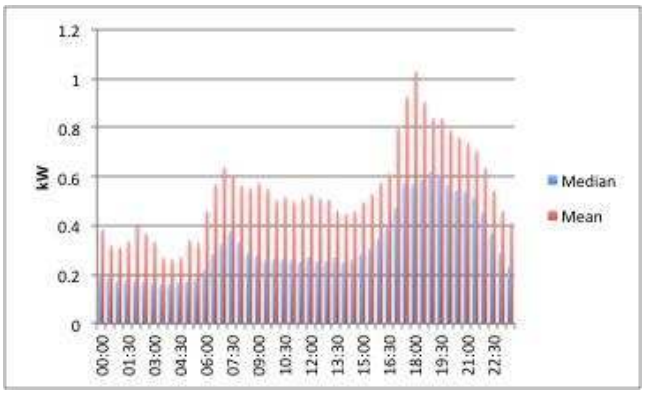

Figure 8: Mean and median power demand by time of day

Figure 8 shows how average (mean and median) power demand is distributed across the day and given the initial working hypothesis that such profiles may be more informative than overall power demand, Figure 9 and Figure 10 show the time of day profile of mean power demand by different household types and by weekday and weekend. As might be expected there are few clear differences between the accommodation types and it is also clear that the evening peak in demand is less evident at the weekend, suggesting that it is partly driven by the temporal constraints of school and/or work during weekdays.

Similarly Figure 10 also shows a more varied temporal distribution at the weekends and a clearer twin peaked temporal distribution for weekdays corresponding to the pre-work/school morning (06:00 – 09:00) and after work/school evening (16:00 – 20:00) periods. Perhaps most interestingly there are also more noticeable differences between the household types when occupancy levels are considered. Thus, whilst there are specific temporal exceptions perhaps driven by the very small sample size, larger households tend to consume more electricity at all times of day and this is particularly the case in the evening. However this pattern is not always consistent with the single person households in the sample consuming more on average than the two person households at peak times. It should also be noted that analysis of the median (not shown) as opposed to mean (shown here) values provides slightly clearer distinctions.

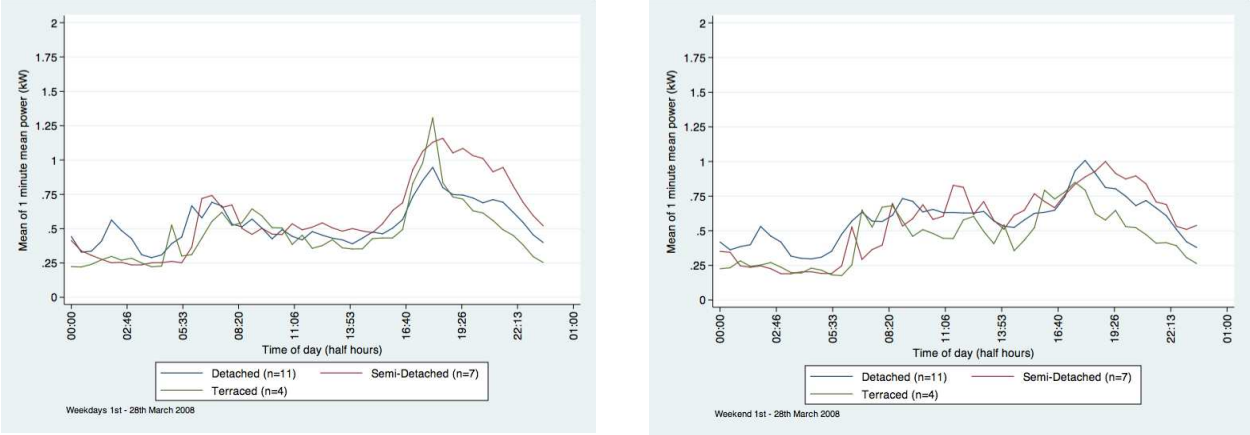

Figure 9: Mean power demand by accommodation type for weekdays (left) and weekends (right)

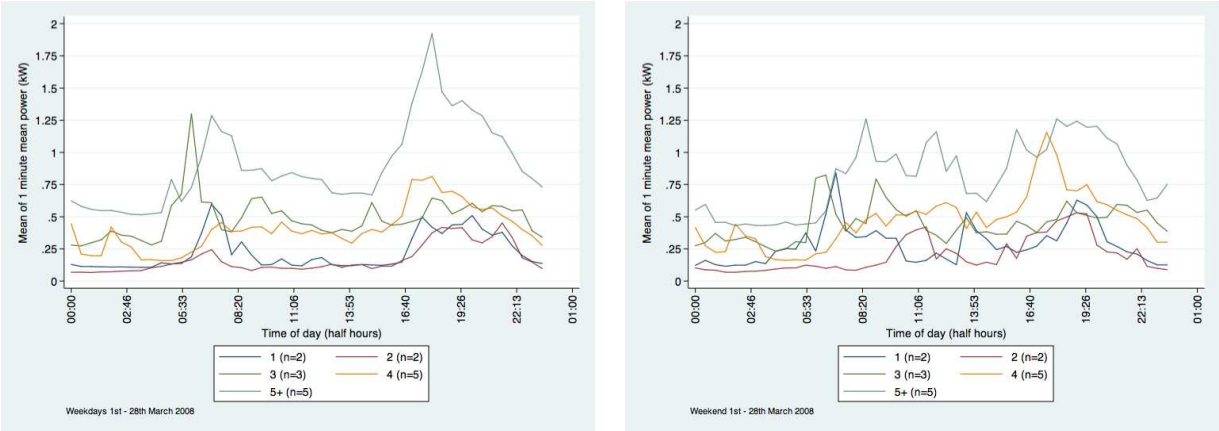

Figure 10: Mean power demand by number of occupants for weekdays (left) and weekends (right)

4.4 Predicting power demand from household characteristics

Given these encouraging descriptive results the first step in the analysis was to test the extent to which accommodation type and household size predicted mean power demand before attempting to estimate the model in reverse and thus predict accommodation type or household size.

We therefore estimated a set of mixed effects multilevel regression models using the full three week half-hourly dataset but selecting only weekdays (Monday – Friday) between 07:00 – 23:00 to allow for the greatest differentiation. Each household was therefore observed 32 times on each weekday of the four-week period producing 640 observations per household.

To account for the nested structure of the data a three level (observation, half hour, household) model was fitted using STATA 12's xtmixed command[4]. This model has the general form

$$Y_{jk} = X_{jk}\beta + Z^{(3)}_{jk}\mu_k^{(3)} + Z^{(2)}_{jk}\mu_{jk}^{(2)} + \varepsilon_{jk}$$

for i = 1,…,n_{jk} first level observations nested within j = 1,…,M_k second level groups (households) which are nested within k = 1,…,M third level groups (half hours). Thus $Z^{(3)}_{jk}$ represents the third level random effects $\mu_k^{(3)}$ (households) and $Z^{(2)}_{jk}$ represents the second level random effects $\mu_{jk}^{(2)}$ (half hours).

[4] www.stata.com/bookstore/stata12/pdf/xt_xtmixed.pdf

Mean power demand values were log transformed to mitigate the skewed nature of the distribution (see Annex 1.1) and the general modeling approach described above was then used to fit the following fixed effects components:

1. Prediction of log mean power demand (over the half hour) by accommodation type (as dummy variables) and number of persons (as dummy variables)
2. Prediction of log mean power demand (over the half hour) by accommodation type (as dummy variables), number of persons (as dummy variables) and time of day (half hours as dummy variables)

The performance of these two models was compared using an LR test. In order to account for the potential non-independence between adjacent half hours Model 1 was then repeated separately for each half-hour period for the times of day when the largest differences were observed in the descriptive analysis above (07:00 – 09:00 and 16:00 – 20:00 on weekdays). In this model each household was therefore observed once per weekday of the four-week period giving a total of 20 observations per household.

The results in Table 3 reports the results of these two models whilst also Table 4 reports model diagnostics. Diagnostic plots for residuals are shown in Annex 1.3 and appear satisfactory. Statistical tests for normality were not used due to the large number of observations.

Table 3: Results of model estimating effects of household characteristics on log mean power demand for all half hours (weekdays, 07:00 – 23:00)

		Model 1				Model 2 (time included)			
Variable	Category	b	95% (lower)	95% (upper)	sig	b	95% (lower)	95% (upper)	sig
Fixed effects parameters									
Accommodation type	Detached (contrast)								
	Semi-Detached	-0.067	-0.184	0.051		-0.066	-0.126	-0.005	*
	Terraced	-0.129	-0.245	-0.014	*	-0.129	-0.188	-0.069	***
Number of persons	1 (contrast)								
	2	-0.468	-0.636	-0.301	***	-0.468	-0.631	-0.306	***
	3	0.753	0.589	0.918	***	0.753	0.594	0.913	***
	4	0.627	0.476	0.779	***	0.627	0.481	0.774	***
	5+	1.300	1.157	1.443	***	1.300	1.161	1.439	***
Time of day	07:00 (contrast)								
	7:30					-0.058	-0.373	0.258	
	8:00					-0.238	-0.553	0.078	
	8:30					-0.371	-0.686	-0.055	*
	9:00					-0.392	-0.708	-0.077	*
	9:30					-0.340	-0.656	-0.025	*
	10:00					-0.433	-0.748	-0.117	**
	10:30					-0.380	-0.695	-0.064	*
	11:00					-0.403	-0.719	-0.088	*
	11:30					-0.450	-0.766	-0.135	**
	12:00					-0.391	-0.707	-0.076	*
	12:30					-0.440	-0.755	-0.124	**
	13:00					-0.475	-0.791	-0.160	**
	13:30					-0.486	-0.802	-0.171	**
	14:00					-0.500	-0.816	-0.185	**
	14:30					-0.446	-0.762	-0.130	**
	15:00					-0.410	-0.726	-0.095	*
	15:30					-0.344	-0.659	-0.028	*
	16:00					-0.229	-0.545	0.086	
	16:30					-0.118	-0.433	0.198	
	17:00					0.161	-0.155	0.476	
	17:30					0.312	-0.004	0.627	
	18:00					0.408	0.092	0.723	*
	18:30					0.367	0.052	0.683	*
	19:00					0.354	0.039	0.670	*
	19:30					0.336	0.020	0.651	*
	20:00					0.312	-0.003	0.628	
	20:30					0.257	-0.058	0.573	
	21:00					0.216	-0.100	0.531	
	21:30					0.180	-0.136	0.495	
	22:00					0.032	-0.283	0.348	
	22:30					-0.149	-0.464	0.167	
Constant		-1.687	-1.854	-1.521	***	-1.559	-1.813	-1.304	
Random Effects parameters									
Time of day	sd(constant)	0.292	0.220	0.388					
Household	sd(constant)	0.455	0.425	0.488		0.440	0.411	0.470	
Residuals	sd(Residual)	0.735	0.725	0.745		0.735	0.725	0.745	
LR test chi sq		3465.25			***	2179.32			***
N		10875				10875			

Note: *: P < 0.05, ** p < 0.01, *** p< 0.005

The results in Table 3 extend the descriptive analysis suggesting that terraced homes have a lower mean instantaneous power demand when the number of occupants is controlled.

Table 4: Model 1 and Model 2 diagnostics

Model	Obs	ll (model)	df	AIC	BIC
Model 1	10875	-12703.17	10	25426.34	25499.28
Model 2	10875	-12655.09	41	25392.18	25691.24

In general the greater the number of occupants, the greater the mean power demand and there is also some evidence that the single person households (contrast category) in this sample demand more power than the 2 person households (negative coefficient) as the descriptive analysis above suggests. Model 2 produces almost identical results (see also Table 4) but confirms the expected time of day effect with the evening periods predicting higher than average power demand. Both models are significantly different from a one level linear regression model (see likelihood ratio test chi sq in Table 3) and a likelihood ratio test comparing them suggests that the models are significantly different (chi sq = 96.16, $p < 0.005$) indicating that including the time of day dummies produced a better fit.

Table 5 shows the results of re-estimating Model 1 separately for each half hour time period on weekday mornings, whilst Table 6 shows results for the same model for half hour periods in the afternoon and evenings of weekdays.

Table 5: Model 1 results for half hour periods in the morning (weekdays) –only fixed effects results shown

		07:00		07:30		08:00		08:30		09:00	
Accommodation type	Detached (contrast)	b	sig	b	sig	b	sig	b	sig	b	sig
	Semi-Detached	-0.421		-0.348		0.030		-0.084		-0.101	
	Terraced	-0.158		-0.038		-0.133		-0.024		0.187	
Number of persons	1 (contrast)										
	2	-1.338	***	-1.119	**	-0.605		-0.685		-0.579	
	3	-0.059		-0.015		0.717		0.717		1.032	*
	4	-0.351		0.136		0.619		0.534		0.630	
	5+	0.815	*	0.983	**	1.542	***	1.277	**	1.397	**
Constant		-0.774	*	-1.100	***	-1.870	**	-0.084	**	-2.062	***
LR test chi sq		28.758	***	29.708	***	58.197	***	54.236	***	61.865	***
N		339		340		340		340		340	

Note: *: $P < 0.05$, ** $p < 0.01$, *** $p < 0.005$

The models for the morning suggest that there is little to choose between the time periods in terms of greater differentiation between the household types although the coefficients for 5+ persons are larger in the 08:00 to 09:00 periods whilst there is a significant effect for 2 person households in the 07:00 and 07:30 periods.

In contrast the evening periods show greater differentiation with the time periods 16:00, 16:30 in particular appearing to offer the potential to differentiate between 3, 4 and 5+ occupancy households, possibly as a result of the presence of children.

Table 6: Model 1 results for half hour periods in the evening (weekdays) –only fixed effects results shown

	16:00		16:30		17:00		17:30		18:00		18:30		19:00		19:30		20:00	
Detached (contrast)	b	sig	b	sig	b	sig	b	sig	B	sig	b	sig	b	sig	b	sig	b	sig
Semi-Detached	-0.013		0.033		0.162		0.110		-0.081		-0.058		0.119		0.113		0.008	
Terraced	-0.311		-0.262		0.108		0.163		0.219		0.026		-0.022		-0.046		-0.204	
1 (contrast)																		
2	-0.272		-0.206		-0.765		-0.833	*	-0.546		-0.363		-0.345		-0.405		-0.535	
3	1.124	**	1.051	*	0.600		0.229		0.232		0.429		0.380		0.461		0.490	
4	1.090	**	1.183	**	1.027	**	0.632		0.519		0.523		0.570		0.486		0.335	
5+	1.657	***	1.702	***	1.497	***	1.246	***	1.211	**	1.103	**	0.977	*	1.065	**	1.046	**
Constant	-2.102	***	-2.047	***	-1.621	***	-1.201	***	-1.049	**	-1.088	***	-1.115	**	-1.136	**	-1.040	**
LR test chi sq	67.397	***	70.992	***	44.374	***	31.559	***	48.282	***	50.206	***	84.950	***	73.736	***	60.003	***
N	340		340		340		340		340		340		340		340		340	

Note: *: P < 0.05, ** p < 0.01, *** p< 0.005

4.5 Predicting household characteristics from power demand

The results discussed above suggest that it may be feasible to predict the number of persons in a household from their electricity power demand. However the results also suggest that the single person households in this small sample may have had higher than expected power demand patterns.

Therefore in order to attempt to predict the number of persons from log mean electricity power demand a mixed effects poisson regression (generally used for count variables) was estimated using STATA 12's xtmepoisson command[5] but excluding single person households. Like the earlier mixed effects linear model this has the form:

$$\Pr(y_{ij} = y \mid u_j) = \exp(-\mu_{ij})\ \mu^{y}_{ij}/y$$

for $\mu_{ij} = \exp(x_{ij} + z_{ij}u_j)$, j = 1, …, M clusters (households) with cluster j consisting of I = 1, …, n_j observations. The responses are counts y_{ij}.

As Table 7 shows the model was estimated for all weekday day-time time periods and then for the 16:00, 16:30 and 17:00 time periods separately as suggested by the results discussed above. The 'All time slots' model (see Table 7) suggested that log power demand could help to predict the number of household occupants but none of the other models produced significant effects for this variable.

However, as might be expected given this very small sample size, none of the models were able to successfully predict the number of household residents. The application of this approach to a larger household dataset is therefore discussed further below.

Table 7: Results of mixed effects poisson models predicting number of household occupants at all times of day and at specific times of day in the afternoon on weekdays (fixed effects part only shown)

	All			16:00			16:30			17:00		
	b	95% lower	95% upper	b	95% lower	95% upper	b	95% lower	95% upper	b	95% lower	95% upper
Log power demand	0.038	0.025	0.051	0.046	-0.033	0.125	0.051	-0.031	0.133	0.069	-0.014	0.153
Constant	1.353	1.330	1.386	1.369	1.210	1.527	1.368	1.215	1.522	1.369	1.234	1.505
N obs	9595			300			300			300		
LL	-16005.130			-500.283			-500.185			-499.633		
LR test chi sq	920.088			24.126			21.906			13.299		
Prob < chi2	0.000			0.000			0.000			0.000		

Note: *: P < 0.05, ** p < 0.01, *** p< 0.005

4.6 Estimating the probability of active occupancy

The second objective of this work was to develop algorithms for predicting the probability that a household would be actively occupied at given times of the day in order to facilitate survey fieldwork and/or census enumeration. It is important to emphasise that this information would only be of use in fieldwork processes. Whilst small area estimates of the number of occupants per dwelling (as above) might be publishable as census small area statistics, there is no intention to publish small area estimates of time-of day occupancy rates at even an aggregate level.

As noted in Section 0 above, the focus on half-hourly aggregated power demand data required the exploration of indicators that could show when power power demand in a given half-hour time period was substantially higher than might be expected were the dwelling to be unoccupied. Since there has been no published prior work in this area, the analysis reported below uses each household's mean night-time power demand (01:00 – 06:00) as the relevant baseline. Three relatively arbitrary indicators of difference from this baseline were then proposed and tested:

[5] http://www.stata.com/manuals13/xtmepoisson.pdf

1. Active occupancy = true if log power demand in a given period is higher than the upper 95% confidence limit of the household's log baseline (01:00 – 06:00) power demand. This is calculated as the mean + (1.96 * standard error of mean) but does not account for the non-independent nature of the observations;
2. Active occupancy = true if power demand in a given period is higher than the 95th percentile of the household's baseline (01:00 – 06:00) power demand;
3. Active occupancy = true if power demand in a given period is higher than the 99th percentile of the household's baseline (01:00 – 06:00) power demand;

To produce a probability indicator the proportion of time periods in which each indicator was true for each household across the 28 day period was calculated for each day of the week. Thus, for example, if household 11 was actively occupied at 08:00 on three out of the four Mondays in the dataset then the probability value for Mondays at 08:00 would be 75%.

Clearly the 99% based indicator would be expected to provide a more effective filter than the 95% based indicator and so might reduce the chances of a false positive prediction of occupancy. The descriptive analysis that follows compares the performance of each of three indicators with this in mind.

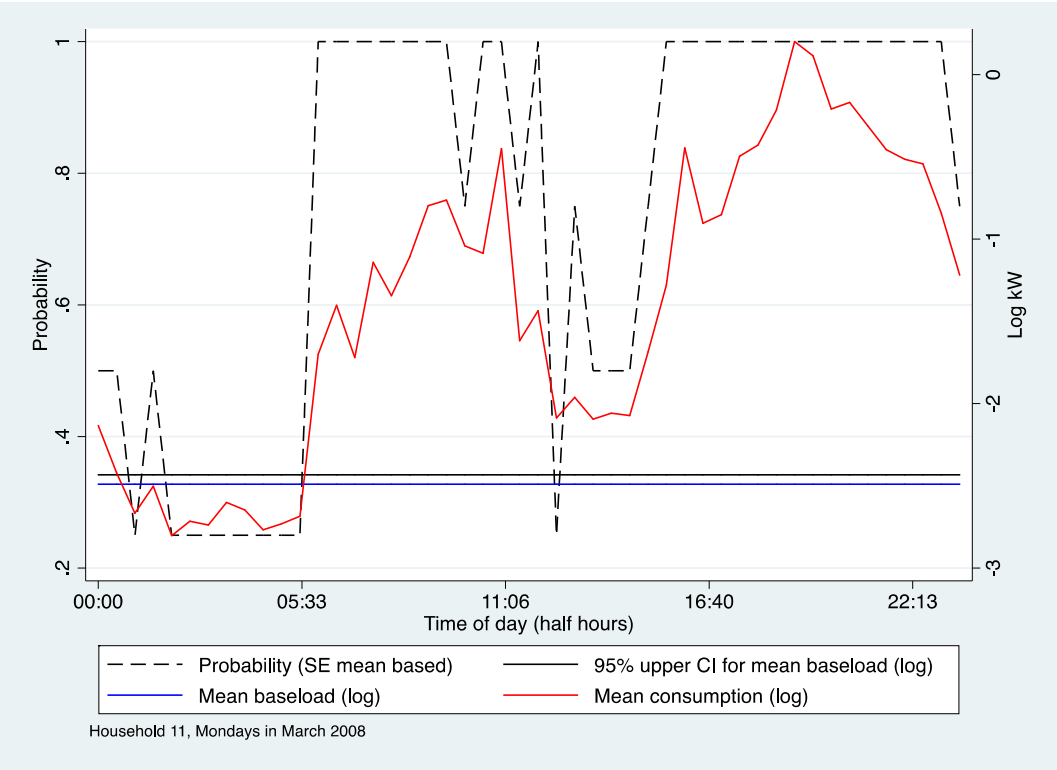

Figure 11: Distribution of power demand thresholds, levels and indicators across time of day on Mondays for household 11 (based on the standard deviation of log mean power demand)

Figure 11 shows the calculated standard error based probability for just one household by time of day for Mondays together with the power demand level and baseline thresholds whilst Figure 12 shows the same calculations but for the 95th and 99th percentile based probabilities.

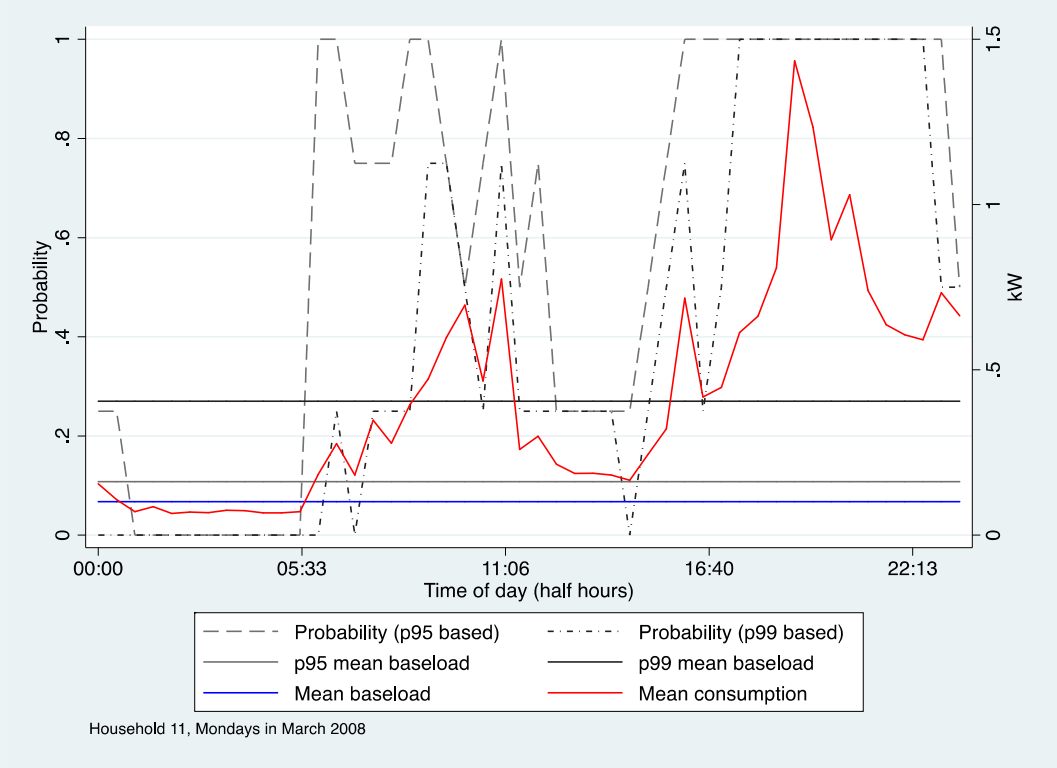

Figure 12: Distribution of power demand thresholds, levels and indicators across time of day on Mondays for household 11 (based on the 95th and 99th percentiles of mean power demand)

In both cases the occupancy indicators suggest varied levels of occupancy between 06:00 and 11:00 but 100% active occupancy between 17:00 and 22:00. This might suggest that were fieldwork to be planned for a Monday in October, household 11 could have been visited between 09:00 and 10:30 or 17:00 and 21:00 with a reasonable expectation of response. Clearly the 99th percentile based indicator provides the highest threshold and so might have an increased likelihood of concluding occupants are absent when they are actually present.

This is confirmed by Figure 13 to Figure 15 which show the distribution of occupancy probabilities for each household by day of the week in October 2008. As expected the standard error based indicator is relatively undiscerning largely due to the lack of dispersal in the log mean power demand distribution. It could therefore be termed a relatively optimistic indicator – there is an increased likelihood of concluding occupants are present when they are not.

On the other hand as would be expected the 99th percentile based indicator is much more discerning and shows relatively fewer time periods with a high calculated probability of occupancy. It could therefore be termed a conservative indicator with much lower likelihood of a false positive. Further it helps to highlight several dwellings which were predicted to be actively occupied during the day on weekdays in March – characterized by horizontal streaks on weekdays in Figure 14 or Figure 15. It can readily be seen how these results could be used as input to a fieldwork scheduling algorithm.

Of course in all cases these indicators simply replicate known patterns of energy power demand above the relatively arbitrarily chosen thresholds and it would require field experiments and/or further analysis of the one minute level data to validate the estimated probabilities of active occupancy.

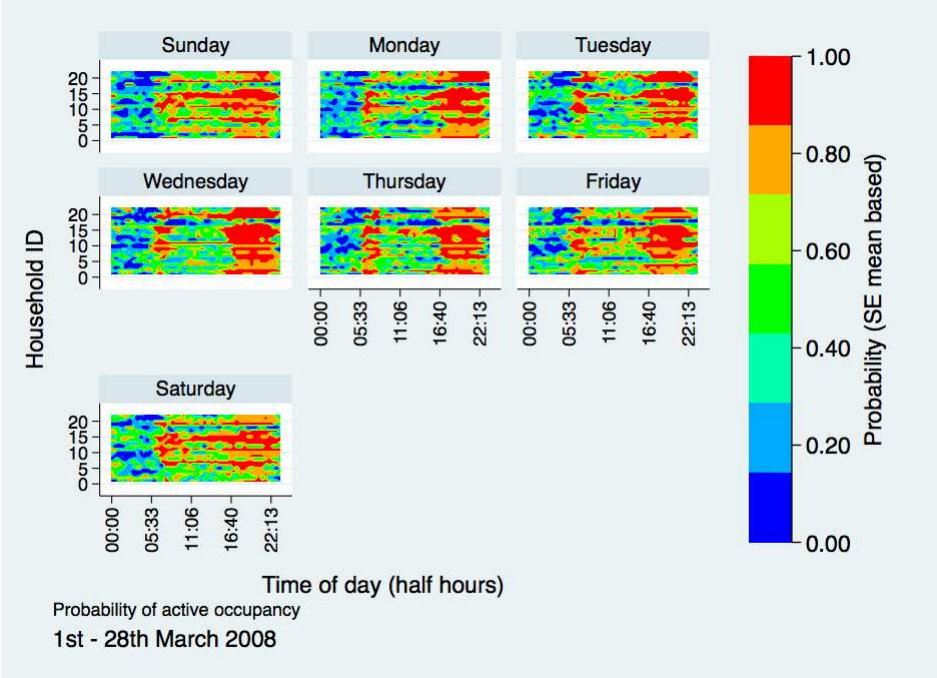

Figure 13: Distribution of mean probability of active occupancy by time of day and day of the week for all households (based on the standard error of log mean power demand)

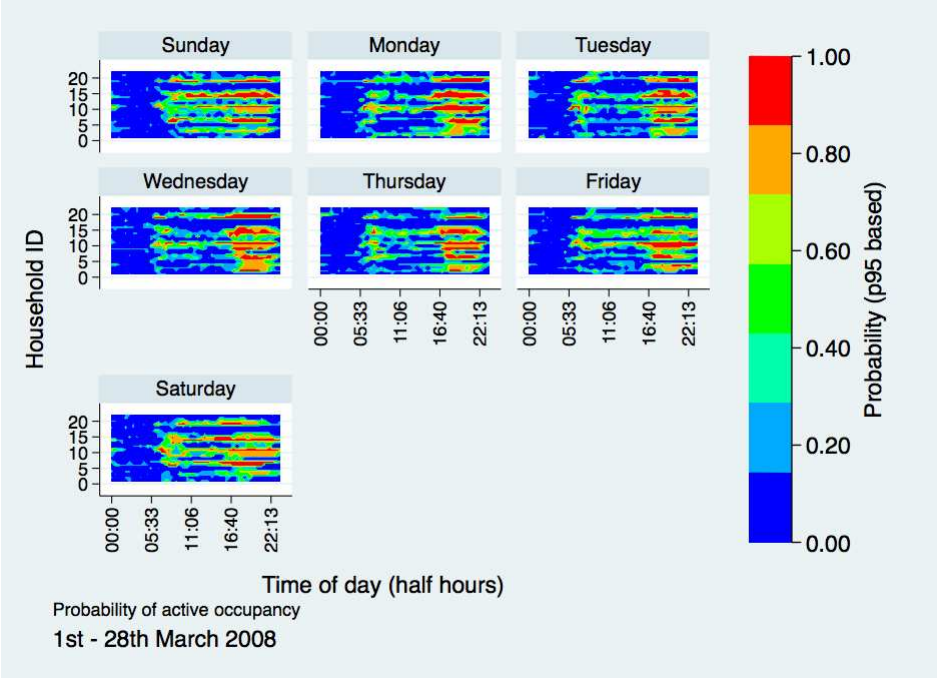

Figure 14: Distribution of mean probability of active occupancy by time of day and day of the week for all households (based on the 95th percentile of mean power demand)

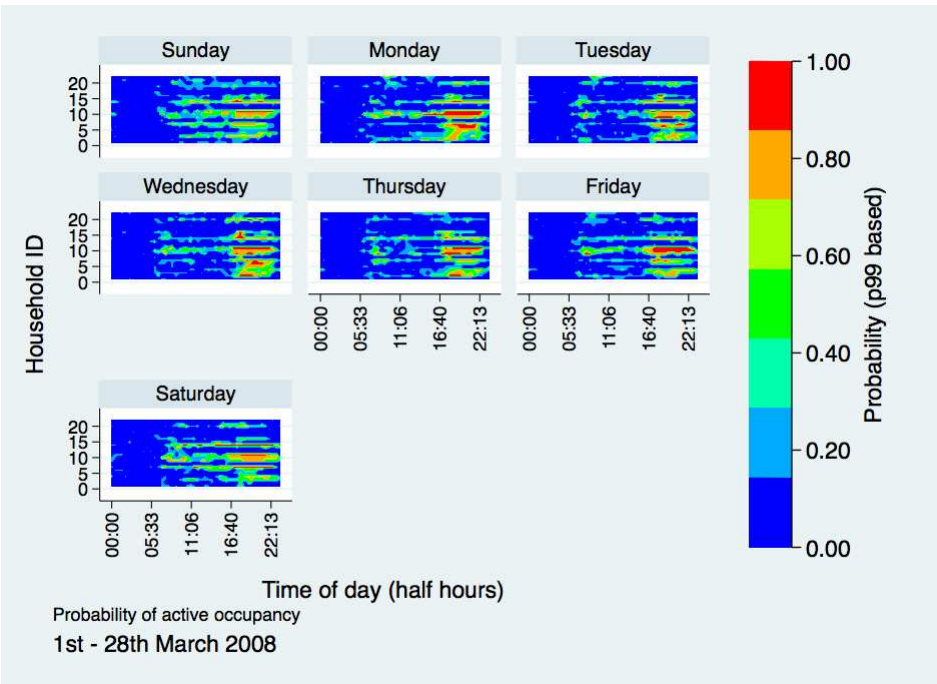

Figure 15: Distribution of mean probability of active occupancy by time of day and day of the week for all households (based on the 99th percentile of mean power demand)

4.7 Summary

Overall the work conducted using the 22 household sample of aggregated one minute level power demand data has provided hints that it may be possible to identify different household types from their temporal electricity power demand profiles. However too few socio-demographic characteristics were present in the data to support more than an explanatory analysis and the small number of households in the sample prevented effective multivariate modeling at particular times of day for the characteristics available.

It also seems clear that some form of threshold based 'occupancy indicator' could be meaningfully calculated from temporal electricity power demand data and that this could be used as input to a fieldwork scheduling algorithm.

In the next section the same analytic approaches were applied to the larger University of Southampton dataset which also provides additional household characteristics from its linked household survey.

5 One-Second Resolution Domestic Electricity Use Data, 2011

5.1 Data Background

This data is being collected by an experimental trial of energy demand reduction innovations being lead by the University of Southampton. The data, which we refer to as UoS-e consists of two linked parts:

1. A household survey that was used to recruit around 180 households in two neighbourhoods (intervention and control areas) and which was then repeated roughly every 6 months.
2. Power demand monitoring using a commercially available broadband hub-based monitoring system. The raw data is collected at 1 second intervals

The household survey data contains considerable information on household characteristics, attitudes and behaviours and allows actual recorded electricity power demand (and thus implied habits) to be analysed alongside self-reported characteristics.

5.2 Data notes

The households surveyed as part of the UoS-E project were drawn from two wards as the basis for a trial of energy demand reduction interventions using a control (~100 households in ward A) and an intervention (~80 households in ward B) group. The study deliberately omitted households that may have unusual occupancy patterns (e.g. where dwellings are unoccupied for part of the year) and any where electricity was the main mode of heating. 83% of the households were in areas (LSOAs) in the lowest deprivation decile as measured by the 2010 Index of Deprivation with 6% in the second decile and 10% in the third. It must also be recognised that households 'opted-in' to the original UoS-E study and received financial reward for doing so, which may have affected the type of household or householder represented by this dataset. As in the previous section, the results reported here hold only for this specific sample and no claims can be made with respect to the general population.

As noted above, electricity monitoring tends to produce a extremely large files. In the case of the one second level monitoring, this produces an expected 473,040,000 rows per month (2,628,000 per household) compared with only 963,600 (43,800 per household) for the Loughborough 1 minute data. To make this more manageable the data for 2011 was selected for use as described below and split into monthly files.

5.3 Data Processing

The majority of round one of the household surveys for this group of households were carried out through the spring of 2011 and the survey data provides crucial household attributes of interest to this feasibility study, namely the presence of children, the presence of elderly (65+) persons and the number of occupants. Unfortunately it cannot distinguish between school age (5+) and younger children.

Whilst it would have been preferable to use data from just after the initial survey period to ensure close correspondence with the survey responses, full sample recruitment was not completed until late Autumn 2011 with a small lag in installation of the monitoring instruments. In order to maximize the data available the period October 1st to 28th 2011[6] was selected as by this time over 95% of households had been interviewed and monitoring instruments had been installed.

The one second level instantaneous electricity power demand data for October 1st to 28th 2011 file was extremely large (293 million records, 22Gb) which needed to be transformed and reduced before statistical analysis could begin. This transformation included checking for duplicate records and counting the number of null recordings (i.e. an observation with null power recorded) before aggregating to one minute intervals for each household (mean power consumed, total number of

[6] Again, this avoided the clock change at the end of British Summer Time (31st October)

observations, total number of null observations). This was carried out using a number of unix commands and processing scripts on the University of Southampton's Iridis4 high performance computer[7]. Even so this processing and aggregation took over 30 minutes to produce a file of 5,233,800 one minute observations (c 110Mb) which could then be read into STATA. As Caroll et al note, a dataset comprising a larger number of households would have needed some form of sampling before aggregation could take place.

At this level completely missing one second level observations were 'filled in' using the same method described above and partially missing one second level observations were flagged by calculating the number missing based on the known number observed (0-60) and the number expected (60). Finally the number of 'OK' one second level observations was calculated for each 1 minute period by subtracting the number of null observations from the number observed.

Despite selecting a period towards the end of the study installation phase, monitoring data was only available for 128 of the 184 households and the 'completely missing' households tended to be those who had been instrumented later in the study (Figure 16). Whilst the 'filling' procedure described above provides some indication of missing data for households where at least one observation is recorded, completely missing households cannot be imputed using this method.

The study team indicated that difficulties with the data logging instrumentation in these relatively early stages might have resulted in the reduced data availability. However the size of the raw data (see above) meant that it was not possible within the resources of this project to conduct an audit of the time periods in which the maximum number of households would have been available.

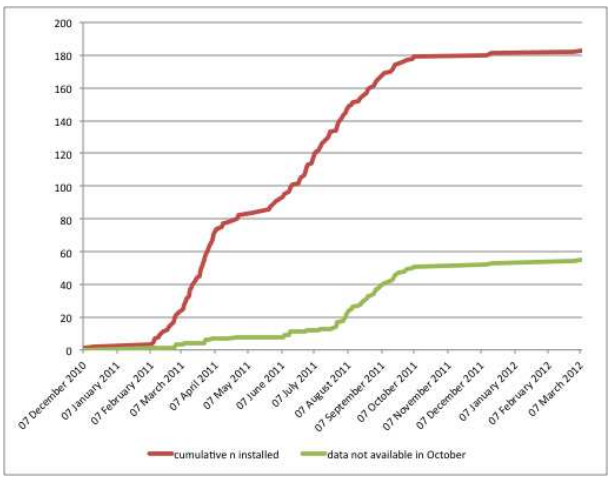

Figure 16: Cumulative installed instrumentation and households with no recorded data in October 2011 by date installed

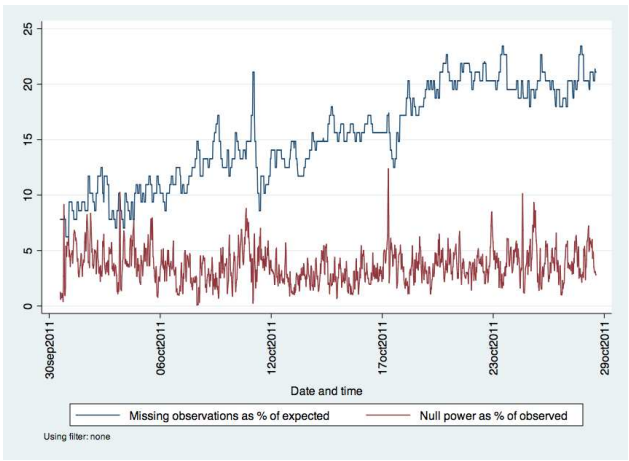

Figure 17: Distribution of missing and null observations during October 2011

The one minute level data was then aggregated to 171,776 half-hour observations for each day for each household (mean power, number of observations, number of nulls) and linked to the survey data.

As with the Loughborough data a number of analyses were conducted to assess the level of missing data (i.e. no observations for a given period) and null recordings. Overall 84% of the half hour periods contained all 1800 1 second level observations whilst 15% had no 1 second level observations at all. The remaining 1% had variable levels of missing observations. The levels of missing data tended to increase through the sample period but the level of nulls remained roughly constant with apparent random fluctuations (Figure 17).

According to the study team, one of the most common reasons for missing data was the temporary loss of the household broadband connection, which required the monitoring hub to be reset. Since the loss of the broadband connection was often driven by demand fluctuations at the local exchange, the

[7] http://cmg.soton.ac.uk/iridis

pattern of missing data across households in the same study area may be 'non-random' (since these households are likely to share the same exchange). Anecdotal reports suggested that missing data was more likely in the ward A area, especially around 8am and 10pm on a Tuesday and an additional ward A issue is indicated in Figure 17 by the sharp 'missing' peak on the 10th October when there was an area-wide broadband failure.

Further analysis by time of day (Figure 18) and by household (Figure 19) suggests that there may indeed be a temporal pattern to the missing data. More significantly most of the missing records for those households who recorded at least one observation were concentrated in just a few households. Indeed 4 households had over 90% missing data and just 11 had over 70%.

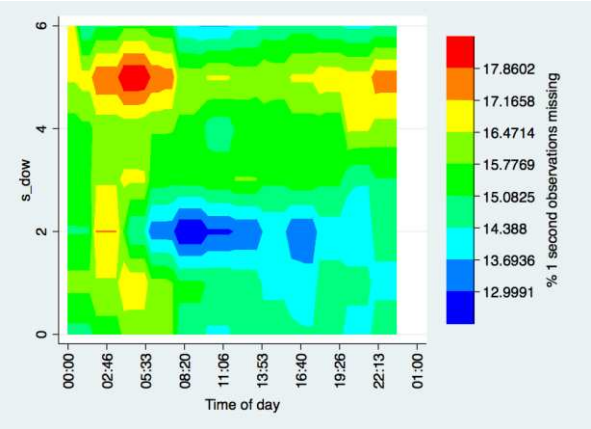

Figure 18: Distribution of missing observations by day of the week during October 2011

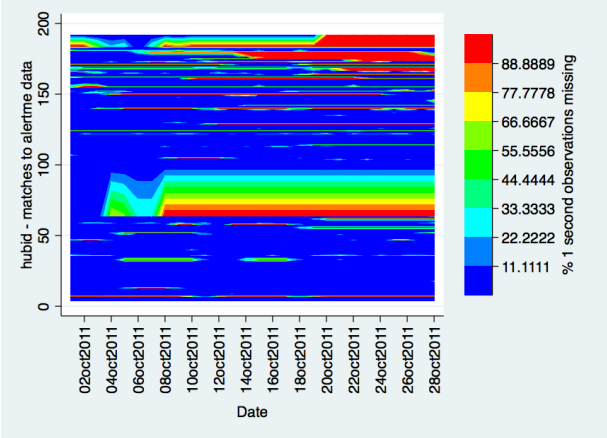

Figure 19: Distribution of missing observations by hubid during October 2011

With this in mind a quality filter was implemented such that observations were only included if at least 50% of the observations in a given half hour were valid (i.e. not missing and not null). This had the effect of removing 18% of the observations and in addition a further two households were removed as their recorded power values were extremely high but there were also randomly distributed peaks and/or constantly high power demand and thus were considered extreme outliers. The remainder of the analysis was therefore carried out on 126 households.

Unlike Caroll et al (Caroll et al., 2013) households whose missing levels were high were not removed completely because to do so would have reduced the sample size still further. However in common with Caroll et al, a relatively high proportion of project resource had to be devoted to processing, checking and cleaning the data before analysis could begin. With the size of data files in use this was a non-trivial activity.

5.4 Power demand by household characteristics

Mean instantaneous power demand at the half hour level was similar to the Loughborough data at 0.475 kW across all 126 remaining households. As before, the distribution of this value was highly positively skewed with a median of 0.309 kW and a skewness of 3.09.

Table 8 shows the number of households in each group whilst Figure 4 shows the overall mean instantaneous power demand for all households by time of day and day of the week and as before reveals the expected pattern of morning and evening peaks with particularly noticeable weekday troughs in demand in between more evenly distributed demand at weekends. However it was also noticeable that the last week of the period showed less differentiation between weekdays and weekends as this period was school half term. Given that the approach under investigation relied on differentiation driven to some extent by work/school patterns the final week of the period was therefore excluded from subsequent analysis. This left a sample of 106,756 half-hourly observations from the 1st to the 21st October 2011 across the 126 households.

Table 8: Household counts by group

N people	N	%	Any senior	N	%
1	6	4.69	No	108	84.38
2	45	35.16	Yes	20	15.62
3	20	15.62			
4	33	25.78			
5+	24	18.75			
Total	128			128	
N children			**Presence of children**		
0	58	45.31	No	58	45.31
1	19	14.84	Yes	70	54.69
2	32	25.00			
3+	19	14.84			
Total	128			128	

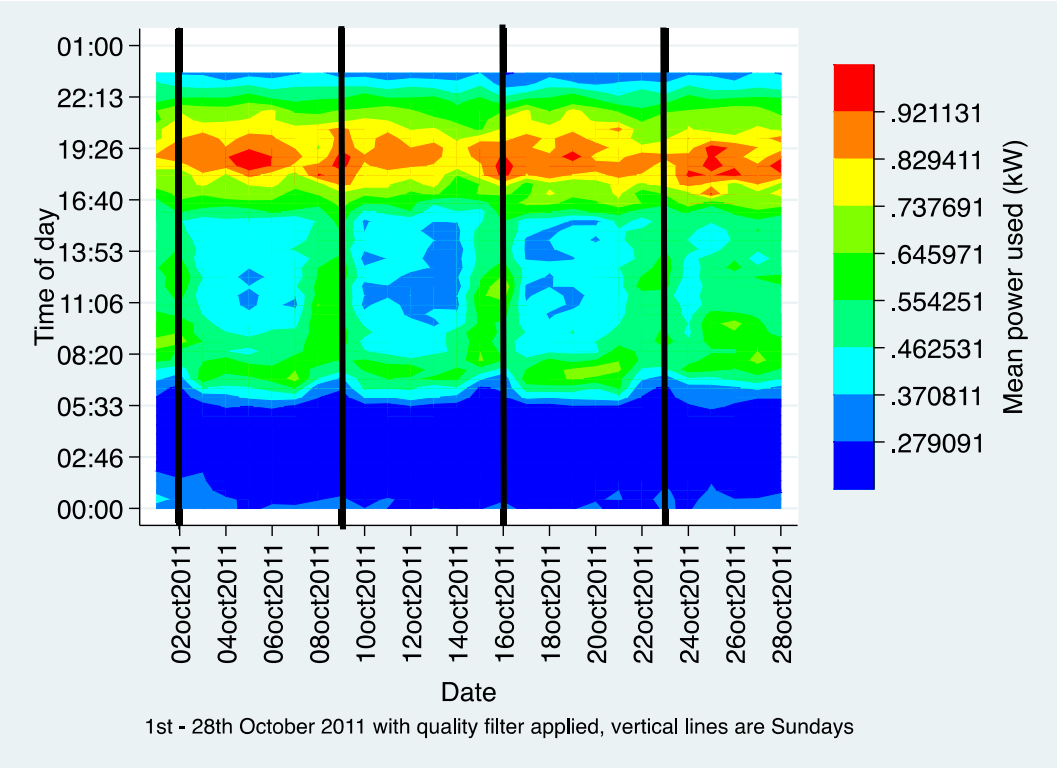

Figure 20: Mean instantaneous power demand by time and day for October 2011

As before, comparison of the weekday and weekend patterns (not shown) suggested that there is more differentiation between the groups during weekdays, again indicating the role of the temporal constraints of school and/or work during weekdays.

Figure 21 to Figure 23 show the distribution of mean, median instantaneous power demand and interquartile (25% - 75%) range by time of day for weekdays for each of the groups of interest. It is immediately obvious that the interquartile ranges are relatively wide for all groups, especially those with multiple occupants (c.f. Figure 21). Whilst this variance could itself provide the basis for some form of indicator it also suggests that for this relatively small sample at least, there may be insufficient within-group homogeneity to easily differentiate between groups on the basis of overall instantaneous power demand alone.

As expected, larger households record higher power demand at both the morning and evening peak periods as do households with no residents under 65 and those with children. Households with

residents aged 65 and over appear to have a far less pronounced morning peak, perhaps due to lack of labour market participation and also a less pronounced evening peak. To some extent this pattern was also true of households without children. Since 17 (30%) of the households without children contained residents aged 65+, there may be confounding effects although Figure 24, which excludes households with residents aged 65+, suggests these may be limited in magnitude. This would suggest that a more complex multinomial household categorization might help reveal such distinctions but, given the small size of the sample, this has not been pursued in the current work[8].

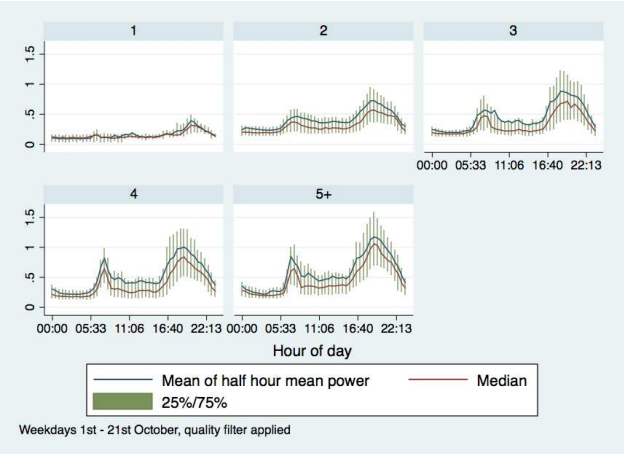

Figure 21: Power demand by number of occupants for weekdays 1st – 21st October 2011 (for totals see Table 8)

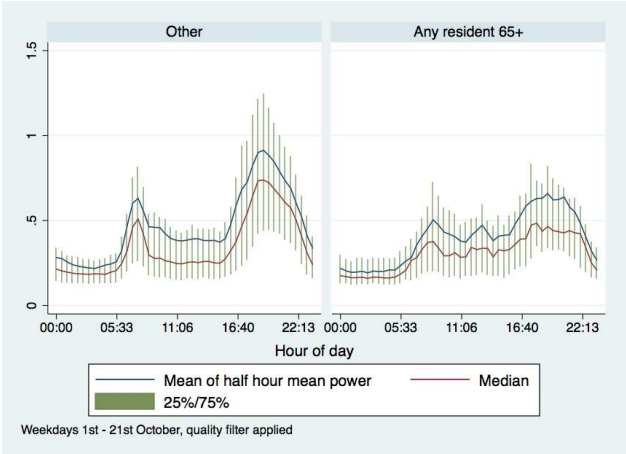

Figure 22: Power demand by presence of seniors for weekdays 1st – 21st October 2011

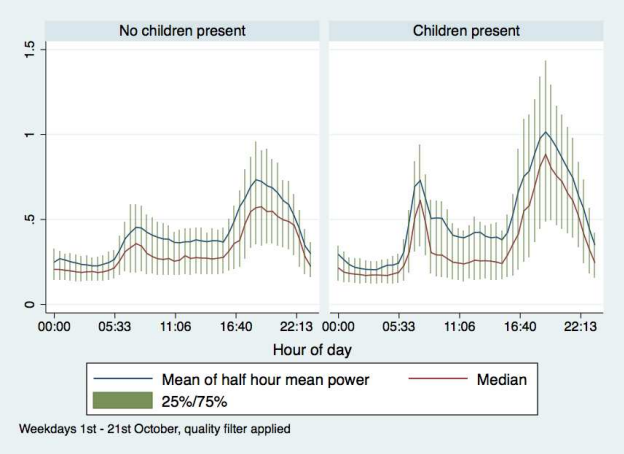

Figure 23: Power demand by presence of children for weekdays 1st – 21st October 2011

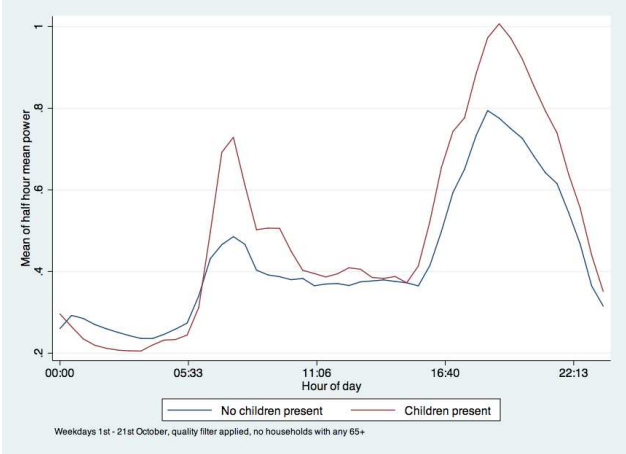

Figure 24: Power demand by presence of children for weekdays 1st – 21st October 2011, households with any residents aged 65+ excluded

Overall the patterns suggest that general power demand levels may not be a particularly powerful way to differentiate between household types although distinctions between the number of occupants may be possible despite the obviously high levels of variation within categories as the width of the middle 50% (25%-75%) bars in the above figures suggests. However the charts do suggest that it may be useful to explore:

- The ratio of morning peak demand to mid-day demand as this may differentiate between households with/without residents aged 65+ and with/without children;

[8] See also Caroll et al's discussion of this issue with respect to the relative accuracy of the prediction of binomial and multinomial classifications (Caroll et al., 2013, p. 10).

- The ratio of mid-day demand to evening demand as this may differentiate between households with/without residents aged 65+;
- The ratio of morning demand to evening demand as this may differentiate between households with/without residents aged 65+ and with/without children.

5.5 Predicting demand from household characteristics

As before, given the descriptive results the same mixed effects multilevel regression approach was used to estimate the role of household size, presence of seniors and presence of children in predicting mean power demand in all half-hour periods in the day (07:00-23:00) for weekdays and for each half hour period at specific times of day when the largest differences may be apparent (07:00 – 09:00 and 16:00 – 20:00 on weekdays).

Mean power demand values were again log transformed to mitigate the skewed nature of the distribution (see Annex 1.2), STATA 12's xtmixed estimation command was used and in the first model (see Table 9) only the household characteristics were included but in the second model time of day was again included as a series of half-hour dummy variables. As before, diagnostic plots of residuals are for Model 1 are reported in Annex 1.4 and whilst the centre of the power demand distribution is unproblematic there may be problems at the tails even when log transformed. Future work could extend the work reported here by filtering out extremely high or low power demand values. Model 2 diagnostics were indistinguishable from Model 1.

As Table 9 shows the number of persons in a household was a good predictor of power demand in Model 1 with larger number of occupants driving increased power demand albeit in a non-linear fashion (four people do not use twice as much as two). These effects were maintained when the time of day variable was introduced (Model 2) and as before the coefficients in this model make clear the shape of the overall weekday temporal power demand profile. Wald tests suggested that the coefficients for 2 and 4 person households were significantly different from 5 person households. In addition Model 1 suggests that the presence of children and of seniors are both good predictors of power demand but this is not the case in Model 2 suggesting that these effects are confounded by time of day demand patterns.

The models for each half hour throughout the day confirmed the result for number of persons with this variable proving statistically significant in all separate half hour slots tested (09:00 – 22:00). However the model for 08:30 (shown with the models for 07:00 and 08:00 in Table 10) suggests an additional positive effect for households with residents aged over 65. Both the 08:00 and the 09:00 (not shown) models estimate a negative but not statistically significant effect for the presence of children as one might expect given that they would have left for school by this time.

Table 11 shows the results for 13:00 and for the early afternoon which were the only models tested in which variables other than the number of occupants were close to being statistically significant. The model for 13:00 estimated a large positive effect for households with residents aged 65+ which perhaps corresponded to lunchtime cooking but this was statistically significant only at the 10% level. The models for 17:30 and 18:00 both showed a significant but unexpectedly negative effect for the presence of children.

Whilst these results suggest that differences in power demand patterns may be driven by different household composition and that this may be detectable at particular times of day, it also implies that multiple effects are likely to confound the attempt to estimate the model in reverse.

Table 9: Results of model estimating effects of household characteristics on log mean power demand at the half hour level (weekdays, 07:00 – 22:00)

		Model 1				Model 2			
		b	CI (lower)	CI (upper)	Sig	b	CI (lower)	CI (upper)	sig
Fixed effects part									
Number of people (1)									
	2	0.711	0.591	0.831	***	0.686	0.089	1.282	*
	3	1.045	0.904	1.186	***	1.027	0.324	1.729	**
	4	0.952	0.806	1.099	***	0.941	0.212	1.671	*
	5+	1.355	1.210	1.500	***	1.350	0.627	2.074	***
Children (no) Yes		-0.255	-0.335	-0.176	***	-0.256	-0.659	0.147	
Seniors (No) Yes		0.121	0.047	0.195	**	0.100	-0.265	0.466	
Time (07:00) 07:30						0.087	0.038	0.136	***
	08:00					-0.038	-0.087	0.011	
	08:30					-0.271	-0.320	-0.222	***
	09:00					-0.297	-0.346	-0.248	***
	09:30					-0.329	-0.378	-0.280	***
	10:00					-0.375	-0.424	-0.326	***
	10:30					-0.406	-0.455	-0.357	***
	11:00					-0.437	-0.486	-0.388	***
	11:30					-0.439	-0.488	-0.390	***
	12:00					-0.396	-0.445	-0.347	***
	12:30					-0.389	-0.438	-0.340	***
	13:00					-0.372	-0.421	-0.323	***
	13:30					-0.399	-0.448	-0.350	***
	14:00					-0.428	-0.477	-0.379	***
	14:30					-0.408	-0.457	-0.359	***
	15:00					-0.420	-0.469	-0.371	***
	15:30					-0.362	-0.411	-0.313	***
	16:00					-0.205	-0.254	-0.156	***
	16:30					-0.028	-0.077	0.021	
	17:00					0.110	0.061	0.159	***
	17:30					0.216	0.167	0.265	***
	18:00					0.372	0.323	0.421	***
	18:30					0.472	0.423	0.521	***
	19:00					0.507	0.457	0.556	***
	19:30					0.487	0.438	0.536	***
	20:00					0.452	0.403	0.501	***
	20:30					0.405	0.356	0.454	***
	21:00					0.318	0.269	0.367	***
	21:30					0.261	0.212	0.310	***
	22:00					0.136	0.087	0.185	***
	22:30					-0.066	-0.115	-0.017	**
Constant		-1.768	-1.933	-1.603	***	-1.678	-2.263	-1.094	***
Random effects part									
Time of day	sd(constant)	0.333	0.258	0.430					
Household	sd(constant)	0.746	0.728	0.765		-0.371	-0.496	-0.245	
Residuals	sd(constant)	0.644	0.640	0.648		-0.353	-0.359	-0.346	
N		49991				49991			
Log likelihood		-54456				-53668			
LR test chi sq		32419				28122			
p(LR test chi sq)		0.000				0.000			

Note: *: P < 0.05, ** p < 0.01, *** p< 0.005

Table 10: Half-hour models (morning) – fixed effects part only

	07:00				08:00				08:30			
	b	CI (lower)	CI (upper)	sig	b	CI (lower)	CI (upper)	sig	b	CI (lower)	CI (upper)	sig
Number of people (1)												
2	0.981	0.344	1.618	**	1.058	0.383	1.733	**	1.082	0.360	1.804	**
3	1.290	0.544	2.036	***	1.345	0.554	2.136	***	1.368	0.525	2.212	**
4	1.226	0.449	2.004	**	1.141	0.316	1.965	**	1.160	0.281	2.039	**
5+	1.672	0.903	2.442	***	1.585	0.770	2.401	***	1.475	0.606	2.344	***
Children (No)												
Yes	0.021	-0.404	0.445		0.036	-2.815	-1.492	***	-0.141	-0.618	0.335	
Seniors (No)												
Yes	-0.134	-0.524	0.256		0.167	-0.247	0.581		0.45	0.008	0.903	*
Constant	-2.071	-2.694	-1.448	***	-0.291	-0.425	-0.156	***	-2.332	-3.039	-1.625	***
N	1,563				1564				1,560			
Log likelihood	-1439.252				-1621.780				-1883.148			
LR test chi sq	1147.937			***	977.333			***	693.209			***

Note: *: P < 0.05, ** p < 0.01, *** p< 0.005

Table 11: Half-hour models (afternoon/evening) - fixed effects part only

	13:00				17:30				18:00			
	b	CI (lower)	CI (upper)	sig	b	CI (lower)	CI (upper)	sig	b	CI (lower)	CI (upper)	sig
Number of people (1)												
2	0.800	0.154	1.445	*	0.820	0.233	1.407	**	0.821	0.333	1.309	***
3	1.030	0.275	1.785	**	1.269	0.585	1.952	***	1.238	0.671	1.806	***
4	0.974	0.187	1.762	*	1.353	0.640	2.065	***	1.408	0.816	1.999	***
5+	1.354	0.575	2.132	***	1.652	0.947	2.356	***	1.618	1.033	2.203	***
Children (No)												
Yes	-0.306	-0.734	0.123		-0.412	-0.795	-0.029	*	-0.390	-0.707	-0.073	*
Seniors (No)												
Yes	0.392	-0.009	0.794	+	0.106	-0.255	0.466		0.031	-0.266	0.327	
Constant	-2.116	-2.748	-1.485	***	-1.622	-2.197	-1.047	***				
N	1,564				1,567				1,560			
Log likelihood	-1737.286				-1829.457				-1715.761			
LR test chi sq	637.745			***	499.103			***	375.409			***

Note: +:p < 0.1, *: p < 0.05, ** p < 0.01, *** p< 0.005

5.6 Predicting household characteristics from power demand

5.6.1 Number of persons

As in the previous work with the smaller Loughborough dataset, a mixed effects poisson model was estimated using log mean power demand first for all time periods and then for peak (07:00 – 08:30 and 16:00 – 20:00) periods when the difference between the curves appeared likely to be highest (c.f. Figure 21). However the only models that produced statistically significant results were those for 07:00 and 07:30 (see Table 12 and Figure 25) suggesting that it may be possible to distinguish between different numbers of residents at these time periods.

Table 12: Mixed effects poisson model results (selected) – fixed effects part only

	All time periods			07:00			07:30		
	b	CI (lower)	CI (upper)	b	CI (lower)	CI (upper)	b	CI (lower)	CI (upper)
Log power demand	0.019	0.012	0.026	0.057	0.009	0.104	0.067	0.017	0.117
Constant	1.092	1.078	1.106	1.124	1.042	1.205	1.128	1.048	1.207
N	49991			1563			1560		
Log likelihood	-79154.620			-2472.862			-2468.922		
LR test chi sq	11804.350			316.883			300.842		
p(LR test chi sq)	0.000			0.000			0.000		

Note: *: P < 0.05, ** p < 0.01, *** p< 0.005

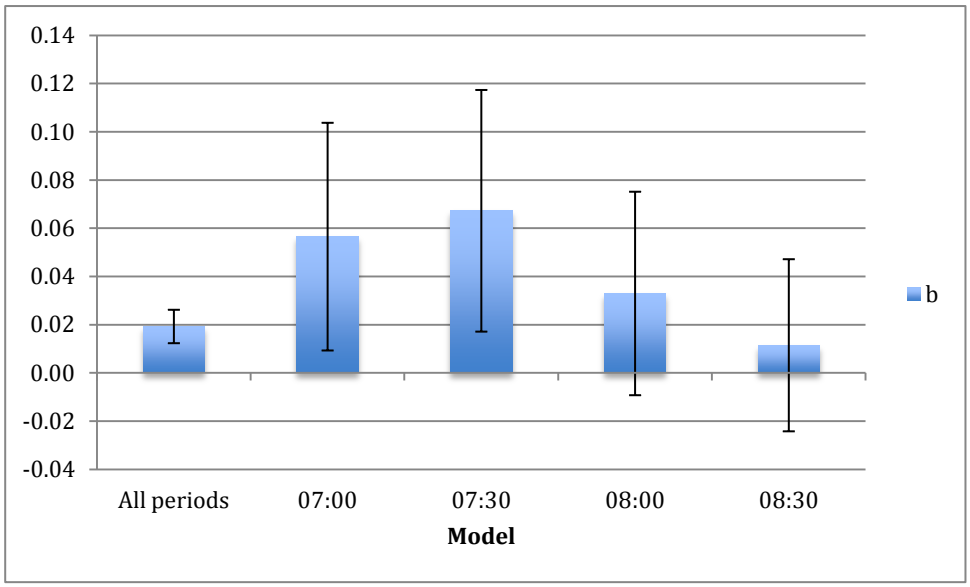

Figure 25: Estimated coefficients and confidence intervals for log power demand in different time of day models. Error bars represent the 95% confidence intervals for the point estimates (b)

Although the degree of imprecision indicated by the 95% confidence intervals for the point estimates (see Figure 25) reflects the degree of variation in power demand discussed above predicted counts were reasonably correlated with the observations (see Figure 26) suggesting that this approach has some value in predicting the number of occupants of a household.

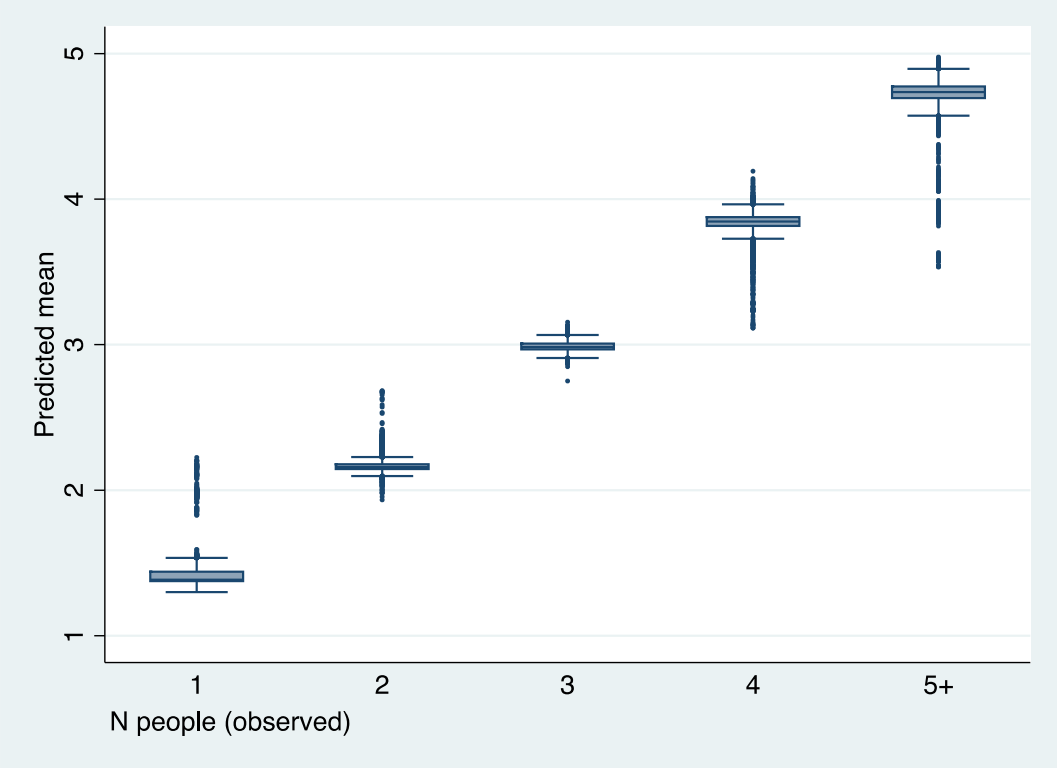

Figure 26: Predicted counts of persons per household by observed number of persons per household (All time periods, model pairwise correlation rho = 0.996, results for time specific models were almost identical)

5.6.2 Presence of children

In order to attempt to predict the presence of children, an initial logistic panel model regressing the presence of children on log power demand was estimated for all time periods and then, as before, for each half hour in peak periods. None of these models produced satisfactory results due to maximum likelihood estimation failure.

Based on the temporal patterns discussed in Section 5.4 a number of comparative indicators were then constructed:

1. The ratio of the mean morning peak (07:00-08:30) power demand to the mean mid-day period (11:00 – 14:30) power demand;
2. The ratio of the 95th and 99th percentiles for the morning mean and mid-day means
3. The ratio of maximum morning power demand to mean mid-day power demand
4. The ratio of mean mid-day power demand to mean evening (17:00 – 20:30) power demand
5. The ratio of mean morning (07:00-08:30) power demand to mean evening power demand (17:00 – 20:30)

Although it would have been possible to calculate the ratio for each day or for each week-day, initially these ratios were calculated at the household level across all weekdays in the sample. A household level logistic model predicting the presence of children was then estimated for each indicator. Only two of the indicators produced statistically significant results (see Table 13 and Figure 27) but they do suggest that it may be possible to use characteristics of the household's daily profile to predict the presence of children. Post-estimation classification tests using a 50% probability threshold suggest that 70% of households with children were classified correctly using the Mean morning: mean mid-day indicator but only 60% for the Mean mid-day: mean evening indicator (see Table 13). False positives and negatives have roughly similar and relatively low rates in the Mean morning: mean mid-day model but there is a tendency (false positive = 56%) for households who do not have children to be classified as households who do in the Mean mid-day: mean evening model.

Table 13: Estimation model and classification test results for the presence of children using the ratio of mean morning peak to mid-day mean power demand and for the ratio of mid-day mean to evening mean.

	b	CI (lower)	CI (upper)			b	CI (lower)	CI (upper)	
Mean morning: mean mid-day	1.142	0.487	1.797	***	**Mean mid-day: mean evening**	-2.724	-4.448	-1.00	**
Constant	-1.543	-2.549	-0.537	***		1.535	0.606	2.464	**
N	121					121			
Pseudo R-squared	0.105					0.066			
chi2	17.47					10.98			
p	0.000					0.001			
ll	-74.6					-77.9			
Hosmer-Lemeshow chi2	13.36 (p = 0.101)			+		13.95 (p = 0.08)			+
Correctly classified	70.25%					60.30%			
False positive rate	25.45%					56.36%			
False negative rate	33.33%					25.76%			

Note: +:p < 0.1, *: p < 0.05, ** p < 0.01, *** p< 0.005

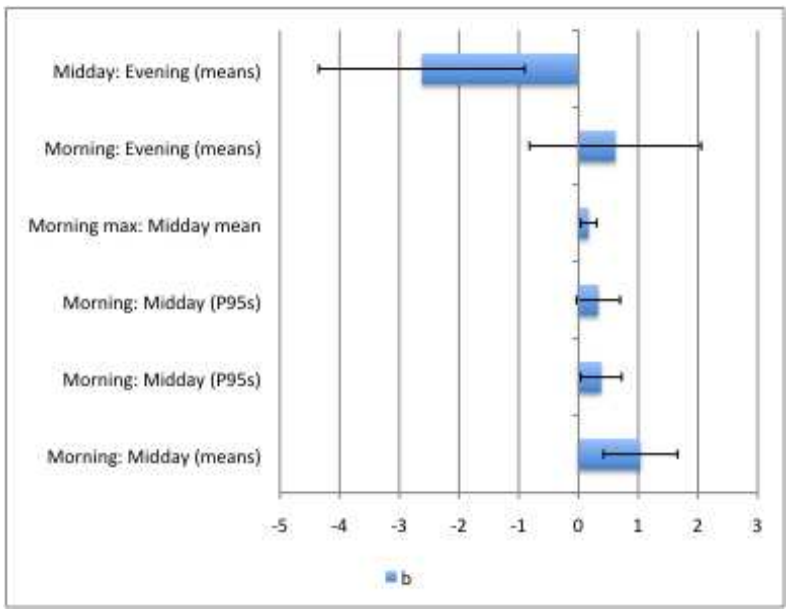

Figure 27: Logistic estimation model results for the presence of children using all ratio-based indicators. Error bars represent the 95% confidence intervals for the point estimates (b)

5.6.3 Presence of residents aged 65+

The presence of residents aged 65+ was tested using the same two approaches as for the presence of children. Thus the first was a logistic panel model using log mean power demand for all time periods and then each time period separately. Although all models reached convergence, none of the half-hour models were able to successfully predict the presence of residents aged 65+ although the 13:00 model came closest with a p value of 0.12 for log mean power demand.

The second approach was to estimate a non-panel logisitic model using the summary indicators described above for the presence of children to predict the presence of residents aged 65+ (cf Section 5.4) at the household level. Only one of these models produced statistically significant results – the ratio of mid-day to evening mean power demand (see Table 14) as might be expected from the discussion in Section 5.4. Post-estimation classification tests for this model suggested a correct classification rate of 84% however there is a false negative rate of 94% suggesting that the approach

tends to classify households as non-seniors when seniors are in fact present. This is perhaps unsurprising given the poor fit of the model as indicated by the Hosmer-Lemeshow test reporting a non significant result.

Table 14: Estimation model results for the presence of residents aged 65+ using the ratio of mid-day mean to evening mean.

	b	CI (lower)	CI (upper)	
Mean mid-day: mean evening	3.700	1.519	5.881	***
Constant	-3.833	-5.289	-2.377	***
N	121			
Pseudo R-squared	0.124			
chi2	15.578			
p	0.000			
ll	-44.598			
Hosmer-Lemeshow chi2	3.14 (p = 0.925)			
Correctly classified	84.30%			
False positive rate	1.94%			
False negative rate	94.44%			

Note: *: P < 0.05, ** p < 0.01, *** p< 0.005

5.7 Estimating the probability of active occupancy

Building on the work reported in Section 4.6 above, the more differentiating p95 and p99 based indicators were re-implemented for this larger dataset. As before these were:

1. Active occupancy = true if power demand in a given period is higher than the 95th percentile of the household's baseline (01:00 – 06:00) power demand;
2. Active occupancy = true if power demand in a given period is higher than the 99th percentile of the household's baseline (01:00 – 06:00) power demand;

As an exemplar, Figure 28 shows the 95th and 99th percentile based probabilities for household 11 on Mondays which was a couple where at least one resident was aged 65+ and the respondent was retired.

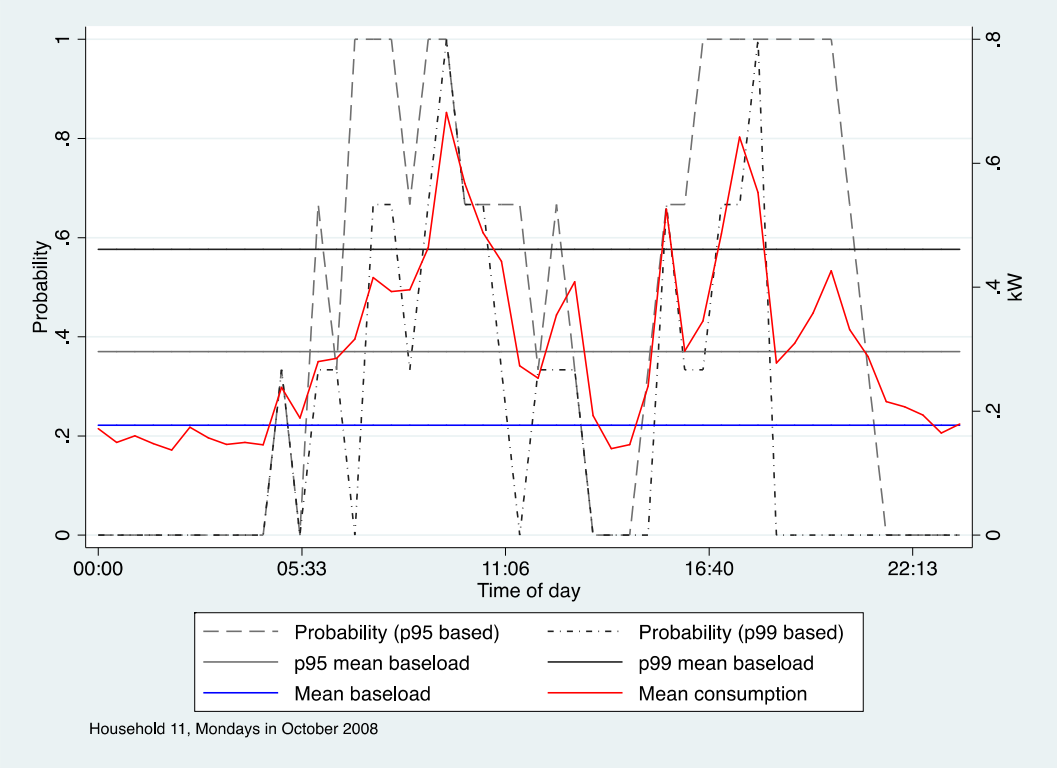

Figure 28: Distribution of power demand thresholds, levels and indicators across time of day on Mondays for household 11 (based on the 95th and 99th percentiles of mean power demand)

In both cases the occupancy indicators suggest varied levels of occupancy between 06:00 and 11:00 but with high probability of occupancy between 07:30 and 08:30 and again between 09:00 and 10:00 with a lower probability at 12:30. As with the previous example the probability increases to 100% between 16:30 and 20:00 on the 95% indicator but only up to 18:00 on the 99% indicator. Again, this might suggest that were fieldwork to be planned for a Monday in October, household 11 could have been visited between 09:30 and 10:30, at 12:30 or 17:00 - 18:00 with a reasonable expectation of response.

Figure 29 and Figure 30 show the distribution of occupancy probabilities for each of the first 30 households by day of the week in October 2011. As expected the 99th percentile based indicator is much more discerning and shows relatively fewer time periods with a high calculated probability of occupancy.

Again, this indicator helps to highlight several dwellings who were predicted to be actively occupied during the day on weekdays in October who could therefore be selected for fieldwork at particular times. As with the previous work however, validation of these estimates would require either experimental fieldwork or, in the case of this particular data, analysis of sub-1 minute level power demand to identify switched appliances that could be more confidently linked to active occupancy.

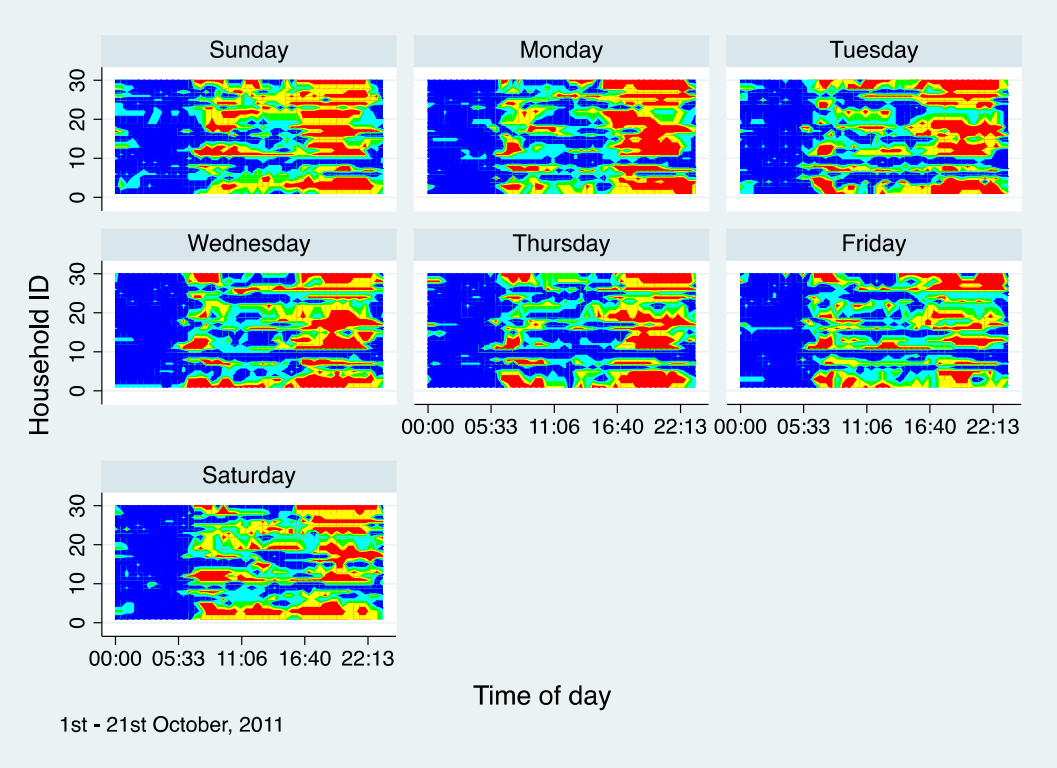

Figure 29: Distribution of mean probability of active occupancy by time of day and day of the week for first 30 households (based on the 95th percentile of mean power demand)

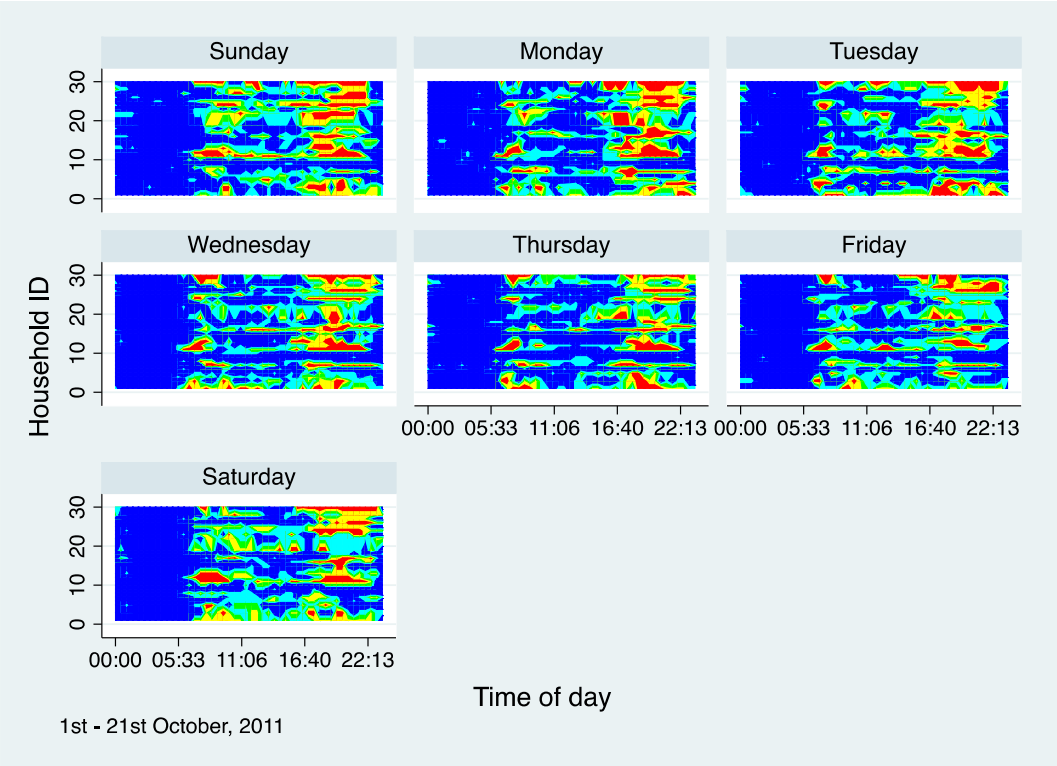

Figure 30: Distribution of power demand thresholds, levels and indicators across time of day and day of the week for first 30 households (based on the 99th percentile of mean power demand)

5.8 Summary

In general the preliminary analysis conducted with the larger sample has given increased confidence that different kinds of households may be detected using half-hour level electricity power demand data. The descriptive results in particular suggested a number of distinct temporal patterns but as was noted the variation to be found within household groups meant that statistically significant differences between them were more difficult to discern. It is possible that this may be caused by the relatively small sample size but it is also possible that the inherent variation in electricity power demand within and between households with very similar socio-demographic characteristics may make such differentiation difficult even with a larger household sample when using data at the half hour level.

There was good evidence that the number of residents in a household correlated with overall power demand and with higher levels of power demand at particular times of day. Both the overall model and a model for two time periods during the morning peak suggested a statistically significant correlation between power demand (as an independent variable) and number of occupants. Analysis of the predicted counts following the fitting of a poisson model suggested that there may be some value in using this approach to estimate the number of occupants of a dwelling.

There was also some evidence that the presence of children affected overall electricity power demand in the afternoon periods (Table 11) but models based on power demand failed to provide useful results. However models based on the ratio of morning to midday power demand and the ratio of midday to evening power demand were able to predict the presence of children with the former having a correct classification rate of 70% and relatively low false negative/positive classifications rates using a 50% threshold.

There was also evidence that the presence of residents aged 65+ affected overall electricity power demand at specific times of day (Table 10 & Table 11) but as was the case for children, models based on overall power demand did not produce useful results. However there was also evidence that the ratio of mid-day to evening power demand was affected by the presence of residents aged 65+ and in this case the correct classification rate was 84%. However there was also a false negative rate of 94% suggesting that the approach tends to classify households as non-seniors when this is not the case.

Finally, as was the case with the smaller sample result discussed in Section 4.6, the 95th and especially 99th percentile based indicators were able to highlight households which could be approached at different times of the day with a reasonable expectation of contact.

6 Conclusions

The work reported here set out to explore the feasibility of developing methods for estimating small area 'census-like' indicators from transactional 'big data' sources. To do this, the project explored the feasibility of using two samples of household electricity power demand data to:

1. Assess the feasibility of predicting at the household level:
    1. the *number* of occupants;
    2. the presence of *children*;
    3. the presence of *single persons or couples aged 65+.*
2. Assess the feasibility of predicting whether occupants at a given address will be 'at home' (*active occupancy*) at given times of the day and days of the week to support census (and other survey) fieldwork processes.

The data to be used were:

- A dataset collected by the University of Loughborough linking aggregated one minute power demand readings to a basic household occupancy and appliance ownership survey of 22 dwellings observed over two years (2008-2009);
- A similar energy power demand monitoring dataset held by the University of Southampton which derives from an ongoing study of around 180 households (UoS-E).

In both cases 30 minute summaries of this data were used to replicate the level of granularity that will initially be available from the proposed national electricity smart meter roll-out.

As was noted a substantial proportion of the resource allocated to this project was expended in processing, cleaning, checking and aggregating the data from it's source state to half-hour periods (see also (Caroll et al., 2013)). This required the use of the University of Southampton's Iridis4 high performance computer as the files containing the one second (UoS-E) readings in particular were too large to be manipulated on a standard personal computer until they had been aggregated to at least 1 minute summaries.

The descriptive analyses reported in Sections 4.3 and 5.4 both suggested that there were differences in household temporal power demand profiles that may be more homogenous within some household groups. Despite the obvious degree of within-group variation there was therefore a basis for exploring methods to predict household types based on patterns of power demand and also to estimate times of day when specific households might best be contacted.

The models predicting power demand based on household characteristics showed that of the variables used, the number of occupants was a major driver of power demand at all times of day. However the presence of children or of residents aged 65+ also played a role at specific times of day. Modelling these correlations in reverse to attempt to predict household occupancy and type from the power demand data proved more difficult but it was encouraging to note that the most effective approaches used indicators that captured some aspect of the temporal profile of use in contrast to overall power demand or power demand at a particular time of day. This confirms that there may be value in pursuing alternative approaches to the analysis of such power demand profiles as a way to develop predictive household classification methods. Such approaches might include the use of time-series analytic techniques to analyse differences in profile shapes and to study the cyclical power demand behaviour of different kinds of households; the analysis of rates of change of power demand at particular times of day and the analysis of higher order variation such as weekly, monthly or seasonal patterns.

An alternative approach that could be considered is that of optimal matching and subsequent cluster analysis (Lesnard & Kan, 2011). However it must be noted that clusters imputed in this way may not match 'census-like' typologies since the methods are likely to reveal more about the nature and extent of 'social practices' across the population than they are to identify specific household socio-demographic types. On the other hand were the classification of households by power demand

practices to be considered of value in the production of novel indicators then this approach would merit attention.

Given the level of variation indicated by the descriptive analysis and the consequential lack of precision in many of the regression coefficient estimates, it may be valuable to pursue access to larger sample datasets in future work. However this will only add value if the increased sample size acts to reduce relative variation between household types and it is not yet clear if this will be the case. Currently it seems equally possible that there may be as much variation in power demand (and power demand habits) within specific household socio-demographic groups as between them as fine-grained analysis of the power demand of water has shown (Shove & Medd, 2005).

Finally, it also seems likely that analysis of finer-grained power demand data might reveal more about household habits and practices which in turn could either provide a basis for 'practice-based' classification (Pullinger, Anderson, Browne, & Medd, 2014) or may be found to correlate with *some* socio-demographic characteristics. On-going research under the ESRC funded 'Census2022' project[9] will investigate the value of using finer-grained temporal data for these purposes.

[9] http://www.energy.soton.ac.uk/tag/census2022/

7 Acknowledgements

This work was funded by the Office for National Statistics who also provided methodological support.

The Loughborough One Minute Resolution data was collected by Department of Electronic and Electrical Engineering. Centre for Renewable Energy Systems Technology at the University of Loughborough with the financial support of E.ON UK; Engineering and Physical Sciences Research Council Grant Numbers: EPSRC Supergen HDPS: GR/T28836/01; EPSRC Supergen HiDEF: EP/G031681/1; EPSRC Transition Pathways to a Low Carbon Economy: EP/F022832/1.

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

8 References

Anderson, B., Vernitski, A., & Hunter, D. (2012). Practice Hunting with British Telephone Call Records. In *Paper presented at CRESI Research Seminar, February 2012.* Retrieved from http://www.slideshare.net/ben_anderson/practice-hunting-with-british-telephone-call-records

Armel, K. C., Gupta, A., Shrimali, G., & Albert, A. (2013). Is disaggregation the holy grail of energy efficiency? The case of electricity. *Energy Policy*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0301421512007446

Caroll, P., Dunne, J., Hanley, M., & Murphy, T. (2013). Exploration of electricity usage data from smart meters to investigate household composition. In *Conference of European Statisticians*.

Claxton, R., Reades, J., & Anderson, B. (2012). *On the value of Digital Traces for commercial strategy and public policy: Telecommunications data as a case study*. Geneva. Retrieved from http://eprints.soton.ac.uk/350352/

DECC. (2013). *Energy Efficiency Strategy 2013*. London.

Dugmore, K., Furness, P., Leventhal, B., & Moy, C. (2011). Information collected by commercial companies: What might be of value to ONS? *International Journal of Market Research*, *53*(5), 619–650. Retrieved from https://www.mrs.org.uk/ijmr_article/article/95197

Hamouz, M. (2012). Disaggregation: Abilities and Limitations. In *Smart Demand (iHeat 2012)*. Cambridge.

Lesnard, L. (2004). Schedules as sequences: a new method to analyze the use of time based on collective rhythm with an application to the work arrangements of French dual-. *Electronic International Journal of Time Use …*. Retrieved from http://core.kmi.open.ac.uk/download/pdf/6231732.pdf#page=67

Lesnard, L., & Kan, M. (2011). Investigating scheduling of work: a two-stage optimal matching analysis of workdays and workweeks. *Journal of the Royal Statistical Society: …*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2010.00670.x/full

McKenna, E., Richardson, I., & Thomson, M. (2012). Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, *41*, 807–814. doi:10.1016/j.enpol.2011.11.049

Pullinger, M., Anderson, B., Browne, A., & Medd, W. (2014). New directions in understanding household water demand: a practices perspective. *Journal of Water Supply: Research and Technology - AQUA*. Retrieved from http://eprints.soton.ac.uk/355502/

Richardson, I., & Thomson, M. (2010). One-Minute Resolution Domestic Electricity Use Data, 2008-2009. Colchester, UK: UK Data Archive [distributor]. doi:http://dx.doi.org/10.5255/UKDA-SN-6583-1

Richardson, I., Thomson, M., Infield, D., & Clifford, C. (2010). Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, *42*(10), 1878–1887. doi:http://dx.doi.org/10.1016/j.enbuild.2010.05.023

Shove, E., & Medd, W. (2005). *The Sociology of Water Use*. London: UK Water Industry Research Limited.

Zimmerman, J.-P., Evans, M., Griggs, J., King, N., Harding, L., Roberts, P., & Evans, C. (2012). *Household Electricity Survey: A study of domestic electrical product usage*. Milton Keynes.

Annex 1 Statistical Annex

Annex 1.1 Aggregated Loughborough 1 minute data: Power data distributions before and after transformation

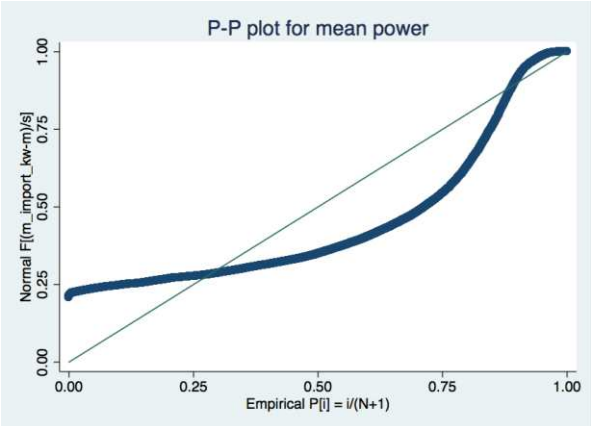

Figure 31: Distribution of aggregated daytime weekday Loughborough 1 minute level data before log transformation (P-P plot)

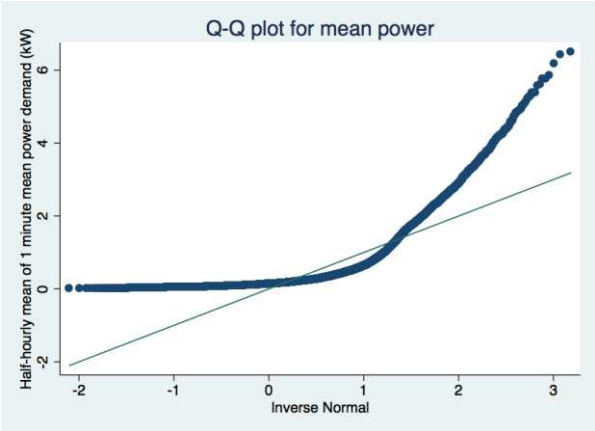

Figure 32: Distribution of aggregated daytime weekday Loughborough 1 minute level data before log transformation (Q-Q plot)

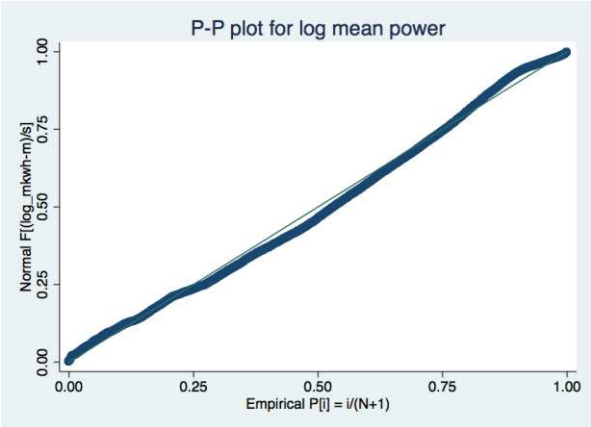

Figure 33: Distribution of aggregated daytime weekday Loughborough 1 minute level data after log transformation (P-P plot)

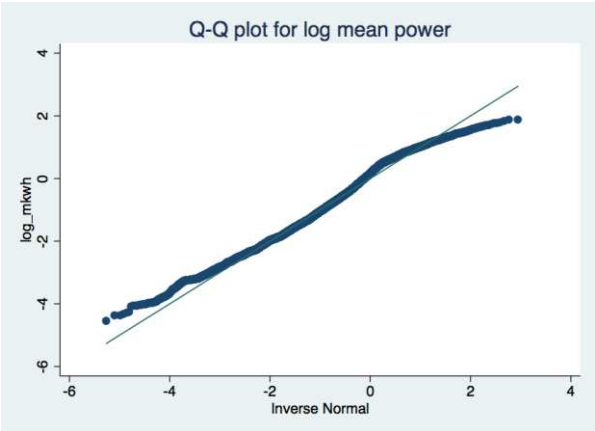

Figure 34: Distribution of aggregated daytime weekday Loughborough 1 minute level data after log transformation (Q-Q plot)

Annex 1.2 Aggregated UoS-E 1 second data: Power data distributions before and after transformation

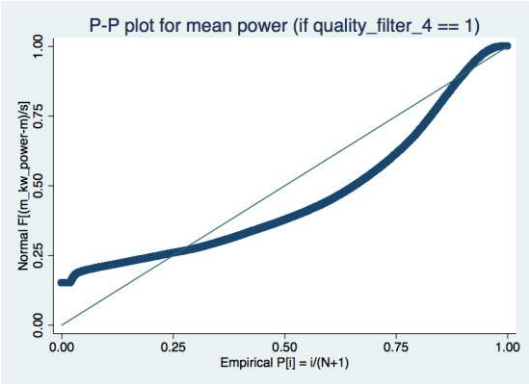

Figure 35: Distribution of aggregated daytime weekday UoS-E 1 second level data before log transformation (P-P plot)

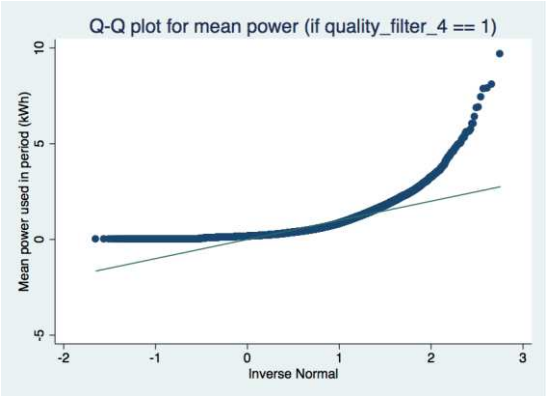

Figure 36: Distribution of aggregated daytime weekday UoS-E 1 second level data before log transformation (Q-Q plot)

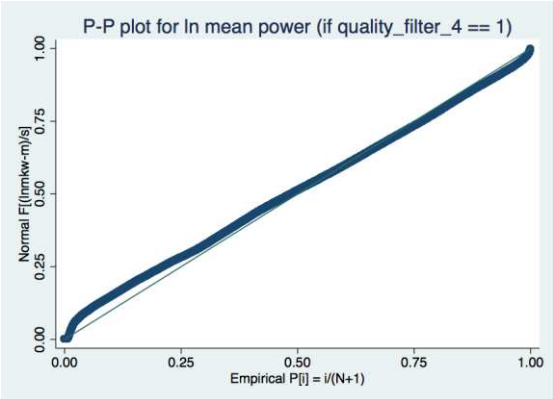

Figure 37: Distribution of aggregated daytime weekday UoS-E 1 second level data after log transformation (P-P plot)

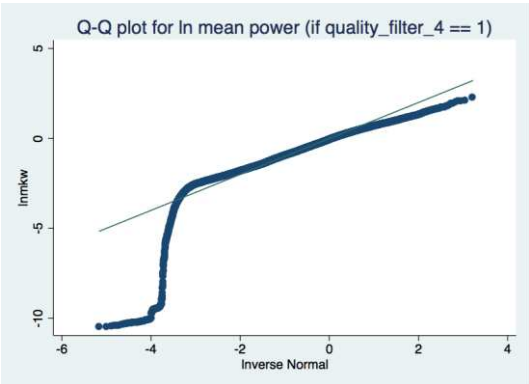

Figure 38: Distribution of aggregated daytime weekday UoS-E 1 second level data after log transformation (Q-Q plot)

Annex 1.3 Model diagnostics: Loughborough aggregated 1 minute data models

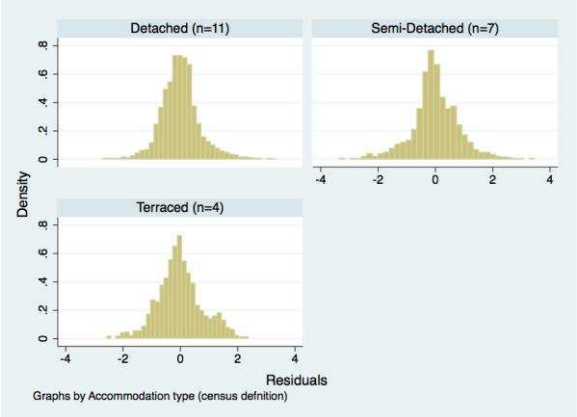

Figure 39: Distribution of residuals by accommodation type (Model 1, all half hours)

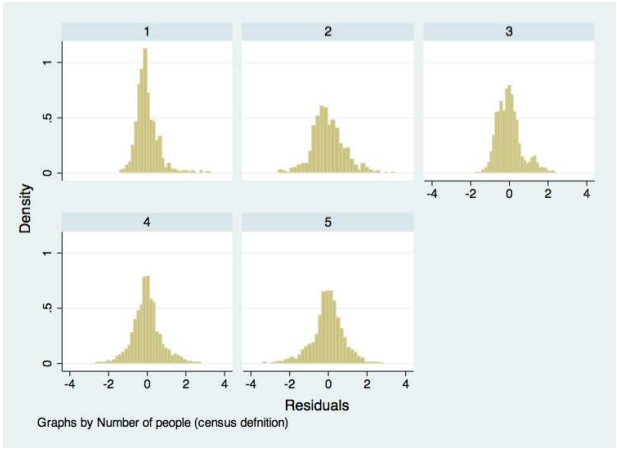

Figure 40: Distribution of residuals by number of occupants (Model 1, all half hours)

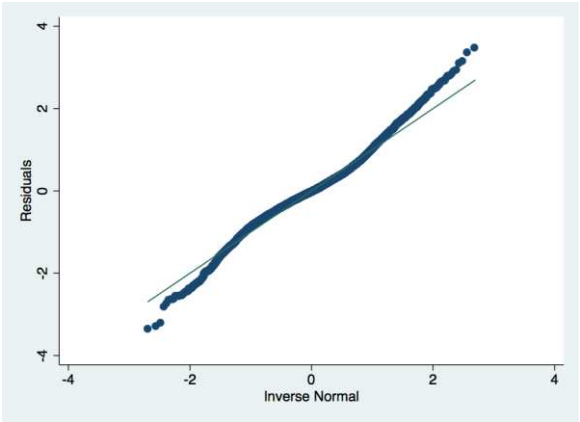

Figure 41: Q-Q plot of residuals (Model 1, all half hours)

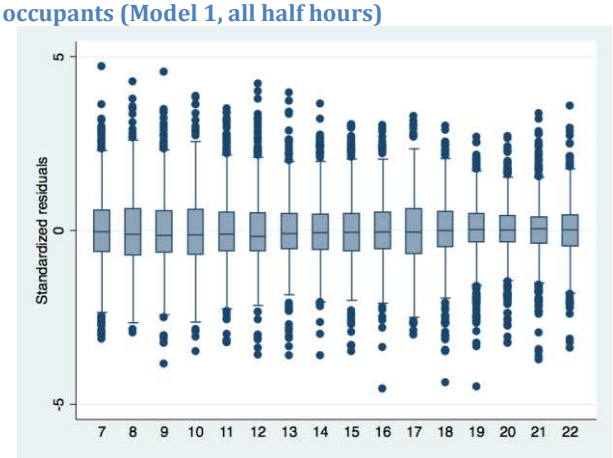

Figure 42: Plot of standardized residuals by hour (Model 1, all half hours)

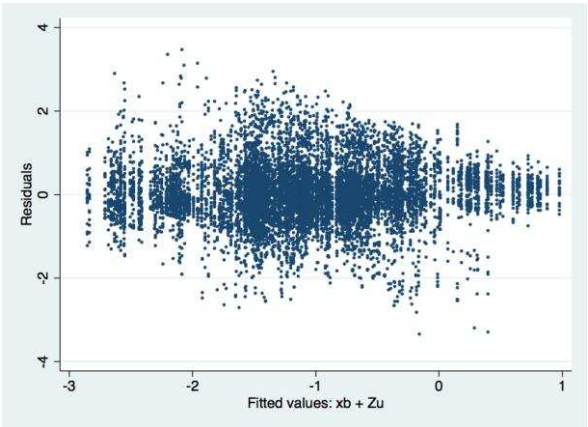

Figure 43: Plot of residuals against fitted values (Model 1, all half hours)

Annex 1.4 Model diagnostics: UoS-E aggregated 1 second data models

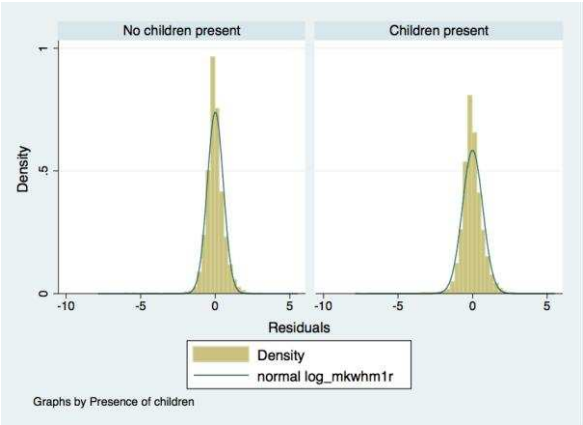

Figure 44: Distribution of residuals by presence of children (Model 1, all half hours)

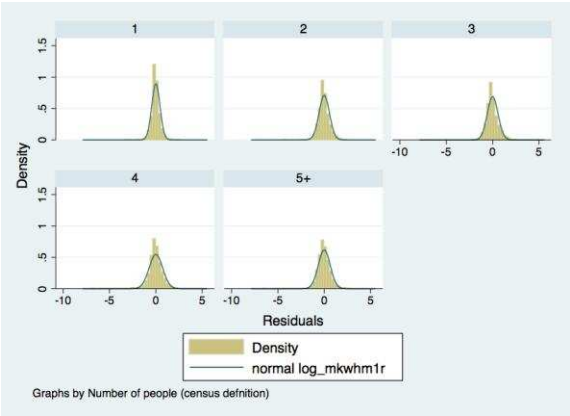

Figure 45: Distribution of residuals by number of occupants (Model 1, all half hours)

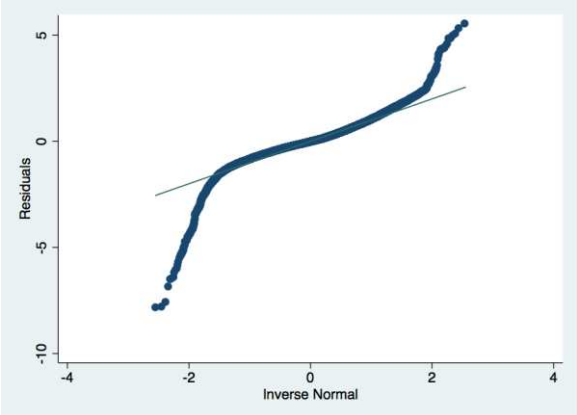

Figure 46: Q-Q plot of residuals (Model 1, all half hours)

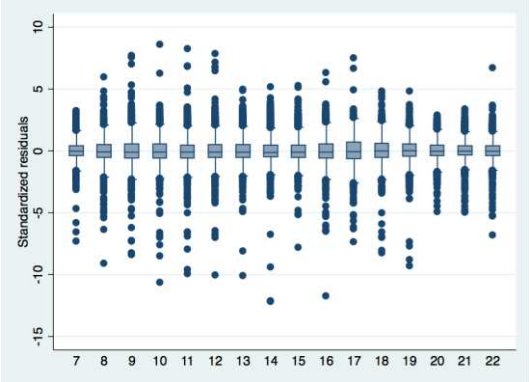

Figure 47: Plot of standardized residuals by hour (Model 1, all half hours)

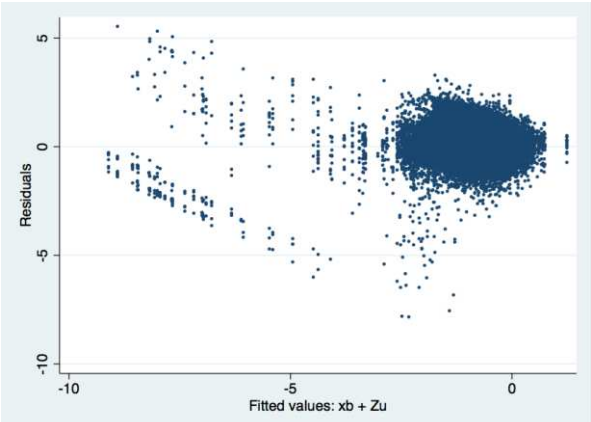

Figure 48: Plot of residuals against fitted values (Model 1, all half hours)