Office for National Statistics

ONS Methodology

Working Paper Series No x

Modelling sample data from Irish smart-type electricity meter trials~~s~~ to assess potential within official statistics

Susan Williams and [redacted]

July 2015

Modelling Irish sample data from smart-type electricity meter~~s~~ trials to assess potential within official statistics

Key Highlights

It is a ~~UK~~ GB wide policy to roll out electricity and gas smart meters to all households by 2020~~,~~ ~~commencing in 2016~~. The meters are capable of recording and storing detailed energy consumption ~~at high frequency~~data. The smart metering Data Communications Company (DCC) will put in place communications across Great Britain to send and receive information from smart meters to energy suppliers, energy network operators and energy service companies ~~before transmitting some information to a central body which will manage the data - essentially for billing purposes but also for other strictly regulated uses.~~ Consumers will have a choice about how their energy consumption data is used, apart from where it is required for billing and other activities that energy companies are legally required to undertake.

Comment [redacted]**:** Although the main rollout starts in 2016, some suppliers are rolling out meters now

Comment [redacted] The DCC is a data processer rather than a data controller.

If ~~access to the~~ suitable access to the data is ~~allowed~~ available in the future, this data may bring significant benefits within the production of official statistics as it is theorised that the patterns in such fine grained energy consumption data may reveal important intelligence about ~~a~~ households occupants' at an aggregate level, such as the average -numbers of people living ~~there~~in small areas, the ~~presence~~ numbers of of retired people or ~~even if a home is occupied at all~~the numbers of occupied homes. This type of information is required for well-informed planning decisions ranging from national policies down to the allocation of local authority services as well as helping to further academic research in various social themes. Currently collected through large scale official surveys which are becoming increasingly expensive to administer, if such features can be modelled into estimates for small areas using smart meter data then they may have a use in helping to validate or even replace official estimates. This will make substantial savings in costs and the burden placed on respondents to take part in surveys. The frequency of such estimates might also be increased allowing decision makers to make more timely and optimal interventions.

This report covers ONS research into the potential of using consumption data from Irish smart-type electricity meter~~s~~ trials to improve official statistics. The research has used samples of data made available from energy trials and has focussed on using this data to develop models which might lead to deriving the likelihood of a home's occupancy.

Key highlights:

Privacy and ethics

~~There is strong concern around which organisations may have access to the data produced by smart meters post 2020.~~ The granularity of the data, ~~possibly~~ as frequent as half hourly intervals ~~and~~ for individual meters, might theoretically be used to identify specific household characteristics or real-time occupancy. The Government has therefore put in place a robust framework to ensure that consumers retain choices over who is able to access their data.

Internationally, there are worries that the data may be mined for marketing and advertising or even price discrimination.

The ONS has engaged with privacy groups to discuss this research and they have given their approval.

Comment [redacted]: We notice from the meeting note you sent us recently that you are not in touch with organisations exclusively representing the interests of domestic consumers. You may find it helpful to engage with them. As a minimum you might consider engaging with Citizens Advice. Other groups worth considering are Privacy International and Big Brother Watch, although they have limited resources for engagement.

Legislation and access

In the UK, consumers will be able to opt out of having a smart meter installed, although it is thought that the uptake will still be high due to the consumer benefits in having one, such as an end to estimated billing and reduced energy use.

Comment [redacted]: This is a key benefit of smart metering alongside energy consumption reduction

~~All smart meters will be required to provide~~ Energy suppliers will be able to access monthly data ~~to their energy suppliers~~ for billing and other regulated purposes. Consumers will have a choice about whether their suppliers are able to access more granular data. All third parties, such as energy service companies, will have to have explicit consent from the consumer to access their consumption data. ~~then have various choices of opting in to providing half hourly data to specific organisations and for specific use.~~

It is unclear whether ONS will have suitable access to this data in future. If access is permitted then it is likely that all data ~~will be~~would be ~~anonymous~~ anonymised so as to restrict linking energy profiles to individual addresses.

Comment [redacted]: Access to data that is not anonymised would require that ONS obtain the consent of each household.

Data for research

There are a number of research datasets available containing high frequency energy readings from smart-type meters.

The ONS has sourced data from consumer behaviour trials of smart-type meters conducted by the Commission for Energy Regulation in Ireland and held in the Irish Social Science Data Archive. This data contains 30 minute frequency electricity energy usage data on 6,445 homes and businesses during 2009-2010. A pre- and post-trial demographic survey was also conducted so it is possible to identify some features of the home and the household inhabitants.

Field Code Changed

Applications within official statistics

There is a growing interest in the role smart meter data may have within official statistics across international statistical organisations as smart meter electricity energy usage data allows investigation at low levels of geography and high levels of timeliness. Additionally within England, when roll-out is finalised in 2020, this data may represent an almost complete coverage of homes.

Previous studies have shown that factors such as household type, the number of occupants and their geo-demographic or socio-economic status are linked to household electricity consumption.

Benefit of occupancy estimates

Low and constant electricity consumption over a period might indicate that a home is unoccupied. This might have application to a single day or a longer period if wanting to identify long-term vacant properties. Feasibly, small area estimates on the likelihood of homes being occupied on certain days and at certain times might be

	achieved which could benefit fieldwork processes in national surveys. Estimates of the number of households unoccupied on Census night can help to verify Census counts and this was chosen as the focus of the research.
Development of methods to identify unoccupied households	Eight methods were investigated to try to automatically identify households unoccupied for a whole day. These methods use different combinations of features such as the total, average or variance of energy consumption for a given day (defined as the 24 half hour periods from midnight to midnight). The methods produce statistics which are then compared against a suitable threshold value to classify days as occupied or unoccupied. Some methods explicitly use a baseline measure of an unoccupied household, such as night time energy consumption. This is taken further in some methods by looking at the energy consumption for a short period before the day being assessed to see how the energy consumption on that day compares with the ‘usual’ pattern of energy consumption for a meter; greatly reduced energy usage to the norm may indicate an unoccupied day.
Challenges of research	To confirm if a day was truly unoccupied it was necessary to conduct a visual assessment of the energy profile. This was labour intensive and to make the work manageable the research was restricted to 10 meters (representing 5,360 days). Although the pattern for an unoccupied day was agreed to be a low and constant energy profile over the 24 hours, slight variations to this meant that it is not always clear whether a household is occupied or unoccupied.
Performance of the methods	To visualise the performance of each classification method a two-way contingency table, also known as a confusion matrix, was set up to contrast the actual and predicted counts of occupied and unoccupied days. The statistical measures of sensitivity and specificity were used to compare methods. Two of the methods performed well, although it is highlighted that all the methods may be improved. A further enhancement to the classification may be to combine methods, so that an unoccupied day is indicated where multiple methods confirm it.
Handling the data	During this research, capability around using big data tools was increased. This resulted in the writing of efficient code to enable the classification methods to be applied to all meters in the data. Further investigation would be needed into efficient ways of processing up to 20 million households as this would be the scale of true smart meter data post 2020.

Future research

It is proposed that machine learning models may be used to identify numbers of unoccupied households rather than explicit methods as conducted in this research. As a simple case, this may involve the use of logistic regression or cluster analysis, although a first step will involve the selection of suitable parameters.

Suggestions for extending this research to identify numbers of longer-term unoccupied households are also given.

1. Introduction

The amount of data that is generally available is growing exponentially and the speed at which it is made available is faster than ever. The variety of data that is available for analysis has increased and is available in many formats including audio, video, from computer logs, purchase transactions, sensors, social networking sites as well as traditional modes. These changes have led to the big data phenomena – large, often unstructured datasets that are available potentially in real time.

Like many other National Statistics Institutes (NSIs) the Office for National Statistics (ONS) recognises the importance of understanding the impact that big data may have on our statistical processes and outputs. So ONS established a 15 month Big Data Project to investigate the potential benefits alongside the challenges of using big data and associated technologies within official statistics. This completed at the end of March 2015. The key deliverable from this proof-of-concept project was an ONS strategy for big data. In taking forward this work ONS is upholding all relevant legal and ethical obligations.

This report covers preliminary research on the potential of smart meter data within official statistics. The data used has been sourced from trials of energy usage using smart-type meters, conducted in Ireland and made available for research through the Irish Social Data Archive.

The report starts with a brief background to smart meters, the ~~UK~~GB roll-out and the data that they record before highlighting the privacy and ethical questions surrounding ~~the~~ access to this data in the GB context. A review of previous studies using this data leads to the identification of possible benefits to ONS in different applications.

Comment [redacted]: Needs to be clearer that the research is based on Irish data, with results then inferred to the GB context.

The focus of the report then shifts to consider the data available from trials of energy usage before introducing the objective for ONS research as the development of methods to automatically identify unoccupied days. The data used in this research is then discussed before addressing the methodological challenges of modelling occupancy. There follows the approach taken to develop and test methods to determine whether households are occupied by examining their electricity consumption profiles. Some ideas for furthering the research are then given before moving to a conclusion.

Comment [redacted]: For post census validation? Missing justification for why targeting unoccupancy.

There is also a comprehensive appendix containing more detail:

Appendix A: Details of the smart metering system
Appendix B: Overview of previous research
Appendix C: Benefits of small area estimates on unoccupied homes within Official Statistics
Appendix D: Definition of an unoccupied dwelling
Appendix E: Initial quality assurance and processing of the data
Appendix F: Details of research into methods 3-8
Appendix G: Increasing the size of the data

2. Background to smart meters

A smart meter is an electronic device that records and stores consumption information of either electric, gas or water at frequent intervals. These data can be transmitted wirelessly to a central system for monitoring and billing purposes.

Comment [redacted]]: The GB rollout is wireless but smart meters can also use power line technology to transmit data.

The Third Package Directives require Member States to ensure that at least 80% of consumers have such intelligent electricity metering systems by 2020.The European Commission's Energy Efficiency Directive (EED 2012) is a common framework of measures for the promotion of energy efficiency within the EU. It supports the EU's 2020 headline target on a 20% reduction in energy consumption and contains a number of smart metering requirements, in particular on the provision of data to energy consumers by energy suppliers. ~~, and its provision~~~~[1]~~ ~~for the roll-out of smart meters requires member states to ensure that at least 80% of consumers have such intelligent electricity metering systems by 2020.~~

Comment [redacted]: The EED does not set out this requirement.

The UK government's Department of Energy and Climate Change (DECC) has one of the most comprehensive~~ambitious~~ roll-outs ~~policies~~ within the EU: to install~~put~~ electricity and gas smart meters in every home in ~~England~~ Great Britain and small business by 2020[2] ~~with roll-out starting in 2016.~~

For electricity, smart meters will record consumption data at a minimum specification of 30 minute intervals and will be capable of storing monthly data for the previous ~~13~~ 13 - 24 months depending on the meter type. The data will be stored on the meter and ~~transmitted at predefined intervals to a central~~accessed by energy suppliers, network operators and third parties, such as energy service companies, using a communications infrastructure overseen ~~provided~~ by ~~body called~~ the Data and Communication Company (DCC)~~ where data access will be permitted for certain specific functions as described in legislation~~. Energy suppliers are obliged upon request to provide domestic smart meters customers with 24 months of daily, weekly, monthly and annual consumption data (or the length of the supply contract if that's shorter) and three months of export data to customers with micro generation installed, where their smart meter is being used to record electricity exported to the national grid.

DECC have produced a leaflet to illustrate the main parts of the ~~UK~~ GB smart metering system and some of the detail from it is provided in Appendix A.

The evidence supporting the roll-out of smart energy meters in the ~~UK~~ GB is based on international and national research. This research highlights the advantages of using smart meters for various stakeholders. For example:

- Smart meter data will enable more accurate billing and energy companies will no longer need to visit homes to read meters.
- Consumers will be able to see how much energy ~~they have used at different times of the day and~~ they are using and how much it is costing them in near real time, helping them to understand their consumption adjust their energy consumption behaviour to reduce bills.

~~Policy makers and energy companies may analyse patterns in energy consumption to devise policies to influence demand so that it fits better with the increasingly unpredictable energy supply associated with the adoption of power generation using renewable energy sources.~~The introduction of smart meters will improve the ability to shift demand to match supply (demand side response) which may be cheaper than building generation capacity to meet future demand peaks

-

~~[1] This provision relates to another EU Directive on smart meter roll-out (2009) which required a full cost/benefit analysis be performed prior to commencing roll-out~~

[2] Wales, Scotland and Northern Ireland have similar policies

3. Privacy concerns and ethics around using smart meter data

In the UK and internationally, there are important privacy concerns around the use of smart meter data. ~~These concerns centre on the large scale collection of personal data which can track what members of a household do.~~ The European Data Protection Supervisor tracks privacy issues in Europe and is concerned that the data could be mined for marketing and advertising, or price discrimination.

Within the UK, and as a result of a government consultation into data access and privacy, DECC ~~outlined plans~~has established regulatory requirements to allow consumers to control how much of their data, beyond that required for billing and regulatory purposes, they supply to energy suppliers or third parties.

Privacy groups such as Consumer Focus suggest that an energy supplier may be able to use its customers' consumption data to give individual advice on the best tariff or to suggest ways to save energy. Other groups such as Privacy International want consumers to understand and give explicit consent for all uses of their data.

Comment [redacted]**:** Consumer Focus has now become part of Citizens Advice.

~~As data access legislation for smart meters is still being devised in the UK, it is currently unknown what level of detail will be available to ONS from this source once smart meters are rolled out across the country.~~ The current regulatory framework does not allow ONS access to smart meter data for the purposes discussed in the paper. However, it is conceivable that such access could be provided in future if suitable arrangements are put in place to protect consumer privacy. ONS has held meetings to discuss research using smart-type meter data, sourced from trials of energy usage, with the Government Digital Service Privacy and Consumer Advisory Group and the ONS Beyond 2011 Privacy Advisory Group. While these groups are content for ONS to conduct research on the utility of this data, more discussion would be required and a strong case made should ONS wish to access ~~true~~ smart meter data ~~within~~ for the production of official statistics.

Comment [redacted]**:** Can you give more details on this? We have had a CAG in DECC, but it is clearly not the same!

4. Research using energy data: Applications within Official Statistics

a. Applications

There is a growing interest in the role smart meter data may have within official statistics across international statistical organisations as smart meter electricity energy usage data allows investigation at low levels of geography and high levels of timeliness. Additionally within England, when roll-out is finalised in 2020, this data may represent an almost complete coverage of homes.

Previous studies, described in more detail in Appendix B, have shown that factors such as household type, the number of occupants and their geo-demographic or socio-economic status are linked to household electricity consumption.

In summary, the current applications of most interest for the production of official statistics are:

1. Energy usage and expenditure which is of key interest to policies concerning the management of energy demand/supply in the longer term. For example, the frequency of smart meter data facilitates analysis of energy demand with weather effects such as temperature and rainfall. If relationships can be identified then weather data may provide a

useful indicator to identifying energy usage trends at a national/regional level, reducing the need to source smart meter data directly or to collect energy spending through surveys.

2. Occupancy status of homes: Low and constant electricity use over a period might indicate that a home is unoccupied. This might have application to a single day or a longer period if wanting to identify long-term vacant properties. Feasibly, small area estimates on the likelihood of homes being occupied on certain days and at certain times might be achieved which could benefit fieldwork processes in national surveys.
3. Household size or structure: It is hypothesised that profiles of energy use during the day might vary by household size or the composition of a household's inhabitants. Small area estimates might again be developed.

b. Data available for research

Over recent years, there have been various trials of smart-type meter devices to investigate energy usage and the data collected from some of them has been made available for research.

The University of Southampton was commissioned by ONS to conduct a small research project to investigate the potential of using smart-type meter data to identify household size/structure and the likelihood of occupancy during the day. The findings from this research have helped to inform ONS internal work. More detail can again be found in Appendix B.

It is highlighted that the ultimate aim for all ONS research is to develop methods to produce small area estimates for use within either statistical outputs or operational processes such as fieldwork. However, as a first step, it is necessary to work at an individual (yet still anonymous) level to understand patterns of energy usage.

4. ONS research

a. Research objectives

It is theorised that an unoccupied household would have a low and fairly constant pattern of electricity consumption across a time period. The focus of ONS research is the development of algorithms to automatically identify when such patterns of electricity consumption exist, possibly implying that a household may be unoccupied. By grouping households in an area, it may be possible to develop estimates for that area of the likelihood that households are unoccupied.

Estimates of the proportion of unoccupied households in an area on Census night would help to verify counts in the Census. Furthermore, knowing which areas have high or low proportions of long-term unoccupied homes would help indicate which areas have high proportions of addresses that are vacant, second addresses or holiday homes which in turn would help in the optimisation of fieldwork for census or surveys. More discussion of these benefits can be found in Appendix C.

Comment [redacted]: This para clearly sets ONS research objectives in the wider context but this is not always as clear in parts of the earlier narrative.

There is no single definition of occupancy or of an unoccupied household across government. Official statistics about unoccupied dwellings come from a range of sources, including the Census, council tax data and household surveys. Different definitions are used in each, as illustrated in Appendix D.

The ONS initial research has focussed on developing methods to identify whole days (midnight to midnight) which appear to be unoccupied. Secondary objectives are to investigate the benefits and drawbacks of using some big data techniques in the processing of the data.

The research could then be extended to examine households which were unoccupied for a longer period of time (weeks or months).

b. Data used

The data used for internal research by ONS came from the 2009 and 2010 Electricity Customer Behaviour trials of smart meter roll-out conducted by the Commission for Energy Regulation (CER) in Ireland.

This trial of smart-type meters ran for a six month benchmark period and a one year test period with 6,445 "opt-in" participants (domestic and non-domestic customers). Half hourly electricity usage was recorded during the benchmark and test periods for each household. During the test period, different time of use tariff structures were used to see how they affected customer behaviour. Pre-and post-trial surveys were conducted and included a range of questions about household demographics and the home itself.

The first observation for all participants is 14th July 2009 and the last observation is 31st December 2010. For each day there are 48 readings – one every half hour starting at midnight of each day. In October when the clocks go back an hour there are 50 readings for that day and in March when the clocks go forward an hour there are 46 readings for that day. Appendix E includes detail of the initial quality assurance and processing of the data by ONS into a more usable format.

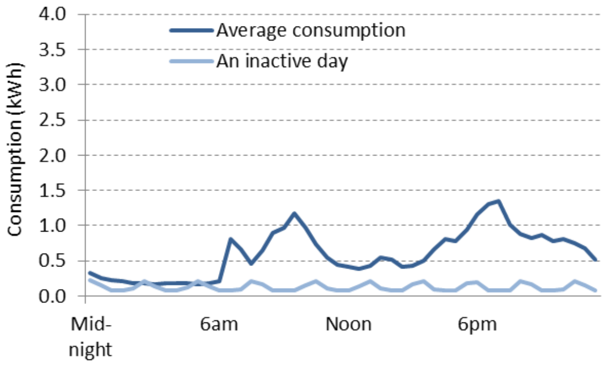
c. Exploration of the data

Electricity consumption is made up of both background loads, such as heaters, fridges and freezers which are driven by automated controllers and do not imply occupancy, as well as physical interactions such as flipping a switch which do imply occupancy. Timed appliances, thermostats, and lights left on may all create confusion as to whether a house is truly occupied when examining consumption alone. Further, some physical interactions such as turning a kettle on last only two to three minutes, making it difficult to distinguish this spike when consumption over 30 minutes is available in the dataset. All of these factors make it challenging to ascertain active occupancy from this data. In the future, determining occupied households will become harder as more and more appliances will be able to be turned on or off wirelessly via smart phones, tablets or other smart devices.

Figure 1~~Figure 1~~ shows the mean daily electricity consumption pattern during the 18 month trial period, as well as the consumption pattern for what is thought to be an unoccupied day for a sampled household.

Formatted: Font: Not Bold, Not Italic

Figure 1: Half hourly consumption for a sampled meter



The mean daily consumption over the 18 month trial period for this household is typical in that both a morning peak and an evening peak can be observed with a dip in the middle when household occupants may be temporarily absent from the home. The unoccupied day has a regular cyclical pattern of electricity consumption, typical of appliances driven by automated controllers such as a fridge or freezer. This is the pattern that has been most often observed during this research for days considered to be unoccupied, although as noted above, it is not known absolutely whether a household is truly occupied or not.

Figure 2: Half hourly consumption on one day for a sampled meter

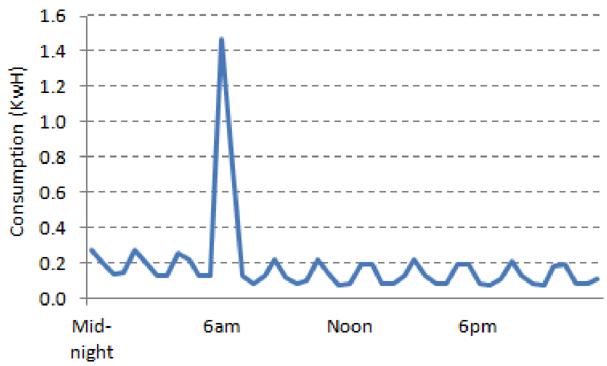


Figure 2~~Figure 2~~ illustrates an example where it is unclear whether the household was occupied at 6am or not: The household may be unoccupied with the spike in consumption being caused by a timed appliance, or the household may have been occupied in the morning around 6am but left unoccupied for the rest of the day. This demonstrates the difficulty in determining whether a household was occupied or not using only the half hourly electricity consumption data.

Formatted: Font: Not Bold, Not Italic

d. Approach

Eight classification methods were developed for determining whole days when a household might be unoccupied. These methods use different combinations of features such as the total, average or variance of energy consumption for a given day (defined as the 24 half hour periods from midnight to midnight). The methods produce statistics which are then compared against a suitable threshold value to classify days as occupied or unoccupied.

Some methods explicitly assume that night time energy consumption is indicative of an unoccupied household, a practice believed to be commonplace in energy demand analysis. By setting such a baseline it is possible to compare day time to night time consumption which might plausibly suggest occupancy if substantially roughly equal. This is taken further in some methods by looking at the energy consumption for a short period before the day being assessed to see how the energy consumption on that day compares with the 'usual' pattern of energy consumption for a meter; greatly reduced energy usage to the norm may indicate an unoccupied day.

In summary, the following list gives the high level description of each method to use half hourly electricity consumption to determine if a day is unoccupied. More detail on the first two methods and the thresholds relating to them follow in this section whilst the remaining methods are discussed in Appendix G.

- Method 1: Variance in energy consumption over 24 hours is low
- Method 2: Average day time consumption is similar to average night time consumption
- Method 3: Total energy consumption for a day is less than the 5th percentile of the daily consumption over the previous 3 months
- Method 4: Day time average is below average of previous 3 months' maximum night time consumption
- Method 5: Day time variance is similar to night time variance
- Method 6: Day time average is below night time average plus 1 Standard Deviation (using log consumption)
- Method 7: Range (maximum minus minimum) of day time consumption is similar to range of night time consumption
- Method 8: Inter-quartile range (IQR) of day time consumption is similar to IQR of night time consumption

As there is no information to confirm if a household was occupied on any specific day, the half hourly consumption data needed to be examined by eye and a subjective assessment on occupancy made for all days. Each meter in the data contains 536 days which led to a very time consuming manual checking process and to keep the work manageable a decision was made to perform initial research on a small sample of only 10 meters.

Most generally, the criteria for assessing a day as being unoccupied required that the electricity consumption on that day to be fairly flat (with up to one spike in energy use for one half hour period during the day) as in Figure 1~~Figure 1~~ and Figure 2~~Figure 2~~.

Formatted: Font: Not Bold, Not Italic

Formatted: Font: Not Bold, Not Italic

To visualise the performance of each classification method a two-way contingency table, also known as a confusion matrix, is set up to contrast the actual and predicted counts of occupied and unoccupied days.

Table 1~~Table 1~~ shows the general form of a confusion matrix. The number of days that are unoccupied and correctly classed by the method is denoted by the letter a; b represents the number

Formatted: Font: Not Italic

of occupied days classed as unoccupied; c is the number of unoccupied days classed as occupied and d is the number of occupied days correctly classed.

Table 1: Example of a confusion matrix

		Actual (by eye)	
		Unoccupied	Occupied
Classification (by method)	Unoccupied	a	b
	Occupied	c	d

A number of different statistical measures can be obtained from the table to test for the performance of a classification method. The three of most interest are:

Accuracy = $\frac{a+d}{a+b+c+d}$

Sensitivity (true positive rate) = $\frac{a}{a+c}$

Specificity (true negative rate) = $\frac{d}{b+d}$

As the proportion of unoccupied days in the data is very small, the classic accuracy measure is not advised. For example, if there were 5% unoccupied days overall and a method identified all days as occupied it would be 95% accurate even though it identified no unoccupied days at all.

Therefore it is more suitable to use the measures of sensitivity and specificity for assessing a method's performance. Sensitivity is also known as the true positive rate and represents the proportion of unoccupied days which were classed correctly by the method. Specificity is also known as the true negative rate and represents the proportion of occupied days which were classed as such by the method.

By using these two measures it is possible to compare the different classification methods, the better ones being those which classify unoccupied days with high sensitivity and high specificity.

e. Illustration of the methods to classify a day as unoccupied

To illustrate the research, two of the best methods are now presented in more detail. The remainder of the methods are to be found in Appendix G.

Method 1: Variance in energy consumption over 24 hours is low

It might be expected that an unoccupied household would have a low variance of electricity consumption. For the ten meters in the sample, the variance for each day was calculated and compared against several different thresholds with a variance of 0.01 found to be optimal in determining occupancy.

Specifically, method 1 states that a household is unoccupied on a given day (24 hours) if:

The variance of electricity consumption for the day < 0.01

Increasing the threshold made the method include too many days which visually appeared to be occupied, while decreasing the threshold excluded too many days which appeared to be unoccupied.

A confusion matrix is a convenient way to illustrate the accuracy of the classification. Table 2~~Table 2~~ shows the confusion matrix for method 1 and reveals that of the 192 days visually classed as unoccupied, method 1 successfully identified 189 of them giving a sensitivity of 98%. Correspondingly, the method correctly identified 5,144 days out of 5,168 as occupied giving a specificity of almost 100%.

Formatted: Font: Not Bold, Not Italic

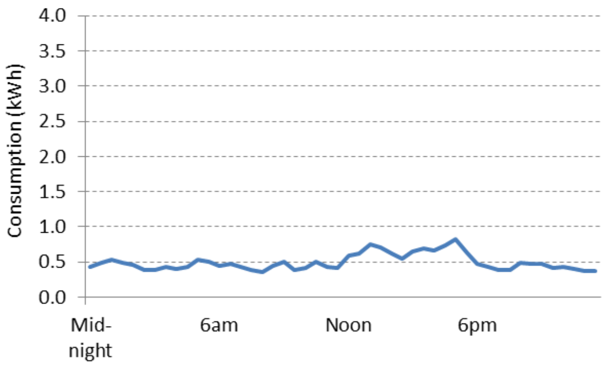
Table 2: Confusion matrix for method 1

Method 1	Examined by eye	
	Unoccupied	Occupied
Unoccupied	189	24
Occupied	3	5,144

Figure 3~~Figure 3~~ gives an example of a day where method 1 classes a household as occupied but the consumption profile visually suggests it to be unoccupied. In this particular case, the variance for this day is just above the chosen threshold of 0.01 so the method does not select it as unoccupied. In theory, this method may not work so well where a household has a more variable background energy use perhaps due to multiple fridges or freezers.

Formatted: Font: Not Bold, Not Italic

Figure 3: Half hourly consumption for a day identified as occupied by Method 1



Method 2: Average day time consumption is similar to average night time consumption

The average day time consumption would be expected to be similar to the average night time consumption in an unoccupied household whereas in an occupied household the day time consumption might be significantly greater than the night time consumption.

This assumption was first tested by forming a ratio of average day time (1am to 5am) consumption to average night time (5am to midnight) consumption and identifying a threshold under which a day may be classed as unoccupied. This form of the method was found to be sensitive to situations where high night time electricity consumption is seen, and not the low consumption low variance typically expected. Therefore the average night time consumption over the previous week was considered instead of the night time consumption on the day being assessed.

Comment [redacted]**:** Should this be pm?

Several thresholds were tested, and a value of 1.1 found to be optimal. As in method 1, increasing this threshold resulted in the method identifying too many occupied days as unoccupied while decreasing the threshold resulted in too few unoccupied days being classed.

As a formula, method 2 states that a household is unoccupied on a given day if:

$$1.1 > \frac{Mean\ day\ time\ consumption\ for\ current\ day}{Mean\ night\ time\ consumption\ for\ previous\ 7\ nights}$$

The confusion matrix for method 2 shows that 173 of the 186 unoccupied days were correctly identified: a sensitivity of 93%. The corresponding specificity is 99% as 5,057 out of 5,114 occupied days are correctly classified. These results show that method 2 is not quite as accurate in its classification as method 1, yet they both produce a good classification.

Table 3: Confusion matrix for method 2

Method 2	Examined by eye	
	Unoccupied	Occupied
Unoccupied	173	57
Occupied	13	5,057

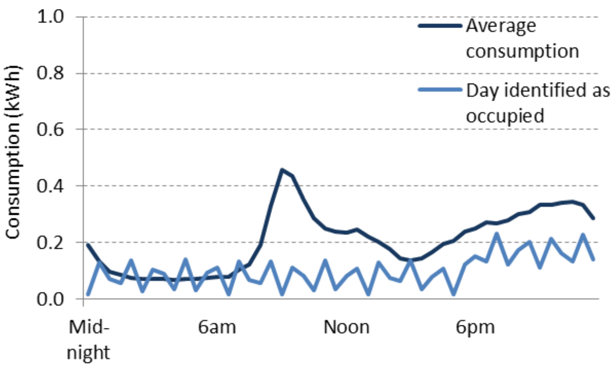
Investigation of method 2 shows that it tends to class days with single spikes as unoccupied whereas method 1 does not (see Figure 2~~Figure 2~~). Therefore method 2 could be used if wanting to define such a profile as an unoccupied day.

Formatted: Font: Not Bold, Not Italic

This method also identifies days as occupied if there is a step change in the underlying electricity load during the day as shown in Figure 4~~Figure 4~~ below where consumption after 6pm appears to be higher than before 6pm. The day under scrutiny was given a visual assessment of being unoccupied due to the low consumption and low variance across the day. Other methods also classify it as such.

Formatted: Font: Not Bold, Not Italic

Figure 4: Half hourly consumption for a day identified as occupied by Method 2



f. Summary of methods tested

Appendix F describes each of the remaining methods in more detail in a similar way to methods 1 and 2 above. For ease of comparison, the sensitivity and specificity of each method tested is given in Table 4~~Table 4~~ below.

Formatted: Font color: Auto

Table 4: Comparison of sensitivity and specificity for all methods

Method and description	Sensitivity (True positive) (percentage)	Specificity (True negative) (percentage)
Method 1 (Variance in energy consumption over 24 hours is low)	98	100
Method 2 (Average day time consumption is similar to average night time consumption)	93	99
Method 3 (Total energy consumption for a day is less than the 5th percentile of the daily consumption over the previous 3 months)	67	96
Method 4 (Day time average is below average of previous 3 months' maximum night time consumption)	99	93
Method 5 (Day time variance is similar to night time variance)	51	99
Method 6 (Day time average is below night time average plus 1 Standard Deviation)	99	86
Method 7 (Range of day time consumption is similar to range of night time consumption)	53	99
Method 8 (IQR of day time consumption is similar to IQR of night time consumption)	57	96

Table 4~~Table 4~~ highlights how methods 1 and 2 are superior to other methods.

Formatted: Font color: Auto

g. Analysis of the full dataset

Following the research on the sample of 10 meters each of the methods were run on all the 4,225 meters in the data to see how the methods would perform. As the full data contained 536 days it was infeasible to conduct a visual assessment for each day so the results focussed on the percentage of days each method identified as unoccupied.

For example, both method 1 and method 2 identified around 4% of days as being unoccupied in the 10 meter sample. In the larger data, it might be expected to see that method 1 and method 2 classing a similar percentage of days as unoccupied – with the same days being identified by each method.

Table 5~~Table 5~~ compares the percentage of days each method classed as unoccupied in the 10 meter sample and the full data representing all 4,225 meters:

Formatted: Font: Not Bold, Not Italic, Font color: Black

Table 5: Number of days identified as unoccupied by all methods in full and test datasets

Method	Number of days identified as unoccupied on full dataset	Percentage of days identified as unoccupied on full dataset	Percentage of days identified as unoccupied on ten test meters
1	179,626	8	4
2	210,390	9	4
3	128,649	7	7
4	449,489	24	10
5	138,628	6	3
6	620,574	27	17
7	131,802	6	3
8	212,414	9	6

Table 5~~Table 5~~ illustrates that, apart from in method 3, a higher proportion of days are identified as unoccupied in the full dataset compared with the ten test meters.

Formatted: Font: Not Bold, Not Italic

Both methods 1 and 2, which perform well in their classification, identify around the same percentage of days as unoccupied – 8% and 9% respectively. In the discussion on these methods, it was observed that method 2 tends to class days where there is a single peak of electricity use as unoccupied. If there are proportionately more days with this profile in the full data then this may explain why method 2 identifies more unoccupied days.

Method 3 is the exception, as it effectively sets the proportion unoccupied to around 5% for any data. In the discussion of this method in Appendix F it is proposed that a more adaptable threshold is needed to make performance better.

Methods 4 and 6 both identify around a quarter of the days as unoccupied which, referring to Table 4~~Table 4~~, is attributable to their lower specificity as this increases the number of occupied days misclassified as unoccupied.

However, it should be noted that all the methods illustrated be considered as only starting points for identifying unoccupied days as improvements to their design may be possible. For example, the formula might be enhanced and/or a more optimal threshold identified. Furthermore, there may be an improved classification for identifying unoccupied days when classed as such by more than one method.

5. Handling the data

The eight methods used on the sample of 10 meters were, with more efficient computer code, replicated on all 4,225 meters in the dataset. However if smart meters are to be rolled out to every household in the country by 2020, and this data made accessible for the production of official statistics, the infrastructure and big data technologies would need to be in place to be able to potentially analyse a dataset covering 20 million households.

ONS has set up innovation labs containing clusters of high specification computers to help facilitate research into new technologies and open source tools, new sources of public data and to develop associated skills. They are completely separate from the main ONS network and therefore provide a route for easily accessing open source tools without compromising ONS security. The innovation labs are a key enabler for the ONS Big Data Project since they allow the team to handle large and complex data sets and to test new big data technologies.

R is a free software package for statistical computing and graphics. It is widely used in the fields of statistics and big data, and for this reason ONS used it to process and analyse the smart-type meter data in the innovation lab. A limitation of R is that it can only address objects that fit in the available virtual memory space so it cannot cope with very large datasets. ONS has examined some R packages and other software which overcome this limitation and may have the potential to analyse larger datasets. The results of this research can be found in Appendix G.

6. Next steps and conclusion

a. Next steps: Using machine learning

Machine learning deals with the construction and study of algorithms that can learn from data. Such algorithms operate by building a model and using that to make predictions, rather than following only explicitly programmed instructions. Machine learning is commonly used in the field of big data to detect and classify patterns in large datasets.

While the smart-type meter data does not include information about whether a household is truly unoccupied or not, it might be safe to say that a household identified as unoccupied by a majority of the eight methods tested is very likely to be truly unoccupied. Information about the energy profile of such households could be used, as well as similar information about occupied households to build a machine learning algorithm which is better able to identify whole days when households are unoccupied.

b. Next steps: Long-term unoccupied households

The research on daily occupancy may be extended to determine long-term unoccupied households such as vacant properties and holiday homes. Such households may be identified by counting the number of consecutive days identified as unoccupied by any appropriate method.

Figure 5: Occupied days identified by method 2 over trial period for one sampled meter

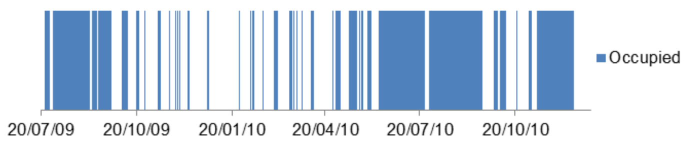


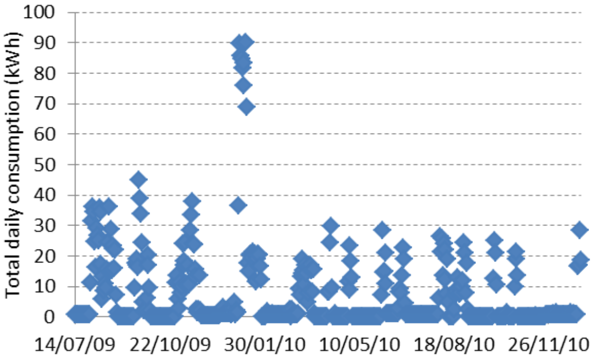
Figure 5~~Figure 5~~ shows whole days which are occupied (in blue) for one household using method 2. It shows that there are periods of time during winter 2009/10 that appear to be unoccupied by this household.

Formatted: Font: Not Bold, Not Italic

Figure 6~~Figure 6~~ shows the total daily electricity consumption for one meter. The sporadic peaks in consumption with low consumption in between suggest that the household may be a holiday home. The very high consumption starts on 31 December 2009 and continues until the start of February 2010.

Formatted: Font: Not Bold, Not Italic

Figure 6: Total daily consumption over trial period for a sampled meter



c. Conclusion

This report has introduced initial research into the potential of using the data from smart-type electricity meters to model household occupancy patterns. Eight methods were developed to automatically identify inactive days with promising results. By aggregating the results from multiple meters, small area estimates of the likelihood of occupancy could be formed. These in turn could be used to validate census returns and would make efficiency savings in a Census operation.

Comment[redacted]**:** Preferred wording around research objectives (as opposed to vacant / unoccupied)

The background to and infrastructure on the ~~UK~~ GB roll-out of smart meters, due to complete in 2020, is highlighted together with the privacy and ethical concerns around access to the fine grained data they will produce. The current regulatory framework would not allow ONS access to smart meter data without explicit consent from each household~~s~~. ~~It is emphasised that ONS does not have access to the half hourly consumption data due to be produced by smart meters. Discussions with DECC have indicated that the legislation for this type of data access would require consumers to explicitly have opted in to providing their data.~~ Engagement has already taken place between ONS and privacy groups to discuss this research using data from trials of energy usage. If this data were to be used in the Census operation or wider official statistics, then much greater engagement would be needed.

Previous academic research using smart-type meter data has shown that patterns in electricity consumption data could indicate household size and some household characteristics. The research by ONS used a similar dataset containing half hourly electricity consumption at 4,225 households in Ireland over an 18 month period. The approach to the research and the challenges encountered are discussed including the necessity of visually checking days identified by the methods. This placed a restriction on the number of meters that could be handled as well as revealing difficulties in deciding whether a household appears to be actively occupied based solely on its consumption pattern.

The performance of the eight methods was assessed using the statistical measures of sensitivity and specificity and two methods appear to perform particularly well although it is recognised that all methods may be improved.

The dataset on which these methods were tested was large and required manipulation using big data technologies. Given that the government intends to deploy smart meters in all 20 million households in Great Britain by 2020, manipulation of a dataset of such an increased size will require significant knowledge of a range of big data technologies.

A logical extension to this research is in the development of methods to determine the likelihood that a dwelling is long-term unoccupied such as being vacant or an infrequently used holiday home. Knowing which areas contain high proportions of such households could offer significant benefits to a Census or survey operation by ONS. Efficiency savings could be made by ensuring follow up responses are minimised in such areas. This was not done in 2011 but given the size of the Census operation, even a small increase in efficiency could have a large impact.

In summary, ~~while there are privacy concerns and access to smart meter data is uncertain,~~ it has been demonstrated that significant value can be extracted from such data. As a result, it is proposed that ONS continues to engage privacy groups in future research and that further engagement with DECC is required around future access to the smart meter data and arrangements that would be needed to ensure that consumer data privacy is suitably protected.

7. References

Anderson B and Newing A (2014a): Smart meters and smart censuses. Paper submitted to *Technology Analysis and Strategic Management.*

Anderson B and Newing A (2014b): Using energy metering data to support official statistics: A feasibility study. University of Southampton technical report submitted to Office for National Statistics.

Carroll P, Dunne J, Hanley M and Murphy T (2013): Exploration of electricity usage data from smart meters to investigate household composition. Paper presented at the Conference of European Statisticians Geneva, Switzerland.

Craig T, Polhill G, Dent I, Galan-Diaz C and Heslop S (2014): The North East Scotland Energy Monitoring Project: Exploring relationships between household occupants and energy usage. *Energy and Buildings* 75:493-503.

Druckman A and Jackson T (2008): Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model. *Energy Policy* 36(8):3177-3192.

Firth S, Lomas K, Wright A and Wall R (2008): Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings* 40(5):926-936.

Haben S, Ward J, Greetham D, Singleton C and Grindrod P (2014): Anew error measure for forecasts of household-level, high resolution electrical energy consumption. *International Journal of Forecasting* 30(2):246-256.

Newborough M and Augood P (1999): Demand side management opportunities for the UK domestic sector. *IEE Proceedings Generation, Transmission and Distribution* 146(3):283-293.

Onzo (2012): Onzo Application Detection Technology. London: Onzo Ltd.

Owen P (2012): Powering the nation: Household electricity habits revealed. London: Energy Saving Trust.

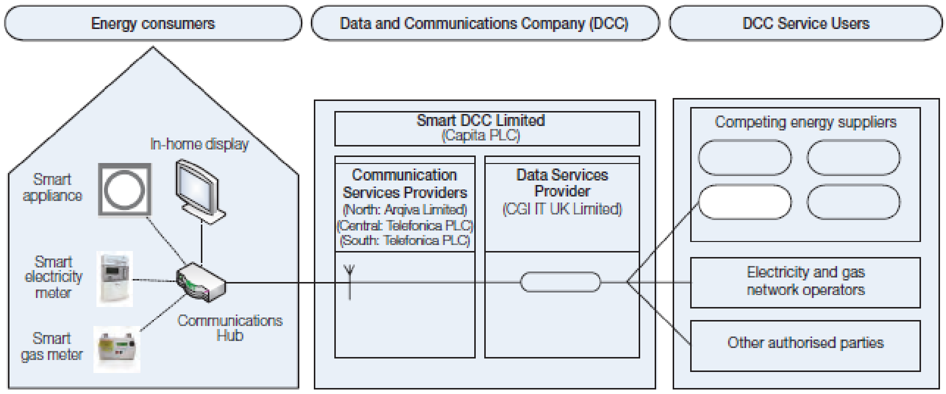
Wright A (2008): What is the relationship between built form and energy use in dwellings? *Energy Policy* 36(12):4544-4547.

Zimmerman J-P, Evans M, Griggs J, King N, Harding L, Roberts P and Evans C (2012): Household electricity survey: A study of domestic electrical product usage. Didcot: Intertek.

Appendix A: Details of the smart metering system

Figure 7~~Figure 7~~ below sourced from DECC[3] illustrates the main parts of the smart metering system. It highlights the role of the Data and Communication Company (DCC) as the entity which will provide the communication infrastructure giving access to ~~holds all the~~ smart meter data ~~and will manage access to the data for~~to other organisations known as DCC Service Users.

Figure 7: Diagram illustrating main parts of smart metering system



The in-home display will allow consumers to see what energy they are using and how much it is costing in near real time. The display can also show information about the amount of energy used in the past day, week, month and year.

For electricity, readings will have a minimum specification of recording consumption data at 30 minute intervals and of storing monthly data for the previous ~~13~~ 24 months. ~~The data will be transmitted at predefined intervals via the communications hub to the DCC. Data access will be permitted for certain specific functions as described in legislation~~[4]~~.~~

[3] See leaflet at https://www.gov.uk/government/publications/the-smart-metering-system-leaflet
[4] ~~Legislation still being devised~~

Comment [redacted] I think it would make more sense just to reproduce the whole content of the leaflet.

Formatted: Font: Not Bold, Not Italic

Appendix B: Overview of previous research

Studies have shown that factors such as household type, the number of occupants and their geo-demographic or socio-economic status are linked to household electricity consumption, Druckman and Jackson (2008); Firth et al (2008); Owen (2012); Wright (2008) and Newborough and Augood (1999). See Anderson and Newing (2014a) for a more detailed review of research using smart-type meters.

The early large scale studies were based on household electricity expenditure while more recent studies are based on the actual power consumption, see for example, Zimmerman et al (2012); Craig et al (2014); Caroll et al (2013) and Haben et al (2014). These studies (with the exception of Caroll et al, 2013) lack comprehensive detail about household characteristics. It is therefore difficult to develop methods which link observed high resolution patterns of electricity consumption to household characteristics.

Carroll et al (2013) uses six months of data from the Irish smart-type meter trial to identify household composition. They use various summary measures of power demand such as mean; maximum; standard deviation; morning maximum and load factor to predict membership of a family type. Carroll et al do not make use of the differences in power demand at different times of the day, but Richardson et al (2010) do.

Onzo (2012), in partnership with UK energy supplier SSE Energy Supply Ltd, has shown that it is feasible to make inferences about household occupancy using energy consumption data. In their analyses they use one second energy consumption data.

The following datasets have been used in ONS/University of Southampton research to date:

University of Loughborough
University of Loughborough's data from energy usage monitoring trials is archived by the UK Data Service for future research use. These data link consumption at one minute intervals to a basic household occupancy and appliance ownership survey. This dataset is derived from 22 dwellings observed over two years (2008-2009). Due to the small sample of properties, this dataset is only useful for testing out various big data technologies and methods to understand the benefits and drawbacks of different approaches to processing.

University of Southampton
The dataset held by the University of Southampton comprises a study of 180 households in the Solent region. The study ran between 2011 and 2014. This dataset consists of energy consumption data at one second intervals which can be linked to repeated six monthly survey data on household occupancy and other variables.

Electricity Customer Behaviour Trial in Ireland
This is data from consumer behaviour trials of smart-type meters conducted by the Commission for Energy Regulation in Ireland and held in the Irish Social Science Data Archive. This data contains 30 minute frequency electricity energy usage data on 6,445 homes and businesses during 2009-2010. A pre- and post-trial demographic survey was also conducted so it is possible to identify some features of the home and the household inhabitants.

Energy Demand Research Project
Data from this project comes from trials of smart-type meters conducted in Great Britain between 2007 and 2010. DECC published this data for research purposes in December 2014. These data represent around 15,000 homes but do not have associated demographic survey data.

University of Southampton research

Research was commissioned by ONS and carried out by the University of Southampton (Anderson and Newing, 2014b) to explore the feasibility of using smart-type meter data to predict:

- The specific household characteristics of:
    - Number of occupants
    - Presence of school aged children
    - Presence of single people or couples aged 65+
- Active occupancy (ie. at home and awake)

The 22 households in the University of Loughborough data were used in preliminary research to help devise methods of handling the larger University of Southampton data representing 180 households. With both sets of data, energy readings were first aggregated to 30 minute intervals so that they better reflected the frequency at which smart meter readings may be available in the future. Various models were then used to investigate the relationship between each household's typical profile of power consumption over 24 hours with features such as accommodation type, and the specific household characteristics.

The models suggested that there may be some potential for smart-type meter data to predict the specific household characteristics tested although the research would need to be continued with a much larger sample to develop more robust methods.

The researchers also provide a method to estimate the probability of active occupancy for any half hour period for each household. This method is untested and would require fieldwork to verify its accuracy.

Appendix C: Benefits of small area estimates on unoccupied homes within Official Statistics

There are several areas of ONS where there are clear benefits for knowing the proportion of unoccupied households.

Census

Knowing that households are occupied on Census night can help to verify counts in the Census. For example, if an estimate of 7% of households appears to be unoccupied in a specific area, then this can be used to validate the counts from the Census on the number of homes which are unoccupied.

Comment [redacted]**:** Reconsider tense as smart meter data will be used retrospectively, not in real-time.

Further, knowing whether a household is long-term unoccupied could help indicate which areas have high (or low) proportions of long-term vacant, second addresses or holiday homes. Knowing this might help in the following ways:

- Optimisation of fieldwork so that enumerators do not keep trying to follow up in areas with high proportions of such households.
- Quality assurance of Census data on ‘household spaces with no usual residents’.
- Producing outputs of vacant dwellings.
- Producing outputs of a seasonal population based on groups of holiday homes which are unoccupied for some parts of the year and occupied at other times.

The Census Transformation Programme has been examining opportunities to reuse existing data already held within government, often by linking different government datasets together. Further information about the programme is on the ONS website.

Survey field work

Improvements to ONS survey operations include:

- Being able to better identify eligible or ineligible households for interviewing. Ineligible households are vacant properties, holiday homes or small bed and breakfasts for example. Currently a lot of addresses targeted by interviewers are deemed to have “unknown eligibility”. Knowing that a high proportion of homes in an area are vacant or holiday homes would be useful as interviewers would not need to spend as long trying to contact households.
- Knowing the pattern of typical occupancy for a household or area which would allow area based targeting of households for interviewing. For example, one area may contain a lot of people who leave for work at 6am, whereas another may indicate higher electricity consumption during the day. Electricity consumption data could be combined with other data for small areas, such as that on unemployment or about the demographics in an area. This would enable a more cost effective calling pattern to be developed for an area.

Appendix D: Definition of an unoccupied dwelling

There is no single definition of occupancy or of an unoccupied household across government. Official statistics about unoccupied dwellings come from a range of sources, including the Census, council tax data and household surveys. Different definitions are used in each, as illustrated in Table 6~~Table 6~~.

Formatted: Font: Not Bold, Not Italic

Table 6: Definitions of unoccupied dwellings in different sources

Source	**Short, medium or long term unoccupied**	**Definition**
Census	Long-term unoccupied	In the 2011 Census, a vacant household space is an unoccupied space that does not have at least one usual resident and is not a second home or holiday accommodation
Council tax	Short-term unoccupied	Some properties which were unoccupied and substantially unfurnished were exempt from paying council tax for up to six months, followed by a discount of 0% to 50%. However the rules changed in April 2013 allowing councils to apply local discounts to such properties of between 0% and 100%
Council tax	Medium-term unoccupied	Some vacant properties undergoing major repair work or structural alteration were exempt from paying council tax for up to 12 months. However, again the rules changed in April 2013 allowing councils to apply local discounts to such properties of between 0% and 100%
Council tax	Long-term unoccupied	A long-term empty property which has been unoccupied and substantially unfurnished for at least two years attracts a council tax premium of 150% of the normal council tax liability
Household surveys	Any length	In ONS household surveys, interviewers are encouraged to understand whether a household with which they cannot make contact is actually unoccupied by looking for signs that someone might be at home or have returned to the address at some time since their last visit. This might include a light on at night, different position of curtains or post having been collected. If the interviewer suspects the address is a holiday home or vacant, they are encouraged to confirm this with neighbours

Appendix E: Initial quality assurance and processing of the data

Quality assurance of the data

The Commission for Energy Regulation (CER) in Ireland cleaned the data before ONS received it. They removed data from households and businesses who left the trial before its end and partial records where readings were missing. Therefore within ONS, a number of additional steps were taken to quality assure the data from the domestic meters as highlighted below.

Figure 8: Frequency distribution of total electricity consumption during trial period (kWh)

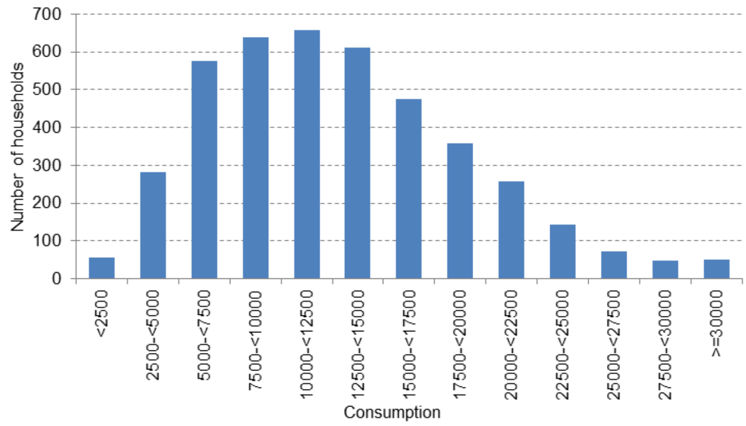


Table 7: Measures describing the distribution of total electricity consumption during trial period (kWh)

Minimum	370
Maximum	57,244
Median	12,146
Mean	12,894
Standard deviation	6,421

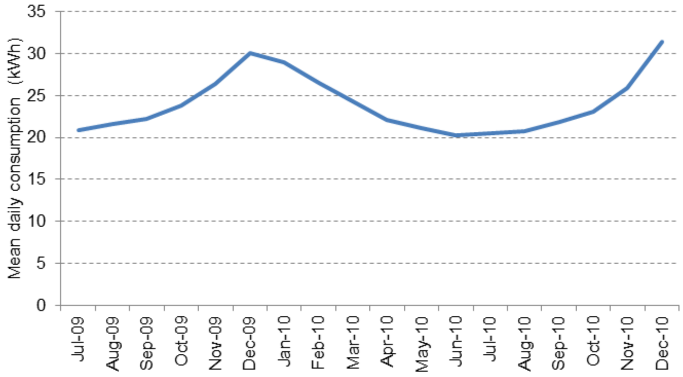
Figure 8~~Figure 8~~ and Table 7~~Table 7~~ show that the distribution of total electricity consumption is reasonable and does not raise any significant quality problems.

Formatted: Font: Not Bold, Not Italic

Formatted: Font: Not Bold, Not Italic

The total number of readings across all households is very consistent with around 6 million readings per month. The number of readings each month varies slightly due to the variation in the number of days in each month.

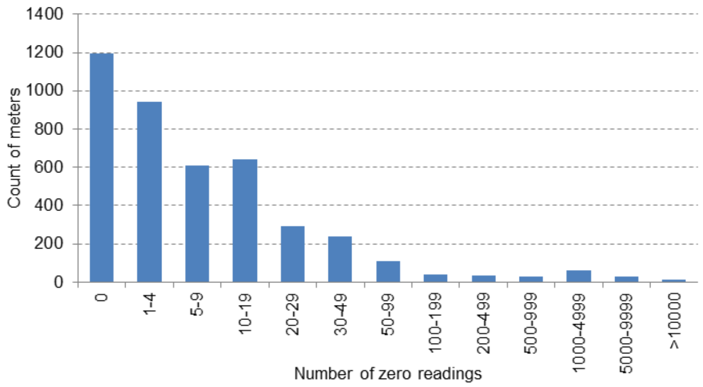
Figure 9: Mean daily electricity consumption over all meters by month (kWh)



As should be expected, there is a seasonal pattern to electricity consumption with the average consumption at its lowest in the summer (20kWh in June 2010) and at its highest in the winter (30kWh in December 2010).

Zeros in the data may be valid values which indicate no consumption (including no underlying loads such as fridges or freezers), or a power cut for example. However zeros may also indicate the failure of the meter to register consumption, perhaps due to a wireless connection being lost.

Figure 10: Frequency distribution of zero readings



There were 565,480 readings in the data (or 0.52%) which contained a zero value. As Figure 10~~Figure 10~~ indicates, the majority of meters contain less than 9 zeros out of an expected 25,730 readings for each meter during the trial period. However, 11 meters contain more than 10,000 zeros out of 25,730 expected readings, with one meter containing 20,749 zeros (81% of all readings). This is correspondingly the meter with the lowest consumption during the trial period. It is impossible to determine whether so many zero readings indicate an error with the smart meter or if there really is so little consumption at this household.

Formatted: Font: Not Bold, Not Italic

There were 30,480 missing readings in the dataset overall (0.03% of all readings). These represent whole days. 539 meters contained one whole day of missing values, 45 meters contained two days of missing values and 2 meters contained three days of missing values. However, as Table 8~~Table 8~~ indicates, some days contained more missing values than others. For example on 20 July 2010, 433 meters (10% of all meters) had no values recorded.

Formatted: Font: Not Bold, Not Italic

Table 8: Number of meters with whole days missing by date

Date	Meters with missing values
03/09/09	2
19/07/10	2
20/07/10	433
15/11/10	1
16/11/10	1
03/12/10	2
04/12/10	43
05/12/10	143
23/12/10	8

In summary, the data appears to be broadly of good quality, with very little missing data and relatively few observations with zero readings. The seasonal pattern of electricity consumption is present, as expected. However a few meters have very low consumption and a very high number of zero readings. While all meters were included in the analysis undertaken, in future such meters may need to be excluded.

Initial processing of the data

The data arrived split into six files which each contained three variables (meter ID, day by half hour time period and consumption). There were 6,445 meters in the survey (for both domestic and non-domestic customers) yielding 158 million rows of data over the 18 month trial period.

Table 9~~Table 9~~ illustrates ten rows of the data. The columns are:

Formatted: Font: Not Bold, Not Italic

- meter_id - This is a unique identifier for each electricity meter (ie. for each household or business)
- time_day - This is the day and time period. The first three digits represent the day where day 001=1 January 2009 and the last two digits represent a half hour period. So 19501 is the half hour period from 00:00 to 00:29 on 14 July 2009
- consumption - This is the electricity consumption over the half hour period in kilowatt hours.

Table 9: Sample of ten rows of the original data supplied

meter_id	time_day	consumption
1063	19501	0.362
1063	19502	0.064
1063	19503	0.119
1063	19504	0.023
1063	19505	0.140
1063	19506	0.036
1063	19507	0.108
1063	19508	0.083
1063	19509	0.056
1063	19510	0.129

An additional file was supplied, indicating the meter type for each meter ID and using this information the consumption data for non-domestic customers was removed, leaving 108 million rows of data covering 4,225 residential households.

For easier onward processing, these data were pivoted in MongoDB[5] from a long thin file (containing 3 columns and 108 million rows as per Table 9~~Table 9~~) to a short wide file (see example in Table 10~~Table 10~~). The short wide file contained 2.3 million rows (one row for each meter / day combination) and 52 columns (meter ID, day, plus 50 half hour timeslot fields containing the corresponding consumption values).

Formatted: Font: Not Bold, Not Italic

Formatted: Font: Not Italic

Table 10: Sample of five rows of pivoted data used in analysis

meter_id	day	t1	t2	t3	t4	t5
1063	195	0.362	0.064	0.119	0.023	0.140
1063	196	0.195	0.167	0.156	0.105	0.115
1063	197	0.275	0.145	0.117	0.082	0.045
1063	198	0.122	0.153	0.130	0.137	0.120
1063	199	0.137	0.076	0.041	0.054	0.064

[5] MongoDB is an open source NoSQL database

Appendix F: Details of research into methods 3-8

Section ~~4.e~~4.e provides detail about two of the methods which were used to try to determine whole days when households are unoccupied. The remaining six methods are detailed below. These use either night time activity as a baseline, or take three months of activity as habitual and examine days which do not follow this habit.

Formatted: Font: (Default) Calibri, 11 pt

In methods 1, 4 and 5 the maximum and minimum values are removed across the relevant time period before obtaining the averages and variances. This is so that odd spikes in electricity consumption do not have a big influence on whether the method selects a day as being unoccupied.

Night time was referred to as 1am - 4am in this paper which examined the feasibility of using smart meter data to detect occupancy, while 1am - 6am was used in the University of Southampton report <LINK TO REPORT> which looked at the feasibility of using smart-type meter data to support official statistics. Therefore 1am - 5am was chosen as a compromise between these two papers, even though it is recognised that this will not be appropriate for all people such as night workers.

Several of the methods require a threshold against which to compare the measure (for example average day time consumption divided by average night time consumption). It would not be expected that this ratio would be exactly equal to 1 due to timed appliances turning on or off. Therefore, the final thresholds were chosen after examining the half hourly consumption profiles for some meters. Slightly narrower and wider ranges were tried but these either yielded too few unoccupied days or too many occupied days.

Methods 1, 4 and 5 were also investigated using log consumption (because log consumption is used in the University of Southampton report), but this did not demonstrate any improved accuracy. Consumption was logged because doing so controls the variance in the data and minimises the impact of very high consumption values.

Method 3: Total energy consumption for a day is less than the 5th percentile of the daily consumption over the previous 3 months

This method was developed after examining daily totals of electricity consumption for several meters and trying to determine an absolute kWh threshold for each, below which it appeared that some days were unoccupied. On average it appeared that this threshold amounted to around the fifth percentile of the three month daily consumption. It also means that around 5% of days in a three month period (ie. around 18 days per year) are classed as unoccupied, which appears reasonable.

This method uses a threshold value in daily electricity consumption, below which a day is classed as unoccupied. For each meter, this threshold is set independently for each day being assessed for occupancy in turn using energy consumption information from the three months around the day in question (45 days before and 45 days after). Specifically, for each meter and any given day, this threshold is set to the value for which only 5% of days have a lower daily energy reading across the three months. The first 45 days could not be assessed due to the absence of a full three month period.

A household is unoccupied if:

Total daily electricity consumption for current day < 5th percentile of the 3 month daily consumption

The use of three months of data attempts to factor in both the usual energy consumption for a meter and also provides a period of time long enough to reasonably observe some unoccupied days. More in question is the impact of seasonal variation in energy usage as three months is possibly too long a period to ensure that all daily energy consumption values can produce a representative threshold for occupancy for the day being assessed.

The approach effectively sets the occupancy rate to be 5% which is observed in the results. These show that for all 10 meters there are 28 or 29 days classed as unoccupied which represent around 5% of the 536 days each meter was in the trial. Clearly, the threshold is set too high for meters which have fewer than 5% unoccupied days (such as people who don't spend much time away) as it will identify days as unoccupied which are in fact occupied, and is set too low for meters that have more than 5% unoccupied days (see Figure 11~~Figure 11~~, for example second homes or homes occupied by weekday commuters).

Formatted: Font: Not Bold, Not Italic

Figure 11: Half hourly consumption for a day identified as occupied by Method 3

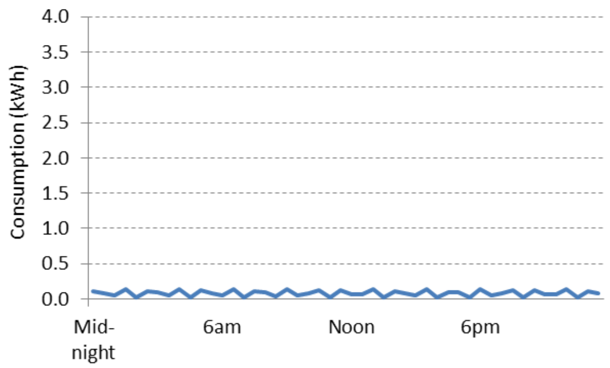


Table 11: Confusion matrix for method 3

Method 3	Examined by eye	
	Unoccupied	Occupied
Unoccupied	108	184
Occupied	53	4,115

Table 11~~Table 11~~ shows that of the 161 days visually classed as unoccupied, this method successfully identified 108 of them giving a sensitivity of 67%. Correspondingly the method correctly identified 4,115 of 4,299 occupied days, giving a specificity of 96%.

An improvement to this method might involve the identification of a suitable threshold based on any step change in daily energy usage over three months that may signify the difference between an occupied day versus an unoccupied day.

Further extension of the method might consider the feasibility of identifying a general threshold value that may provide a good classification rate across all households.

Method 4: Day time average is below average of previous 3 months' maximum night time consumption

This method again uses a threshold energy consumption value below which a day is classed as unoccupied. The threshold for each day being examined is set to the average of the maximum half hourly consumption during each night over a three month period.

A household is unoccupied if:

Mean day time consumption for current day < Mean of maximum night time consumption over a 3 month period

Of note is that night time is defined as the eight half hour periods between 1am and 5am, and daytime represents the remaining 40 half hour periods in a 24 hour day. Furthermore for the calculation of both the means in the equation above, minimum and maximum values were first removed to prevent outliers in energy consumption influencing the derivation of the averages.

This method requires night time usage to be indicative of low activity such as would be expected in an unoccupied household. By using the average of the maximum half hour consumption values for each night it is considered that a high enough threshold is produced to adequately identify unoccupied days. For example, for 24 hour periods where the electricity consumption profiles are low and fairly flat, suggesting an unoccupied household, the average of the maximum night time consumption over three months should be slightly higher than the mean day time consumption.

Figure 12~~Figure 12~~ and Figure 13~~Figure 13~~ provide examples where method 4 has not worked well. In both cases the method has classed the days as being unoccupied even though evidence of active occupancy can be seen.

In Figure 12~~Figure 12~~ a single spike of high energy consumption is present in the morning. Here method 4 fails to register this energy peak as it is excluded from the calculation of the daily mean energy consumption. In Figure 13~~Figure 13~~ there is quite a lot of active occupancy seen across multiple half hour periods. Method 4 identifies this day as unoccupied because consumption is zero between 11am and 3pm, which moves this meter slightly below the chosen threshold.

Figure 12: Half hourly consumption for a day identified as unoccupied by Method 4

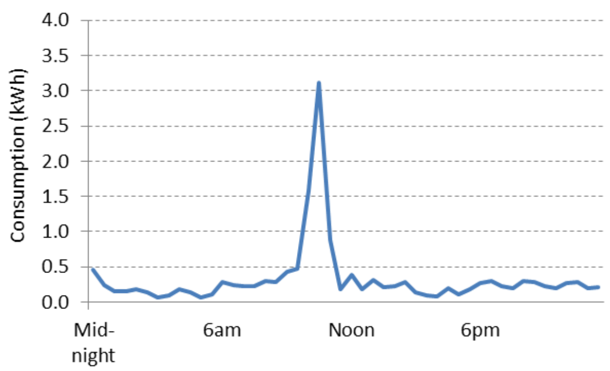


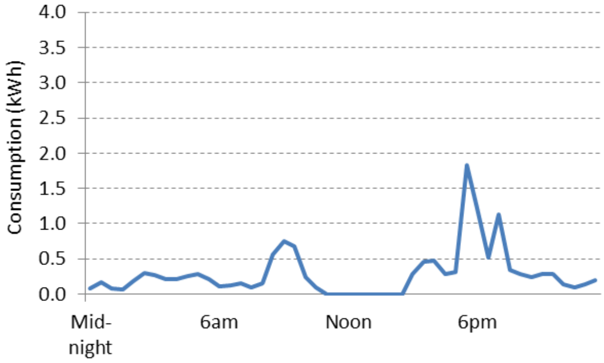
Figure 13: Half hourly consumption for a day identified as unoccupied by Method 4

Formatted: Font: Not Bold, Not Italic

Formatted: Font: Not Bold, Not Italic

Formatted: Font: Not Bold, Not Italic

Formatted: Font: Not Bold, Not Italic



~~Table 12~~Table 12 illustrates that this method classifies too many days as being unoccupied. It might show some improvement if the maximum half hourly energy consumption during the day being assessed is retained within the average calculation as single high energy peaks within a half hour period would then be registered. Correcting for high usage night time use is more difficult as it challenges the assumption that night time energy use is a benchmark for inactivity required by the method.

Formatted: Font: Not Bold, Not Italic

Table 12: Confusion matrix for method 4

Method 4	Examined by eye	
	Unoccupied	Occupied
Unoccupied	159	301
Occupied	2	3,998

~~Table 12~~Table 12 shows that of the 161 days visually classed as unoccupied, this method successfully identified 159 of them giving a sensitivity of 99%. Correspondingly the method correctly identified 3,998 of 4,299 occupied days, giving a specificity of 93%.

Formatted: Font: Not Bold, Not Italic

This method was also tested on logged smart meter data to further reduce any impact had by extreme values within the calculation of average energy consumption but no improvement in its functionality was seen.

Method 5: Day time variance is similar to night time variance

In an occupied household the variance of day time consumption is expected to be greater than variance at night time, while in an unoccupied household this ratio should tend towards 1. However it would not be expected that the ratio should be exactly 1, so various tolerances around 1 were tested on ten meters and 0.5 and 1.5 were found to be optimal. This is because widening the threshold appeared to identify too many occupied days while narrowing the threshold appeared to identify too few unoccupied days.

A household is unoccupied if:

$$0.5 \le \frac{Variance\ of\ \ day\ time\ consumption\ \ for\ current\ day}{Variance\ of\ \ night\ time\ consumtion\ \ for\ current\ night} \le 1.5$$

Method 5 does not work well when the variance of the night time electricity use is much smaller than the variance of the day time use. This is because the ratio (of night variance to day variance) is too big and outside of the chosen threshold of 0.5 to 1.5. For example at one sampled meter, it appears that four consecutive days are unoccupied as they all have flat electricity consumption profiles. However, unlike most of the other methods, this method only selects the last of these days as being unoccupied. The first three days are not selected because the night time variances for these days are smaller than the day time variances, resulting in large ratio values which are outside the chosen threshold of 0.5 to 1.5.

Table 13: Confusion matrix for method 5

Method 5	Examined by eye	
	Unoccupied	Occupied
Unoccupied	98	64
Occupied	94	5,104

Table 13~~Table 13~~ shows that of the 192 days visually classed as unoccupied, this method successfully identified 98 of them giving a sensitivity of 51%. Correspondingly the method correctly identified 5,104 of 5,168 occupied days, giving a specificity of 99%.

This method would not work well in correctly identifying occupied days at households where there is a high variance at night and during the day, but the ratio of these is around 1. Further, as information about the current day and night are used in this method, it is sensitive to sudden changes in energy usage between day and night.

This method could be improved by examining night time consumption over a longer period of time than one night, or by combining it with other methods.

Method 6: Day time average is below night time average plus 1 Standard Deviation (using log consumption)

This method identifies days as being unoccupied when the day average is similar to or smaller than the night average. This was one method tested in the research undertaken by the University of Southampton <INSERT LINK>. Again, night time is used as a baseline for low activity. The values are logged in order to reduce the influence of very high values, and minimum and maximum values were retained in the calculation.

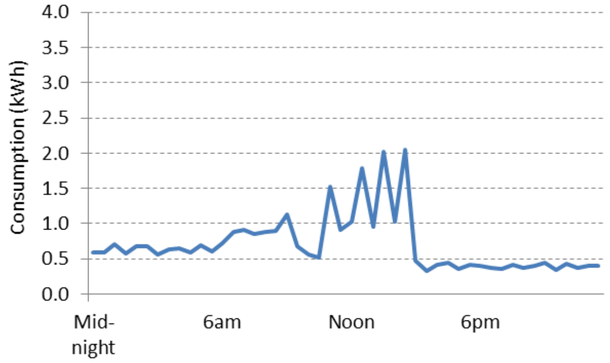
A household is unoccupied if:

Mean of day time logged consumption for current day < Mean of night time logged consumption for current night + one standard deviation of the night time logged consumption for the current night

Like several of the other methods, method 6 selects three consecutive days as being unoccupied at one sampled meter, but it also selects the day before even though there appears to be a lot of electrical activity during the lunch time and early afternoon (Figure 14~~Figure 14~~). In this case this day is identified as being unoccupied because the day and night averages are similar, so the day average is below the night average plus one standard deviation.

Formatted: Font: Not Bold, Not Italic

Figure 14: Half hourly consumption for a day identified as unoccupied by Method 6



This method identifies days as being unoccupied when the day and night averages are very similar. This could be at truly unoccupied households or for households with night workers who spend the day sleeping or households where someone is in during the day but does not use many electrical appliances. Again, as information about the current day and night are used in this method, it is sensitive to sudden changes in energy usage between day and night.

Table 14: Confusion matrix for method 6

Method 6	Examined by eye	
	Unoccupied	Occupied
Unoccupied	191	712
Occupied	1	4,446

Table 14~~Table 14~~ shows that of the 192 days visually classed as unoccupied, this method successfully identified 191 of them giving a sensitivity of 99%. However the method only correctly identified 4,446 of 5,158 occupied days, giving a specificity of 86%. Because specificity is much lower than the other methods tested, it is suggested that no further work is conducted into it.

Formatted: Font: Not Italic

Method 7: Range (maximum minus minimum) of day time consumption is similar to range of night time consumption

Like other methods, the range of the day time consumption would be expected to be similar to the range of the night time consumption. Several thresholds were tested, and 1.6 found to be optimal.

A household is unoccupied if:

$$1.6 > \frac{Range\ of\ day\ time\ consumption\ for\ current\ day}{Range\ of\ night\ time\ consumption\ for\ current\ night}$$

This method does not identify as many truly unoccupied days as other methods (Table 15~~Table 15~~). This is likely to be because this method is sensitive to particularly low or high values. However it is the third most accurate method and it works well in Figure 15~~Figure 15~~ where the other methods do not work so well, possibly due to a change in the underlying electricity load.

Formatted: Font: Not Italic

Formatted: Font: Not Bold, Not Italic

Table 15: Confusion matrix for method 7

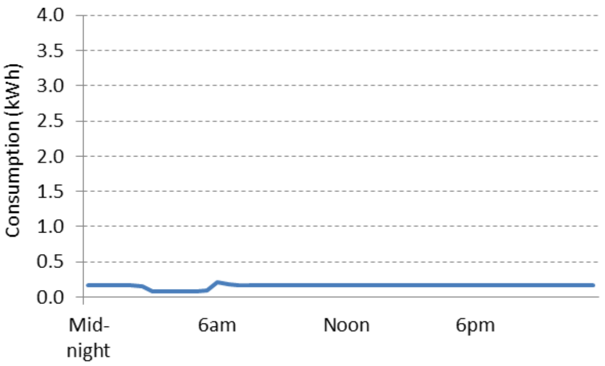
Method 7	Examined by eye

	Unoccupied	Occupied
Unoccupied	101	61
Occupied	91	5,107

Table 15~~Table 15~~ shows that of the 192 days visually classed as unoccupied, this method successfully identified 101 of them giving a sensitivity of 53%. Correspondingly the method correctly identified 5,107 of 5,168 occupied days, giving a specificity of 99%.

Formatted: Font: Not Italic

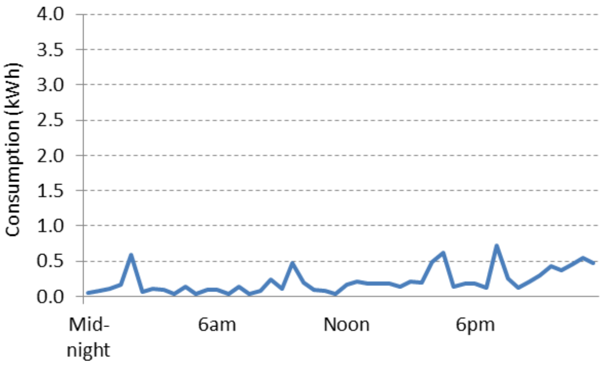
Figure 15: Half hourly consumption for a day identified as unoccupied by Method 7



However this method does not work well in instances where there appears to be some night time activity, for example returning from night clubbing or staying up late before going to night work. That is because in these cases the range of the night time consumption can be similar to the range of the day time consumption, as illustrated in Figure 16~~Figure 16~~:

Formatted: Font: Not Bold, Not Italic

Figure 16: Half hourly consumption for a day identified as unoccupied by Method 7



Like other methods, this could be improved by examining night time consumption over a longer period of time than one night, or by combining it with other methods.

Method 8: Inter-quartile range (IQR) of day time consumption is similar to IQR of night time consumption

The idea of using this method is that there may be one spike in usage during the day as perhaps a cleaner arrives at a household, or a timed appliance turns on. Method 7 above may not identify such a household but using this method would.

A household is unoccupied if:

$$1.3 > \frac{IQR\ of\ day\ time\ consumption\ for\ current\ day}{IQR\ of\ night\ time\ consumption\ for\ current\ night}$$

Table 16~~Table 16~~ shows that this method incorrectly identifies more days than other methods and correctly identifies fewer truly unoccupied days than other methods. It shows that of the 192 days visually classed as unoccupied, this method successfully identified 109 of them giving a sensitivity of 57%. Correspondingly the method correctly identified 4,979 of 5,168 occupied days, giving a specificity of 96%.

Formatted: Font: Not Italic

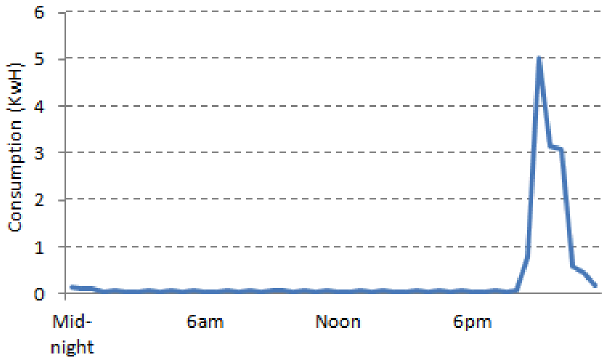
Table 16: Confusion matrix for method 8

Method 8	Examined by eye	
	Unoccupied	Occupied
Unoccupied	109	189
Occupied	83	4,979

It was not possible to find an example where this method worked well but other methods did not. Conversely, there were many examples of where this method did not work well. This was caused by effectively not considering half of the data points each day (the top and bottom 25%). One example of where this method did not work well is in Figure 17~~Figure 17~~:

Formatted: Font: Not Bold, Not Italic

Figure 17: Half hourly consumption for a day identified as unoccupied by Method 8



It is suggested that no further work is conducted using this method.

Appendix G: Increasing the size of the data

The eight methods investigated for the sample of 10 meters were tested on all 4,225 households in the dataset (a 700MB file) (results in the main body of the report). However if smart meters are to be rolled out to every household in the country by 2020 and this data used within the production of official statistics, the infrastructure and big data technologies would need to be in place to be able to potentially analyse a dataset covering 20 million households. It is estimated that such a file providing half hourly readings over a year would be 2.2TB in size. This section explores how this could be achieved.

About ONS' innovation labs

The innovation labs have been set up to help facilitate research into new technologies and open source tools, new sources of public data and to develop associated skills. The innovation labs are a key enabler for the ONS Big Data project since they allow us to handle large and complex data sets and to test new big data technologies.

The labs consist of a number of high specification desktop computers with some additional network storage. The hardware is configured using OpenStack cloud computing technology. This provides a very flexible environment to deploy different virtual environments depending on the processing and storage requirements of different projects. In particular, this approach provides a flexible framework for experimenting with big data parallel computing technologies such as Hadoop. The labs have been designed to provide a route for accessing open source tools.

Potential of R packages ff and ffbase

R is a free software package for statistical computing and graphics. It is widely used in the fields of statistics and big data, and for this reason ONS used it to process and analyse the smart-type meter data. A limitation of R is that it can only address objects that fit in the available virtual memory space so it cannot cope with very large datasets. Therefore the potential of R packages ff and ffbase which are designed to overcome this limitation have been examined. They extend the R system by making use of data which is stored elsewhere (such as on a file share) rather than in the main memory. These packages are used in big data research.

As an example, one particular operation (joining two tables together by a common key field) was tested in base R in the innovation lab and returned an error "Cannot allocate a vector of size 4034.6 Gb". However the same operation was possible using the packages ff and ffbase.

Potential of parallel R package

Another limitation of R is that it carries out operations using only a single computing core. R has packages which enable parallelisation of some tasks utilising the existing multi-core infrastructure in the innovation lab. The R package called "parallel" was briefly tested and it produced promising results: after increasing the size of allocated RAM of the OpenStack instance, R successfully carried out operations using 8 computing cores and parallelisation sped up processing time considerably.

Potential of Hadoop

Hortonworks is a platform which supports Hadoop, a framework that allows the distributed parallel processing of large datasets across clusters of computers. It is able to handle much larger datasets than R.

One of the methods was successfully implemented using Hortonworks using a sample of 10 meters. The next step is for Hortonworks to be set up so that it can implement the methods on the full dataset. This has the long-term potential to be able to process the smart meter data which could be obtained from 20 million households over a year.