Generating Insights from Smart Meter Data: Challenges and Opportunities

Anastasia Ushakova

Thesis submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Geography and School of Public Policy

University College London

January 21, 2019

Declaration

I, Anastasia Ushakova, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The introduction of smart meter technology has been central to recent innovations in energy provision for the UK residential sector. Smart meters have the potential to give greater insight into energy consumption behaviour for energy providers and researchers alike. For example, they may aid our understanding of how the consumption of gas and electricity may be replaced by the energy from renewable sources, or how consumer behaviours can be changed to reduce overall energy consumption, increase efficiency, and lessen the pressure on the national grid networks. The advantage of a thorough understanding of the insights generated from smart meter data for policy issues may sound obvious at a first glance. However, there are significant challenges associated with the availability of methods and computation necessary to perform a complete analysis of the available data. The thesis provides an in depth look at the nature of energy consumption through an analysis of big data that is recorded by around 400,000 smart meters installed at residential properties across the UK. It further discusses how this data is different from perhaps more conventionally collected retail consumer data, and in what way does the temporal nature of these data reveal information about the customers dynamics without compromising their anonymity. Various machine learning methods are applied and surveyed against more conventional methods often used by researchers and industry practitioners. Some extensions to improve the accuracy and reliability of methods for both segmentation of the behaviour, and prediction are also suggested. Lastly, a case study looking at identifying the fuel poor from smart meter data is presented as an illustrative example of potential research questions one may answer with smart meter data records.

Impact Statement

Since 2005, demands to reduce the CO2 emissions nationally and internationally are growing, and thus the need to replace current gas and electricity energy resources by renewables. Consequently, the global investment in energy research has seen a rise (Skea, 2014). Smart meter data is the first resort for exploitation to make the sustainable energy goals achievable. The work in this thesis, in partnership with a UK-based domestic energy provider, focuses extensively on the analysis of spatiotemporal Big Data of UK residential energy consumption, constructed from thousands of smart meters and constitutes the most extensive data set of smart meters available to date. The findings of this research were shared and communicated with the industry partner who provided the data, and the methodology used in the thesis was attempted on energy company premises. The thesis thus shares emerging industry and academic research impact. The thesis provides a survey of methods and offers a toolkit to anyone who may hold smart meter data and wants to make sense of it. The main contribution of the work in a broader sense is a provision of a comprehensive guide that researchers and industry practitioners may refer to when smart meter data arrives at their hands in nearest future. It is the first attempt to combine together a step by step strategy for data analysis that allows for building a holistic picture of what insights these data may provide for various tasks that are associated with description and visualisations, segmentations and data organisation, forecasting, and data reduction. Smart meter data is an example of a complex time series data structure that holds information on the dynamics of individual activities. Given the rise of big data and devices that collect information of similar type in real time, methods and analysis presented in the thesis may be well transformed to other

datasets that may have a similar structure. This thesis suggests that smart meter data cannot be looked at solely through a social science lens, neither can it be completely understood by computer science and statistics approach. The marriage of two is necessary for a complete explanation of variation among smart meter users, areas where they live and the variation within individual users energy consumption. This thesis is the first attempt to bring these two together in application to smart meter data and suggest ideas for data interpretation and detailed exploration.

Acknowledgments

I would like to say thanks most of all to my supervisors Paul Longley and Slava Mikhaylov. To Paul, I am grateful for our cheering conversations at George Farrah cafe down the road and occasional disagreements which always pushed me to do more. Little did he know what it would be like when he took on board non-geographer, but I believe it was more than just an interesting journey for both of us. To Slava, I am indebted for introducing me to this opportunity. For always inspiring and encouraging me to try things which may seem impossible at first and for being there whenever all seemed a bit all over the place. To both of them I am thankful for finding time to meet and talk about both work and life and letting me do my own trials and errors. Also for their great sense of humour and pragmatism, different in every way, but certainly capable of cheering me up at any stage of the PhD process.

To the Consumer Data Research Centre (CDRC) and Economic and Social Research Council (ESRC) and people who approved the funds to be given for me to spend during my PhD work. To all amazing people I met along the way of research and teaching: Keith Dugmore, Bruce Jackson, Stephen Haben, Jennifer Hudson, Ozan Aksoy, Jack Blumenau, Philipp Broniecki, Tom O'Grady, Orlanda Siow, Liam Quirke, David Lee, Andy Simpson, and many many others. I am grateful to CDRC management team represented by Sarah Sheppard and Navta Vij who were always there to assist with us any administrative issues and also take care of all our important dates such as birthdays and weddings.

I am grateful to my examiners, Colin Provost and Shaun Bevan, for reading my work and providing me with the encouragement to continue my research. Their feedback was valuable and I appreciate their pragmatic and critical way of accessing

my thesis, given the fact that nature of such experimental work is not very common in social science practice.

To my colleagues in Chorley: Alyson, Guy, Karlo, Ollie, Ffion, Alistair, Kira and Rad for occasional drinks nights and reminding ourselves of the beauty of PhD timetable, for playing badminton in the office as a new form of break. To my dear friends living all over across the world now and more specially, to Paolo for failing in attempting to take me for writing breaks, but being there to talk about teaching adventures and walk to Euston back and forth from our homes, to Preeti, for knowing that I may need some time out, sometimes to sit outside in Barbican, to Emilia, for being source of constant inspiration even when being so far away. To Paul, to talk about literature and struggles of being human, to James and Dan, for being there any time for a great laugh and coffee too and to Francesco, for always saying that this PhD will turn out to be extraordinary without even knowing what it was all about.

I would like to say thanks to all my London family and to Alex, my partner who over the course of this PhD became my husband. I am thankful for our walks, days offs in the middle of the week, resistance to my stress and for giving me support with any trouble that may come out of during PhD process, including proof-reading this piece! To my giant family in-law for supporting me throughout all this time, and most of all to my mother, Oksana and my grandmother, Galina, for being so understanding of me coming back home not as often as I hope I will do once the PhD is over.

To the PhD itself, I am really grateful for what it made of me and the way it turned my life upside down. It made me someone who through inevitably growing levels of procrastination managed to run half marathon, learned how to be a great cook, read an absurd amount of fiction and non-fiction and began to write one herself. What a journey it was!

Thesis Outputs

Some of the work presented in this thesis have appeared in the peer reviewed conferences and published material. This is particularly relevant for part of Chapter 3, Chapter 4 and Chapter 6. Most of the thesis outputs are currently under submission or in preparation for being submitted to peer reviewed journals.

Work in Progress

- Ushakova, A. and Mikhaylov, S.J. 'Predicting Energy Customer Vulnerability from Consumption Behaviour' *Submitted*

- Ushakova, A. and Mikhaylov, S.J. 'Learning to Predict: Estimating the Structure of the Non-Stationary Spatiotemporal Profiles for Behaviour Prediction' *R and R*

- Ushakova, A. ' Generalised Additive Models for Residential Energy Load Prediction' *In Preparation*

Book Chapter

- Ushakova, A. and Murcio, R. 'Interpreting Smart Meter Data of UK Domestic Energy Consumers' (pp 120-137) in Longley et al (2018) *Consumer Data Research*

Peer Reviewed Conferences and Presentations

- UCL Energy Institute Seminar Series, 2018 , *(Invited Talk)*

- Home Office Research Methods Briefing, 2018, *(Invited Talk)*

- American Political Science Association (APSA) Annual Meeting, 2017, San Francisco, CA *(Talk)*
- Theory of Big Data Workshop, 2017, London, United Kingdom, *(Poster)*
- Women in Machine Learning as a part of NIPS, 2016, Barcelona, Spain, *(Poster)*
- Consumer Data Research Centre (CDRC) Meeting, University of Leeds, 2017, Leeds, The UK *(Talk)*
- Research meet of UK Data Service (UKDS) and UCL Energy Institute under the umbrella of Smarter Household Energy Data project, 2016, London, The UK
- Data Science meets Social Science, LSE, 2016, London, The UK *(Talk)*
- CDRC Meeting, University of Oxford, 2016, Oxford, The UK *(Talk)*
- Research Meet of CDRC and Administrative Data Research Centre (ADRC), UCL, 2016, London, The UK *(Talk)*

Contents

List of Figures

List of Tables

Chapter 1

Introductory Material

'If you torture the data long enough, it will confess'

- Ronald Coase in Tullock (2001),

The field of energy research, whether it is household energy consumption, decision making on renewable sources, energy supply or energy economics, is becoming of growing importance for both academic and industry communities. Censoring energy by using smart meters and the ability to quantify energy use at high temporal granularity offers numerous opportunities for research communities that focus on policy, economics, resource management, geography, the built environment, and statistics. Other research domains choose to focus on the sensitivity of energy consumption to prices of electricity and gas, customer habits and weather, using these to better understand how consumption varies from one user to another.

Over the last two decades, a vast amount of research has aimed to show the relationship between energy variation and factors such as property and household characteristics, with the end goal of guiding policymakers and energy suppliers in their delivery of efficient and fair provision of resources (Beckel et al., 2014; Kavousian et al., 2013; Albert and Rajagopal, 2013; McLoughlin et al., 2012b) . Despite this body of research, the dynamic nature of energy use and often high variation among users means there is still a range of research needed, with unexplained variation in energy consumption being on average about 54 percent, according to Huebner et al. (2015), even when considering a sample with a sufficient amount of additional variables on housing and sociodemographic characteristics. Samples of

energy data may differ as may levels of aggregation. Samples of energy data may differ, levels of aggregation may differ and different types of predictors occur, depending on which behaviour researchers are interested in observing. All of which can lead to the use of different methods and thus, differing results. This thesis will show that there is, as yet, little agreement on which methods are optimal for classifying or predicting consumer energy usage using smart meter data as the field is still underdevelopment and a vast amount of experiments need to be performed before research in this highly interdisciplinary area can be considered state of the art.

In this thesis, an attempt to unlock some of these issues is presented through surveying the various methods a researcher considers using, taking into consideration that these decisions often depend on data resolution and the final research aims.

It is highly important to reach an understanding of how much energy is being consumed by UK residents, and how much of their current consumption of gas and electricity may be replaced by energy from renewable sources. It is equally important to understand how consumer behaviours can be changed nationally to reduce energy consumption overall, lessen the pressure on the national grid networks and thereby increase energy efficiency.

The advantages of a thorough understanding of the insights generated from smart meter data for policy issues may appear obvious at first glance. However, the respective challenges and disadvantages associated with the computational power necessary to perform a complete analysis is a major obstacle that needs to first be overcome. This thesis aims to address these obstacles and present some of the possible ways in which they may be tackled.

1.1 Big data and new forms of data in social science research

Not just the energy research was expanding over past decade, but other phenomena have entered the stage, namely Big Data. The emergence of big data[1] and its associ-

[1]Please note that the term 'Big Data' is capitalised only in cases when referring to the phenomena as in Boyd and Crawford (2012). More often the term 'big data' used as it refers directly to the data of characteristics that are associated with the phenomena

ated potential disruptions to traditional methods used in social science research was acknowledged by Kitchin (2013) and Barnes and Wilson (2014). Kitchin (2013) has largely warned the field about the important changes and adaptations social science methodology will face after the arrival of big data, while the latter has provided a rather optimistic vision of the future of social science in the light of big data. Barnes and Wilson (2014) looked primarily at how social scientists, to be more specific, social physicists such as George Zipf and William Warntz, unlocked the lid of large datasets potential for the social science field putting themselves forward among the researchers who could make sense of big data to study society. Overall, it is still hard to say whether big data is a curse or blessing for the social science field. The attempt to analyse big data on energy consumption and study of the potential insights that can be observed about the social dynamics from this data can let this thesis to take a stance in the debate.

Some of the most obvious challenges associated with this emerging notion is certainly the definition. There is still no defined description of big data due to the novelty of the phenomena and associated with it concept. Some may suggest that term 'big data' is rather vague and used so widely in different domains and applications that it became rather meaningless (Goes, 2014). Some may say that in a nutshell it is the complexity and the size what defines big data and separate it from other more conventional forms of data (Taylor et al., 2014). One of the most famous and widely used definitions of big data is that of Laney (2001) who uses the concepts of volume, variety and velocity as main ingredients of the big data. The recent literature takes this definition further by introducing key characteristics of big data as exhaustive, relational, flexible and high in resolution. These are the features that makes these data not just a unique form of data but an innovation which is highly disruptive in nature. Smart meter data is not an exception as the arrival of smart meter data has changed fundamentally how energy consumption can be studied by moving away from energy consumption estimate at monthly or annual level, we are now dealing with highly granular temporal dynamics and able to study consumption behaviour throughout the day during any month, any season

of year. Below, smart meter data is described with respect to each of the big data characteristics that are available to date (Boyd and Crawford, 2012; Dodge and Kitchin, 2005; Mayer-Schönberger and Cukier, 2013)

- **Volumous**: the data source is large in volume as each smart meter user annually would have around 17, 520 readings.

- **Variable**: the data tends to come in various forms. For instance, it may be described by time series that reflect the energy use over time together with the data of categorical nature that describe different type of energy source (i.e. electricity, gas) and geographical locations of the users. Where more detailed data on smart meter users is available , more variety in data structure can be expected.

- **High Velocity**: smart meter data is recorded automatically and offers real-time updates, typically recorded at half hourly intervals;

- **Exhaustive** : the data coverage is striving to include as much as possible of the population and aims to cover the whole country, this is different from traditional sample collection tailored for small sample case studies (Kitchin, 2014)

- **Relational**: having various geographical reference for smart meter data makes it suitable for co-joining with other datasets

- **Flexible and Scalable**: new data can be added easily to smart meter data as well as smart meter data available from different source can easily be joined together due to homogeneous nature of the way this data is recorded.

- **High in resolution**: in principle, smart meter data readings provide a detailed picture on consumption behaviour compared to conventional ways to record energy consumption. Half-hourly resolution can be considered as high enough. Besides, it is anticipated, that the resolution of minutes or even seconds will be available in the future to governments and energy companies.

1.1.1 End of Theory?

A debate on whether big data brought 'the end of theory' remain relevant for this thesis as once again, so far the available research have seen times of trial and error, experiments and constant exploration. Anderson (2008) has dedicated his book to discussion of how the correlation analysis in big data may replace the theoretical and scientific approach to the relationships observed in the data. Kitchin (2013, 2014) discussed extensively how big data gives a rise to the new epistemological standpoint that social science researchers need to take. Mainly, to embrace the paradigm shifts that result from the need of rethinking how new forms of data could be incorporated into the existing methodologies. In this thesis, the attempt to take a very exploratory approach is taken to study smart meter data. There is no defined methodology to analyse these data yet, likewise there is no defined state of the art strategy to evaluate the big data recorded by smart meter data due to unavailability of datasets of such magnitude in the past.

Fortunately, we are living in the age of a significant expansion of the applied data science field that allow the social science researchers interested in big data to borrow valuable techniques that may help them in tackling some of the most pressing challenges associated with smart meter data. For example, data mining and the field of knowledge discovery from large databases have become popular in terms of the technology useful for analysing smart meter data. This is in part due to both researchers and industry practitioners being attracted to big data, data that have transformed largely how the individual behaviour and decision making can be studied given the various forms of the digital footprints produced by the individuals each day.

At this current moment, data mining in the energy sphere is considered mostly for its utility for decision making in business strategy research, whilst other forms of big data has spread out to other disciplines such as the biomedical science, physics, engineering and various behavioural research fields within economics and political science (Wu et al., 2014; Swan and Ugursal, 2009).

Big Data is certainly everywhere, with both business and universities seeking

to create departments dedicates to its analysis and use (Hand, 2016). According to Hand (2016), big data-based research aims can be associated with two streams of tasks. One of these is data management and data manipulation such as matching data, sorting it, cleaning it and perhaps, also, linking it with other datasets. Another type of big data exercise is concerned primarily with determining what big data can tell us about the future. This is where prediction tasks for various scenarios can be considered.

This thesis will thus look in more depth at the nature of energy consumption by analysing big data recorded by the smart meters at residential properties across the UK. It will further look at how this data differs from any other retail collected consumer data. In what way does the temporal nature of this data, which records timed responses, reveal information about the customer without compromising their anonymity? This may mean inferring working hours or the hours someone spends at home, or studying the seasonal responses of households on their energy loads.

Machine learning, computational statistics and artificial intelligence (AI) are growing in popularity as common streams of approaches for tackling big data sets that have similar granularity as energy consumption readings. This thesis is not an exception as some of the machine learning methods will be adapted for smart meter data segmentation later on in the thesis. To give a brief idea, machine learning is characterised by automated methods which can perform statistical tasks. Some examples may include applications in healthcare as well as remote sensing technologies often using in geo sciences to study satellite images of the Earth. Developed primarily by academic researchers, these methods are used in services that rely largely on the extensive collection, processing of data records and computation. They have also been widely adapted in those private sectors that use the data for delivery of services (i.e. Amazon, etc.). However, when considering the provision of services within the public sector, such as essential energy resources, artificial intelligence still seems quite a long way from replacing traditional methods of data analysis. Not because they may be less useful, it is in fact driven primarily by the limitations in access to advanced computing technologies as well as the need to

train energy companies analysts to be able to use these new novel methodologies. This thesis aim to address this by providing a detailed step by step guide on how some examples of such advanced methods can be used and will survey some of these methods to identify those that can be useful for energy consumption analysis. They will also be discussed in the comparison to more conventional methods often used by social science researchers.

1.2 Data Collection

The introduction of smart meter technology has been central to recent innovations in energy provision in the UK residential sector. Smart meter data has the potential to give greater insight into energy consumption behaviour, not just for energy providers, but also for the wider research community. The data generated by smart meters is an example of the emerging concept of Big Data. This concept started to gain popularity during the last decade as a revolutionary source of streaming information on people and objects which negated the need for routine survey collections, such as censuses. Smart meters offer a temporal breakdown of energy consumption data and therefore offer immediate advantages for research in terms of increased temporal granularity and sample size of data. However, as this data is new to both the research community and industry analysts, there is a need for a greater awareness of certain aspects that require caution during both data collection and analysis.

The way this data is recorded and transmitted can be considered pretty straightforward. The user, which is a household or a mix of households, receives a meter that is connected to their traditional meter systems. Gas and electricity are now under the same umbrella of the smart meter. The interface of the meter is now in real time and users can see how much energy has been consumed in the past half hour or past day, month or year. An illustration of a device prototype is presented in Figure 1.1. It is about the size of a standard digital e-reader, and it can immediately show the consumer the cost of the energy they use during the day. Each company may have their own variation on the smart meter, with some offering more functionalities than others.

In a nutshell, smart meters are expected to bring an end to energy bill estimation, used widely in the past, so that both customers and suppliers can have a more accurate understanding of how much energy is consumed or saved through various short-term alternations in energy use. It also offers greater precision in regards to how much customers pay for their energy usage and avoids issues such as overpayment or overestimation of associated costs.

Like any digital device, there may be both anticipated and unanticipated issues with smart meters such as missing recordings or faults of the meter that may miscount the actual energy use. While these are rare, they are still under discovery by energy supplier analysts and some examples of these (i.e. missing reading or smart meter sudden turn off) can be easily seen from smart meter records and detected where necessary.

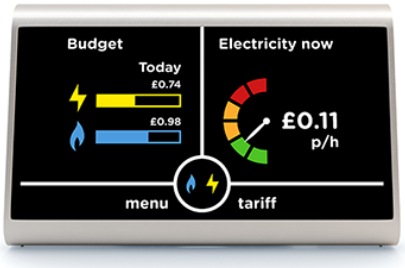


Figure 1.1: Example of a smart meter display *Source: Smart Energy GB*

1.2.1 Nature of the Data

As was shown earlier, smart meter data are commonly assessed using the big data characteristics - volume, velocity and variety - and which effectively characterise its nature.

While there are challenges with smart meter data for both energy company analysts and researchers, it has opened up new possibilities to analyse residential energy consumption. It has been suggested by Swan and Ugursal (2009) that previous research involving energy consumption analysis had tended to focus more on major private sectors that had more incentive and expertise for consumption reduction, as well as operating under tougher regulatory requirements. This may become

less of a problem, with growing numbers of initiatives aiming to bring academia and industry together, such as the Consumer Data Research Centre that sponsored this project. Indeed, motivations for various research projects are expanding to include long term gains for societal wellbeing brought by this or similar research using smart meter data.

Another reason for the smaller proportion of developments dedicated to understanding household energy consumption was the issue of privacy associated with the collection and distribution of detailed data on households (Swan and Ugursal, 2009). In the UK, after installation of smart meters in 2015, energy customers agree by default that their data may be accessed by their energy supplier [2]. However, they retain the right to opt out of their data being shared with energy providers or government agencies. Smart meter users can decide themselves which temporal granularity of data they want to be shared with the energy provider. If they are not happy with detailed information about their consumption being shared, they can choose whether they prefer daily, weekly, monthly or even only annual figures to arrive straight into the hands of energy company analysts.

The UK Government aims to ensure that every domestic and non-domestic property will have been offered a smart meter by 2020. The regulatory environment encourages providers to roll out installation as quickly as possible to meet the obligation of complete installation by 2020. By the first quarter of 2017 there were a total of 6.78 million smart meters installed by energy suppliers across residential and business addresses in the UK; of which six million had been installed in domestic properties by the Big Six energy providers (EON, British Gas, EDF Energy, Scottish Power, SSE and npower). Smart electricity meters account for more than half of the total of these installations due to electricity being more widely available than gas. The Department for Business, Energy and Industrial Strategy (BEIS) (2017) reports that despite an acceleration of smart meter rollout, most domestic properties nevertheless still have traditional meters.

[2]This was true for 2015, year that corresponds to the time frame of the data collected and the time when this research project was initiated.

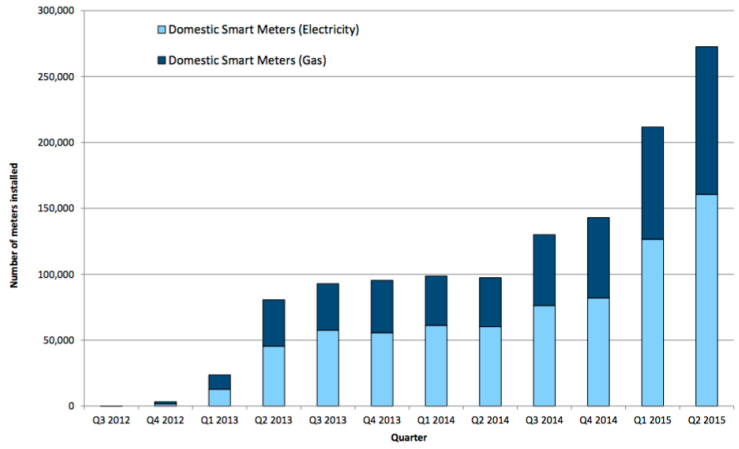


Figure 1.2: The speed of the smart meters roll out by last quarter of 2015*Source: Smart Energy GB*

Figure 1.2 illustrates the speed with which smart meters have been installed in domestic properties by large energy suppliers in the UK and the break down of meter type (electricity or gas) reported by the beginning of 2015. Traditional meters remain a dominant technology for gas (97% of meters) and slightly less for electricity (95.9%.). By the first quarter of 2015 there were only around 250,500 smart meters installed since the start of the rollout. In general, a survey conducted by BEIS (2015) found that the immediate perception of smart meter installations and user experiences was quite positive; with an average of about 80 per cent of the circa 2,000 surveyed customers expressing satisfaction and a sense of being well informed on how to use and benefit from the technology.

Smart meter introduction may be promising in enriching the analysis of energy consumption in the UK and acknowledgement of this can be clearly seen from the statistics of the Smart Meter roll out. However, it may still be a long wait before smart meters replace conventional technology for energy use recordings. This is driven by the fact that some properties may not be suitable for smart meter installations and by the fact that there is a number of vulnerable customers who may need extra support in getting the value out of this technology (i.e. visually impaired

customers or customers with lower mobility). Nevertheless, where data from smart meters is available, greater consideration must be given to analysis. This will be explored in more detail in the rest of the thesis.

1.3 The Thesis Aims and Data

The aims of this thesis are threefold. Firstly, it assesses the methods that are suitable for grouping patterns of energy consumption to provide a better metric for describing the temporal profile variability among energy customers. Secondly, once the patterns that consumers form are understood, these are extrapolated to create temporal profiles to further discuss whether the dynamics of energy consumption can be predicted given the periodicity of past consumption. Lastly, the thesis looks largely at issues of spatial heterogeneity, the uncertainty that may arise from sample selections and aggregation levels of energy data which then can be categorised as external bias (i.e. the data source) and internal bias (i.e. heterogeneity among individual records). The issues of selection bias, data quality and the usefulness for answering various research questions are discussed throughout each chapter so as to remain critical and cautious of many issues, the size of which may grow in parallel with data magnitude.

The data analysed in this thesis fall into five groups. The first group of data correspond to the largest dataset that was available for this research. It consists of about 400,000 meters that have annual records for gas and electricity and are available at high temporal resolution (half-hourly readings) but low geographical resolution (Postcode Sector). The second group of data consists of the Bristol dataset that has records for about 2,000 meters at similar temporal resolution but slightly narrowed geographical resolution (Census Output Area). The third group of data consists of aggregated data derived from a large sample in order to reduce the magnitude of the raw data. For more than 8,000 postcode sector levels, half-hourly averages are used to determine the spatial distribution of the smart meters and then to investigate any variations that might be associated with the geographical locations of the customers. This data is also used for clustering of energy use and the thesis will

look at the effects of such aggregation for final differentiation of energy use. The fourth group of data consists of individual smart meter users picked up from the Bristol sample, which are used for more detailed analysis of yearly consumption through regression analysis. For this group, a number of feature transformations are also used to carry out a spectral analysis. Lastly, the fifth group of data constitutes a sample on customers that have additional qualitative labels . Available at limited resolution due to privacy concerns, this data is used for experiments that determine whether smart meter data alone may predict customer characteristics such as financial vulnerability with respect to the costs of energy.

It is important to note the time period over which the data used in this work was collected, mainly from the years 2014-2015. Relevant government reports and population characteristics are selected from a similar time frame to minimise bias that may arise from any subsequent changes in population or smart meter rollout figures. The main data is described in more detail in Chapter 3, while samples that are used for more specific experiments, such as the third, fourth and fifth groups of data, are discussed in the chapters that focus on their analysis.

1.4 Applications

It is important to note that work in this thesis is driven by the nature of the data rather than by particular research questions. Social science research often starts with a question in mind. Given that question, the researcher then collects the data that can answer it. The primary task of this thesis is to define which kinds of questions can be answered with smart meter data or, in other words, what the data may be used for? Given the novelty and very recent availability of this data, one can see how critical assessment of the usability of this data for researchers may serve as a contribution to the energy and social science research fields in its own right. A number of possible applications associated with the analysis of energy consumption in the thesis are suggested below. For instance, the results presented in this thesis work towards increasing understandings of how stable or predictable British consumers' behaviours, on average, are in terms of their energy consumption. Understanding

predictability on an individual level may improve the forecasting of national consumption as well as give a more precise idea of the confidence intervals from which uncertainty is derived when it comes to energy consumption estimates. The thesis further looked at whether consumption profiling may be achievable with this data. The remaining part of the thesis is centred around forecasting methodology and data reduction techniques. All of these may play an important role in the research agendas that focus on the tasks such as :

- a) lending an insight into national energy consumption volume and scale;
- b) contributing towards an understanding of the energy resources required for moving towards renewable pathways for energy provision;
- c) gauging the scale of the differences among energy customers;
- d) providing a framework of for studying how much pressure the average energy consumer puts on the national grid, or in other words, how much energy is needed to supply the average UK consumer every day at peak hour;
- e) identifying energy customers who consume less than expected on average so as to provide a necessary support ;
- d) designing strategies that can help study the periodicity of energy consumption to gain an insight into how predictable smart meter users are.

1.5 Thesis overview

The thesis is organised into eight chapters in total with Chapter 1 being an introduction followed by the seven chapters outlined below.

- **Chapter 2 (Literature Review)**: An overview of the energy consumption pre determinants and associated with these uncertainties are presented in this chapter. These are further complemented an account of the research which has been done, up to date, on understanding energy consumption variation using both smart meter data and data from other sources. One of the most

pressing and challenging issues that current research attempts to address with energy data is fuel poverty. Current research in the area of fuel poverty and debates around its definition and measurement are also discussed. The literature on methodology and analytical strategies previously used to study energy use is presented along with a discussion of which methods can, in fact, be borrowed from other disciplines (i.e. computer science).

- **Chapter 3 (Data)**: The data available at the national scale is introduced in this chapter, together with the sample on Bristol from the Census Output Area and the National Data aggregates at a Postcode Sector level. A third, experimental sample that used in the thesis to study fuel poverty is also presented briefly in this chapter. Basic visualisations and descriptive statistics are provided with the reference to official statistics. Discussion on what this data may be used for and its possible limitations are given in this chapter. In addition, this thesis section looks at the issues of aggregation for energy data and selection of unit of analysis is discussed along with an introduction to some other approaches that may help us to describe the data while avoiding excessive generalisations.

- **Chapter 4 (Methodology : Clustering/Load Profiling)**: This chapter looks at the various methods and metrics available for grouping temporal profiles. The suitability and limitations of both are discussed and the methods are compared in terms of their accuracy and reliability. The chapter concludes with a treatment of the issues of heterogeneity, bias and uncertainty that are associated with different types of analyses when applied to energy data. At this stage preliminaries on probability, Gaussian processes and smart meter data as a time series process are introduced.

- **Chapter 5 (Methodology: Regression Analysis of Smart Meter Data and Forecasting)**: This chapter looks more closely at energy dynamics and how they can be described using generalised additive models. Various subsamples of data were selected for regression analysis and studying the periodicity of energy consumption. Such patterns are associated with the relationship be-

tween certain hours of the day and regular or cyclic behaviour. Consideration is given to the potential to rebuild the data, thereby undermining the data series structure. Additionally, this chapter addresses the issue of seasonality in the data and attempts to recover the estimated time of consumption by looking solely at the variation.

- **Chapter 6 (Methodology: Label Prediction)**: In this chapter, the incorporation of other variables is examined to define how useful they may be in understanding what contributes to differences in energy consumption among individuals. The chapter attempts to answer a very specific question: can energy consumption vulnerability be identified using smart meter data? If this is not so, then which additional data is required and which methods may be useful in answering this or similar questions? This chapter looks at defined customer label prediction using solely smart meter data. Limitations and opportunities associated with such prediction are discussed in a wider context of operational and public policy research.

- **Chapter 7 (Scaling Up: Data Reduction and Transformations Techniques for Smart Meter Data)**: The issues of scaling up analysis are addressed through an introduction of methods from the area of signal processing known as spectral analysis. Examples of such analysis include Fourier and Wavelet transformations. As will be seen from the thesis, while one may be lucky enough to have big data on smart meters, processing and analysing this data in one turn is quite challenging, if not impossible at this stage. Data reduction and compression of the series such that it can be presented with fewer features yet can still hold all the vital information about the uniqueness of the pattern will be presented in this chapter.

- **Chapter 8 (Conclusions)**:This thesis concludes by investigating future research paths where smart meter data research design may target consumers and survey their behavioural habits rather than details on who they are or where they live so as to understand what drives the residual or uncommon

behaviour in the data that cannot be possibly captured by the quantitative approaches presented in this thesis. The methodology is summarised and an overview of the observed results and possible contributions is given.

Appendix: Note on Terms and Definitions

Pattern

The word ‘pattern’ is used in the thesis to refer to the pattern of consumption that is represented by the sequence of half-hour readings. Such is different from the use of word ‘pattern’ when referred to the behavioural patterns used in statistical literature that can be found across and within readings over time using solely statistical and pattern recognition methodologies.

Behaviour

The term ‘behaviour’ used throughout the thesis is a primary tool to describe temporal behaviour that is presented by energy consumption records. It is therefore subjective to only one dimension which is available from smart meter’s records and can suffer from biases when used to describe household behaviour and their activities as there may be way more going on behind the scenes in the household or at the address to which smart meter is attached.

Segmentation

’Segmentation’ is used as another way to represent the ways in which energy consumption readings can be split into distinct groups using clustering or classification methodologies. The word ’segmentation’ used here to highlight another type of grouping that can be applied to units of analysis such as those that refer to their socio and economic characteristics. This is different from relying solely on a statistical algorithm.

Random

The word ’random’ is used fairly frequently in this thesis to define the selected readings for visualisation and analysis. Definition of ’random’ needs to be narrowed to that fact that the sample of data we are dealing with may be non-random on a population scale, but the random selection of energy readings are taken from this non-random sample. Given the magnitude of the dataset, it may be fair to assume that random selection of patterns may be fairly representative of larger sample

diversity, yet this is something that will be challenged in the subsequent sections of the thesis.

Fuel Poverty and Vulnerability

'Fuel poverty' is perhaps the most challenging concept to define in the current energy research literature given the disparities of definition among academics looking at the concept from various angles and disciplines. More details on the use of the concept will be given in Chapter 2 (Literature Review) and Chapter 6 (Classification). In this thesis, there is a number of obvious limitations on how fuel poverty and vulnerability could be quantified. These arise primarily due to data quality reasons as well as ethical considerations that are associated with preserving the anonymity of individual users.

Chapter 2

Literature Review

'...those of us who call ourselves energy analysts have made a mistake . . . we have analyzed energy. We should have analyzed human behavior'

- Schipper in Cherfas (1991),

2.1 Introduction

Energy may be compared to economic wealth when thinking of its pre-determinants and variations. As an important aspect of the subsistence of human beings it can be used differently depending on the overall wealth and economic preferences that individuals may have. Furthermore, energy product value, such as the price of gas or energy tariff, may have a similar effect on the quantity of resource consumed. This chapter looks in more detail at the various factors that could contribute to variations in energy consumption. They include household socio-economic characteristics and various property attributes. Cultural (Lutzenhiser, 1992) and social environments that may affect consumption intensities are also touched up briefly. that may affect consumption intensities are also touched upon briefly. Empirical studies of the connection between energy consumption and peoples lifestyles also deserve consideration (Druckman and Jackson, 2008; Druckman et al., 2011). All of this is presented with the aim of describing energy consumption as highly complex and integrated from various factors and processes that go well beyond simple smart meter readings. As an introductory visualisation, in a broader sense the dia-

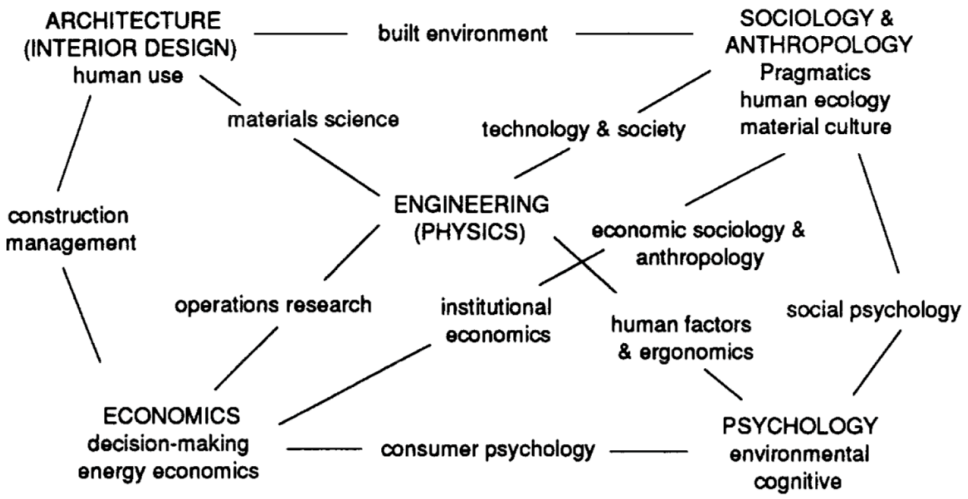


Figure 2.1: Disciplines involved in the study of human/ environment and human/technology relations (Lutzenhiser, 1992)

grams constructed by Lutzenhiser (1992) and Lutzenhiser et al. (1997) give an idea of the interconnectedness among disciplines that may be useful in attempting to understand energy consumer behaviour Figure 2.1 presents a number of disciplines that are involved in the investigation of human vs environment and human vs technology relationships. With engineering and physics undoubtedly being central and essential to energy production and supply, the use of resource is determined by economics (decision-making process) , psychology (environmental awareness as well as consumer attitudes and beliefs towards risk and behavioural norms), sociology (material cultures and technology awareness) and lastly, architecture (human use of buildings and how properties are designed).

The importance of the interdisciplinary approach, which this thesis cannot emphasise enough, when dealing with energy data is driven by the heterogeneity of the consumption that is observed when looking at different smart meter users. As will be shown throughout the remainder of the thesis, a large amount of the residential variation that would not meet the criteria of the expected or typical temporal

profile remains ambiguous, unexplained and in need of more interdisciplinary research analysis that could integrate spatial, temporal and social components. With an aim to find a middle ground for generalisation and interpretation of the temporal differences, this chapter reviews the literature that analysed energy expenditure and consumption pre-determinants in the past and looks briefly at consumption trends described in the UK government overview reports to provide an idea of the context in which smart meter users analysed in the thesis may reside and be shaped by.

The first part of the chapter will thus be dedicated to mainly pre-determinants of energy consumption variation previously found by researchers across various disciplines including public policy and engineering. Starting with the discussion of typical or expected profiles of energy consumption, the chapter then moves on to discuss how various socio-economic characteristics of smart meter users and attributes of the properties in which they reside can shape the dynamic and magnitude of residential energy load. Fuel poverty identification as one of the prevalent issues arising from slight outlier behaviour and ways in which energy is consumed is further discussed to provide a background for the case study the thesis will return to in Chapter 5. The second part of the chapter is dedicated primarily to methodological advances in the energy research that uses smart meter data. Some examples of methodology from other disciplines that were designed for similar types of data will also be reviewed. The methodology review will have a more centralised focus on various strategies that were developed and surveyed with an end goal of customer characterisation, outlier detection and effective forecasting of energy use. These sections will set a preliminary stage of the analysis in the remainder of the thesis which is, as will be observed, primarily methodological.

2.2 Typical Profile

The availability of smart meter data may give a greater precision for understanding the differences in temporal patterns for households that may have similar total energy consumption but a very different way of consuming. However, it also creates further challenges when it comes to the generalisation of temporal profiles. How

can one know what is the average or expected energy consumption profile? Would it depend on location of consumers or their property or household characteristics?

'The variability in residential consumption reported in the literature suggests that there is hardly a 'typical' level of consumption for any energy end-use" (Lutzenhiser, 1993). Nevertheless, a typical behaviour derived from the data is often assumed for benchmarking and anomaly detection. This behaviour also underpins the design of energy efficiency measures. An example of a double-peaked temporal profile is shown in Figure 2.4. Often, to derive such a shape and to represent the most common trend in the consumption of an individual or a group: mean or median consumption are used. The profile below can often be described as a full time employed smart meter user, leaving home early in the morning to come back around 5-6pm and resume activities in the household that are associated with energy use.

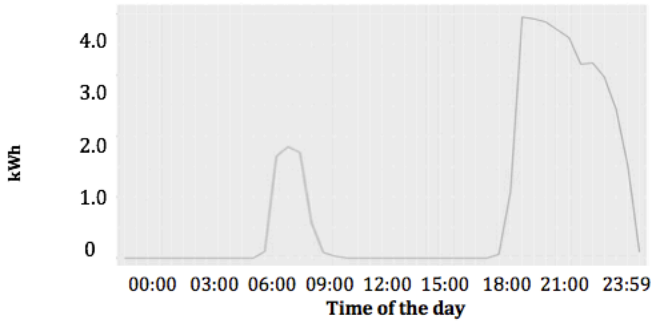


Figure 2.2: Example of median consumption derived from the smart meter data *This shape was derived using a sample of data available for this research*

The UK Housing Energy Fact File 2012 produced by DECC (2012) reports average electricity use broken down for appliances that represent quite a similar to above shape. Besides, on the individual household level variations are expected to be more heterogeneous even when 'typical' shapes of consumption are observed. DECC (2012) further has classified the factors that may affect household energy use as: (a) related to investment decision making (for example home upgrades or purchase of new appliances); (b) infrequent actions (for example temperatures to be set in the rooms or settings for running the appliances); (c) repeated actions (for example taking a shower or standby appliances use) and (d) spontaneous reactions

to the events (for example reactions to extreme weather events or using lights in the rooms).

Property type, household income and tenure have been reported as being able to explain only forty per cent of the variation in the gas consumption among households in the UK (DECC, 2012). Extreme cases of very high and very low consumption profiles are not easily identifiable. These extreme behaviours, as suggested by DECC (2012), could be classified using the following clusters - physical properties of the houses such as additional extensions, conservatories, open plan spaces, consideration of how the temperature is managed, and by looking at people at home. Further to that analysing who is present, and when and which types of activities are predominant in the house may help to understand a bit more about outlier behaviours present in various consumption readings.

One may think of some further classifications of energy usage such as heating, lighting, entertainment or comfort. Furthermore, the minimum and maximum expected energy use could be observed for different types of households, living in various property types. To highlight the broader research that investigated most of these factors, the next sections overview some of the major pre determinants of diversity in energy use.

2.3 Spatial, Temporal and Social Determinants of Energy Use

Figure 2.3 provides a useful overview of the collection of research that has attempted to explain energy consumption. As may be seen, historically, dwelling type, appliances holdings, household size and income tend to dominate in terms of the attention that they receive as pre-determinants of energy consumption variation, in particular, for electricity. United under the umbrella of the physical- technical economic models (PTEM), physical attributes of buildings have been used extensively in energy demand forecasts and policy in the past (McLoughlin et al., 2012a). Through an understanding of how these dominant factors contribute to the intensity of energy use, we may further inform the policy of the best strategies to alter con-

sumer behaviour and increase energy savings. Besides, such a skewed distribution of research focus (centred mostly around studying property attributes and household income) may be driven not by the realisation that certain factors are less important than originally thought but by unavailability of the data or limitations posed by the covariates that are hard to quantify such as social environments, habits and culture. Some of these factors are discussed below.

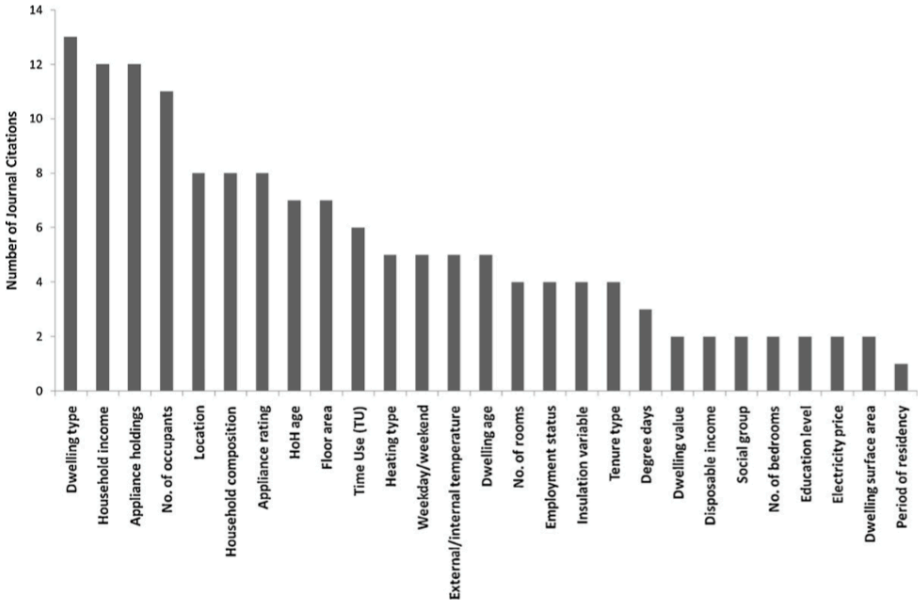


Figure 2.3: Number of citations looking at property attributes and household characteristics as explanatory factors for domestic electricity consumption (McLoughlin et al., 2012a)

2.3.1 Financial incentives

Perhaps one of the most straightforward factors affecting energy consumption is the price of energy. Meier and Rehdanz (2010) 2005, and revealed that heating behaviour is affected by energy price increases but may also be affected by policy measures, such as those that target carbon emission reduction. To show this they used a regression analysis with heating expenditure being the dependent variable. They further tried to define how different household types would respond to price and policy modifications (i.e. the elderly are associated with lower heating expen-

ditures). Research has suggested that policy may consider focusing more on rural households as they tend to spend more on heating. Nevertheless, lack of information on dwelling and built environment characteristics was acknowledged by the authors as a priority for further research. Price variables when considered as an influencing factor also need to be studied alongside knowledge and information about pricing in real costs and in regards to price elasticity. A study of consumer behaviour may be performed at both individual and household level. However, depending on the type of data that we are dealing with, the household level may be far more accessible, especially when working with consumer data, which tends to be associated with one registrant representing the household (i.e. the energy customer agreement). Thus it may be more reasonable to take a household as a unit of collective consumption, accounting for differences in how energy price may affect a consumption pattern for the whole household rather than an individual. Information or knowledge about the energy market that customers have may also play an important role for determining their behaviour (Lutzenhiser, 1993). Various studies further looked at the effect of 'feedback such as recent consumption for example, as well as the effect of direct information delivery to the customers through their community role models or local networks (Winett and Ester, 1983). Lastly, economic psychologists and behavioural economists tend to omit further complications associated with the marketing system, limiting their approach to the effects of price, while consideration of the effects of non-price elements of the marketing mix such as advertising and sales promotion is considered to be influential (Alhadeff, 1982; Herrnstein, 1988)). It has been suggested by Bolton (1998) that apart from the price, satisfaction with the service is no less powerful in altering consumer attitudes/beliefs, and as a consequence may lead to certain home improvements that will affect the ways in which service is used and have an impact on consumption continuity.

2.3.2 Consumption environments

In addition to associated costs, patterns of domestic gas and electricity consumption may often be driven by constantly changing consumption environments such as temporary guest visits, sickness, change of employment shifts or change in the

size of the household. This would be particularly important when we consider the dynamics of fuel poverty. An example of how such a change could be measured was presented by Bernard et al. (1988) , who studied variations in natural gas consumption introduced through occurrence of events or non-routine activities at the household property. They differentiated between the components of consumption that sum to total consumption and suggested that each customers consumption has three components: (a) structural consumption; (b) habitual consumption, which may further be complemented by unconscious habits (Lutzenhiser et al., 1997; Wilhite and Wilk, 1987; Hackett and Lutzenhiser, 1991) and (c) daily variation consumption (i.e. holidays, sickness, having visitors).

They further suggested that to understand variations in consumption between households with more variegated profiles, it is necessary to study not just daily but also weekly patterns. This may inform us about the temporal pre-determinants that could be inferred by studying a longer period of time. A simple example of this could be the extensive use of energy for preparation of food a few days before hosting visitors. Studying the time before and after an event taking place in the household may give opportunities for a more accurate inference about the level of change in energy use that occurred because of a particular event. As a consequence, when modelling such behaviour, researchers may test a hypothesis where there is a certain level of structural or fixed consumption for each household as well as a component which is variable.

2.3.3 Dwelling characteristics

The effect of dwelling characteristics has been analysed in more detail by Nguyen and Aiello (2013), who showed that occupancy activities and building characteristics have an impact upon consumption in both residential and retail sectors. They also looked more in depth into differences between household activities and how those could affect households' choices of appliances that seem to have a very direct relationship with the dwelling type. Furthermore, Guerra-Santin and Itard (2010) concluded that despite dwelling attributes, the temperature in the housing stock may contribute to further diversity in the variation of consumption during the heating

hours.

In the UK, to quantify the energy efficiency of domestic buildings, primarily Energy Performance Certificates (EPC) are used together with the Standard Assessment Procedure (SAP) ratings. They provide an extensive overview of the property with an attached rating which ranges from A (most efficient) to G (least efficient). EPC data is quite a useful starting point for analysis of housing stock, as there is quite a strong relationship between EPC band and both household energy efficiency and the probability of being fuel poor, which will be discussed in more detail later. The EPC band also helps to differentiate between rural and urban areas. Additionally, there is also an interesting relationship between EPC band and health indicators. For example, it has been suggested that the most inefficient and hard to warm houses in the bands G and F also tend to be quite unhealthy for individuals (Boardman, 2010). Interestingly, it appears that there is some kind of vicious circle for inefficiency. Those who are likely to live in fuel inefficient houses are expected to live in rural areas and to have old large properties that are the hardest to heat. More efficient and healthy housing would be located rather in the city centres, would be less affordable and may be associated with more intense energy consumption.

2.3.4 Income and wealth

Sovacool (2011) performed an extensive analysis of how urban households energy consumption could be conceptualised using income divisions such as low, middle and upper income groups. He further identified the differences that may be observed in energy consumption due to the urbanisation and direct vs indirect uses. As may be seen from Table 2.1, low income consumption tends to be associated with the minimal levels of consumption that are necessary for subsistence energy use such as cooking, heating and hot water. Middle income groups may have spend larger proportions of their energy consumption using electrical appliances such as TVs, computers, washing machines. High income groups have multiple appliances and larger spaces for heating or cooling. This group may also have extra space , such as gardens with lighting and heaters outside the property, that require further energy consumption for use and maintenance.

Table 2.1: The urban household energy services ladder. Adapted from Sovacool (2011)

Household Type	Primary fuels	Primary technologies	Primary energy services	Broader Driving Factors
Low Income	Wood, dung, kerosene, charcoal, coal, biomass, liquefied petroleum gas, paraffin, candle wax, bio gas, agricultural waste, diesel, coconut oil, sunlight	Cook stoves, open fires, candles, solar cookers, small solar home systems	Cooking and lighting, occasionally television, telephone, radio, mobile phone charging, space, heating, refrigeration and hot water	Satisfying subsistence needs
Middle income	electricity, natural gas, coal, liquefied petroleum gas, kerosene, fuel oil	Large solar home systems, televisions, radios, DVD players, air conditioners, refrigerators, water heaters, dishwashers, clothes washing machines, computers, printers, other modern appliances	All low income services plus some heating and cooling, hot water, cooking, entertainment, and lighting, refrigeration and freezing, clothes washing and drying, computing and surfing the internet, watching television, advanced telecommunication	Convenience, comfort and cleanliness
High income	Electricity, natural gas, fuel oil	multiple air conditioners, refrigerators, water heaters, dishwashers, clothes washing machines, other advanced appliances	All of the middle income services above plus luxury practices such as swimming in a heated pool, going in the bathroom with a heated toilet to the sound of music, and watching television while one cooks	Conspicuous consumption and social signalling

2.3.5 Household type and size

While household type and size may influence consumption dynamics, energy consumption may also be informative of the household size in samples where data is lacking details about smart meter users or where there is a sudden household expansion. As was presented by Ushakova (2015) we may often observe differences in yearly consumption patterns between families, single owners and young house shares, with families being less variable and young house-shares being rather random in the nature of their energy use. The study ofGuerra-Santin and Itard (2010) also showed that the use of heating and ventilation might be different for different household types. For instance, elderly customers may have longer hours of heating with few hours of ventilation, while families with children are more likely to use

longer hours for ventilation yet use less energy for heating.

2.3.6 Occupation

Sociological research has suggested that occupation, and more specifically specialisation of the household, may have an impact on the energy consumption behaviour and attitudes toward sustainability and efficiency. In the case of fuel poverty, it was suggested that unemployment is one of the potential covariates of fuel poverty (Boardman, 2010). To give an example of more trivial implication, working hours may be inferred from visibly regular absence of the household. Employed household for instance was successfully inferred from the energy patterns in Beckel et al. (2014) by using multiple linear regression.

2.3.7 Health

The relationship between energy and health is relevant when considering energy vulnerability and winter cold-related diseases. While this is discussed more thoroughly in the section 2.4 , only some of the immediate relationship between energy and health are considered in this subsection. These are differences in the health of the population that may affect the ways in which energy is consumed. Yet, the way energy is consumed may adversely affect health as was seen in the section that discussed dwelling characteristics. For large energy companies, one of the obstacles to smart meter roll-out is the health status of the households. If the customer is blind for example, such a condition may bring complications for both marketing campaigns as well as maximising the benefits that can be brought by smart meters, including other energy services products that can be used at home.

2.3.8 Geography and culture

Bouzarovski et al. (2014) conclude that to date the academic literature has focused little on understanding how domestic energy provision is regarded to be sufficient in different cultural and geographical settings. Thus, these limitations may serve as a potential fundament for this research as it attempts to address, for example, how differences in geographical location may impact upon consumption classifications. Adding the dimension of climate and weather would also offer us the potential to

enrich explanatory power of the models we may use to describe energy consumption variations. Later on it will be shown that geodemographic classification such as Census Output Area Classification (OAC) may provide some guidance into why consumption may differ across some areas of the UK. This relationship will be rather suggestive as it is quite challenging to gauge any causal relationships given the limitations of the sample and the bias which is inherited from the way smart meters are being rolled out.

2.3.9 Society and behaviour

The final factor affecting the variability of energy use is concerned with uniqueness of behavioural norms across countries and regions. As the UK Housing Energy Fact File (2012) suggests over the past years, energy consumption in the UK has had a tendency to be shaped more significantly by consumer habits and by lifestyle than by household size or dwelling type.

Raaij and Verhallen (1982) pioneering study of a behavioural model of residential energy use provided a basis for understanding energy consumption predeterminants. Their study was based on an investigation of energy demand in the Netherlands. Consumer behaviour types associated with energy were outlined as purchase related, maintenance and operating behaviour and usage related behaviour. Purchase related behaviour has close links with the ways the heating equipment is used by the households, and its relative importance in consumer budgeting. Such usage is related to how households use their home and appliances. This can be described using intensities and frequencies of usage over time. Maintenance behaviour is characterised by ways in which households tend to maintain their household equipment, including servicing and financing repairs and home improvements.

Attitudes formed through energy usage could be further aggregated into price concerns, environmental concerns, health concerns and personal comfort(Raaij and Verhallen, 1982). To look at this more broadly, philosophical perspective of Bourdieu (1984) may be considered. Bourdieu (1984) offers an interesting account of how social practices of people are pre-determined by their environments and wealth. He suggests the idea of *habitus* based on the theory that interaction with capital and

field (i.e. a day to day environment of an individual) would determine the differences in various practices among the individuals in society, in the example of energy use, this would be the variety of energy consumption behaviour.

2.3.10 Summary

This sections aim was to overview broadly the various factors that may contribute to the variation in energy use. The secondary goal was to highlight the variety of the research that has looked into energy consumption and used it as a proxy of individual well-being. Lastly, this section was set up to draw on disciplines that can contribute to a comprehensive understanding of why energy consumption can vary so much, even in cases where the samples are drawn from the same region or particular area of the country.

The research in this thesis aims to find and survey the approaches which may help in segmenting different behaviours across energy customers such that they can be grouped together based on the described earlier factors. It will further look at predictability of these diverse behaviours using smart meter data readings. Given the limitations associated with the sample of data available for this research, it was challenging to explain why energy usage varied across the users in the UK. Nevertheless, understanding of various reasons that contribute to these differences outlined in past research have allowed for some speculation as to why customers may have shown certain patterns of usage. The next section, will specifically look at fuel poor customers and various characteristics that may help in determining those smart meter users that may need financial or social support to be provided by the energy supplier. Being one of the most pressing issues on the UK energy policy agenda, this issue is addressed later on in the thesis by introducing smart meter data as a possible proxy of ones vulnerability.

2.4 Who are the Fuel Poor and Why Do They Need to be Found?

How the energy vulnerable and the customers that are at risk of becoming fuel poor can be found and subsequently supported? This question has occupied UK government and energy suppliers for the past twenty years and has thus been explicitly stated by OFGEM in Energy Company Obligation under the Home Heat Cost Reduction Obligation component. With the pioneering research of Boardman (1998), academics have also started to take fuel poverty under closer attention, making it a separate area for energy research projects concerned with household energy consumption. For this research the definition of vulnerability is aligned with the one that is defined by the UK government and is based on income measures. These will be discussed in more details later in the section. Nevertheless, it is important to acknowledge that even within the UK the definition does vary across various government departments. This became an obstacle for fuel poverty elimination that was targeted to be achieved by 2016 (Boardman, 2010).

This section looks at who the fuel poor energy customers are, how fuel poverty may be defined and in what way fuel poverty is linked to energy consumption vulnerability. Fuel poverty or energy poverty impacts both individuals and the household as a whole when it comes to consuming energy collectively. The effect that fuel poverty may bring on a households wellbeing may well manifest interaction between property attributes and household characteristics; it is also expected to vary spatially. Furthermore, individuals that are considered vulnerable may experience acute impacts of fuel poverty (e.g. customers with disabilities). This section revisits the concept of vulnerability in the UK and its primary characteristics, before moving on to a discussion of vulnerability in the context of energy provision.

2.4.1 The concept of vulnerability

The following section addresses the concept of vulnerability with initial focus on social vulnerability, before narrowing down to the specifics of energy consumption.

As was acknowledged in Kandt (2015) , conceptually focused geodemographic

classifications may inform policy-making and interventions in a structured way, especially when it comes to targeting social groups in need of support. An example of this could be the classification of social vulnerability. Kandt (2015) characterised social vulnerability from a health perspective, and identified that preventive health care is definitely one among many important pre-determinants for group or neighbourhood vulnerability. Health tends to be one of the central determinants of social vulnerability, yet the spatial context and area characteristics may have adverse effects on health and policy interventions.

Simultaneous causal relationships between health and social vulnerability are not uncommon.To consider social vulnerability more broadly, Cutter et al. (2003) underlined a definition of vulnerability as a potential for loss. In the context of energy research, this definition may be transferred to the loss of resources that an individual household may have used for adequate heating. The inherent difficulty of quantifying social vulnerability is often ignored due to ambiguity associated with calculating social costs. Nevertheless, this thesiss research may present an opportunity to use energy as one, amongst many, ingredients of social well-being.

2.4.2 Fuel poverty and its causes

The definition of fuel poverty and the ambiguity that is associated with the concept is considered to be an important aspect in policy formulation as well as policy evaluation, targeting and monitoring (Moore, 2012). According to Moore (2012), the definition of fuel poverty may date back to the 1980s, Brenda Boardman's definition (Boardman, 1998), which relates fuel poverty to the proportion of disposable income that is devoted to energy expenditure. Fuel poverty as a concept was born mainly in the UK and Ireland yet was subsequently considered in other European countries as an indicator that contributes to the wealth of the countrys population. A number of research papers have thus considered the relationship between energy expenditure and the prevalence of low-income households (Foster et al., 2000; Santamouris et al., 2007; Roberts, 2008).

In the UK, a household is defined as fuel poor if spending on energy services as a proportion of their incomes exceeds ten per cent. Severe energy poverty is

attributed to households who spend more than twenty per cent of their respective income. Such a definition may sound very straightforward, but one of the main problems with this measure is the uncertainty in regards to quantifying the income of households.

Income variables tend to be associated with measurement error and bias, often due to common exclusion of other forms of capital held by an individual. These can be reflected in terms of savings or inheritance, property ownerships to name but a few (Boardman, 2010). Another factor which is complementary to energy services spendings is an overall household expenditure that is composed of rent or maintenance costs. Housing benefits may be able to soften the impact of these on household risk of becoming fuel poor. Likewise, specific payments like winter fuel amounts may improve the circumstances of potentially risky households.

To reduce the error associated with solely using income as an indicator, research on fuel poverty has started to shift gradually towards more non-income identifiers of fuel poverty, greatly emphasising the fact that those who are income poor are not necessarily fuel poor. However, low income remains an important ingredient in estimation, as the interaction of income for example with energy inefficient housing will be a valuable pre-determinant of fuel poverty. Such a combination was proposed in Boardman (2010) : **Fuel poverty = inefficient housing + low income + high energy price.**

For Boardmans formula, there is no need for smart meter data to be included to reveal the probability of risk of becoming a fuel poor customer. Nevertheless, the obvious issues is that the data on all three components is rarely available as a linked combination at the level of the individual household and dwelling. It is possible to observe the type of housing and income group of the household at the expense of keeping their energy bill and pattern of consumption unknown. For energy companies, prior to installation of smart meters, keeping a record of household changes was not as easy and it would normally be limited to the information received when the household became a customer. Consequently, the customer databases available to energy suppliers will lack an updated record of the changes that may have oc-

curred over the customer journey, making it incredibly hard to find ways to identify potentially vulnerable energy customers.

To address the issue above, some researchers have attempted to consider factors other than income. For instance, fuel poverty and energy customers vulnerability were addressed and discussed in Bouzarovski et al. (2014); Hills (2012a); Legendre and Ricci (2015); Middlemiss and Gillard (2014); Sefton (2002); Boardman (2010) and Rosenow et al. (2013). Rosenow et al. (2013) has provided a very substantial and critical assessment of fuel poverty, as defined by policymakers in the UK. The authors used various analytical frameworks provided by the UK government, and consultations and statements by energy suppliers, supplemented by fifteen interviews with main representatives of governmental and non-governmental energy organisations in the UK. A history of the framework as well as an assessment that can be of use to energy suppliers were provided. Rosenow et al. (2013) have also suggested various methods for estimation of fuel poverty using non-income measures such as property characteristics. It was further concluded that measures of fuel poverty would greatly benefit from the inclusion of customers geographic location too.

In the study of Moore (2012), variables on tenancy have shown a correlation with fuel poor indicators. It was observed that private rental sector or housing associations residents tenants are more likely to be fuel poor due to greater house inefficiency. Furthermore, it was reported that approximately fifty per cent of all fuel poor households in England were elderly, singles and couples, while seventeen per cent of the fuel poor are those with children. In addition to that, paper of Legendre and Ricci (2015) estimated the scale of fuel poverty in France using different definitions of the phenomenon such as issues related to income or energy inefficiency. Using various econometrics models authors have found that the greater probability of being fuel poor highly related to customers being retired, living alone, rent their home, cook with butane or propane or have poor roof insulation. These variables may thus be no less important for designing a fuel poverty metric.

Middlemiss and Gillard (2014) have provided a further contribution to research

on fuel poverty and understanding of what the main drivers of energy vulnerability are, specifically for the UK population. They performed a qualitative study, results of which have shown that among other factors, energy costs, health, social status and income stability may impact the risk of being fuel-poor.

2.4.2.1 Urban and Rural

To add on this, work of Roberts et al. (2015) emphasised that the differences that may be observed in fuel poverty across the UK can be explained using urban and rural comparison. When looking into predictability of energy consumption in more detail later in the thesis, this distinction will become more apparent. While research often tends to suggest that rural areas are more associated with fuel poverty because of the structure and the character of rural housing stocks or limited connection to the grid, fuel poverty in urban areas tends to be more persistent and has lasts longer on average than in rural areas. Often, this may be caused by the fact that rural fuel poverty may be easily fixed with efficiency measures such as insulation or boiler replacement. In urban areas, on the contrary, fuel poverty tends to be associated with income which may be harder to modify at least in the short term, yet can be somewhat addressed with the introduction of financial support by energy suppliers (i.e. the Warm Home discount).Nevertheless, to support the claim that rural households are also likely to be fuel poor, researchers have shown that the fuel poor in rural areas are more sensitive to changes in energy prices. The authors suggested that monitoring how fuel poverty changes over time is vital if we are interested in the effectiveness of targeting the fuel poor and ensuring that policy goals for fuel poverty reduction are achieved (Roberts, 2008). To support Roberts (2008),Walker et al. (2012) performed an area-based study on targeting fuel poverty in Northern Ireland. Significantly clustered areas were recognised using the Moran's I coefficient of spatial autocorrelation. The majority of those at high risk of fuel poverty were identified, as may be expected, in rural areas (Walker et al., 2012). Besides, it was noted that there is higher variability of risk of fuel poverty among smaller geographical areas (i.e. at neighbourhood level). Low risk neighbourhoods may be part of a broader high risk area and vice versa. This could be due to there being different

concentrations of households in adverse circumstances in smaller areas (e.g. a larger number of the elderly in privately rented accommodation). The authors have thus emphasised that it may be important to focus on studying the demographic outliers as well as considering a more holistic methodology to study energy consumption. Shared engagement in this process by different sectors may offer more possibilities to aid the policy implementation strategies through data sharing, for examples see Walker et al. (2012).

2.4.3 Summary

To summarise this rather brief but nevertheless relevant fuel poverty discussion, it is important to consider a trade-off between understanding and modelling individual circumstances and the wider political implications of fuel poverty analysis. As pointed out by Hills (2012a):

'With any practical approach to tackling fuel poverty there will be some households who are assisted that do not come into a strict definition of what fuel poverty means. In reality, policies have to have a broad spread and cannot be designed to adhere narrowly to precisely drawn boundaries.' (Hills, 2012b)

The above has been widely supported by current research findings. Moore (2012) similarly contends that actual expenditure is a poor indicator to use. In order to estimate the required expenditure one needs detailed knowledge of the housing stock, and its energy efficiency, as much as these are needed for explaining any pattern of energy use. Hills (2012a) argued in favour of collecting this information, emphasising that it is important to ensure that all households are comparable. Equalising households and standardising the data means putting energy users on the same scale by taking into account the different energy service needs for different household sizes [1]. While sounds straightforward, to achieve such equalisations is no less challenging than explaining what contributes to an individual energy use patterns. This section and the preceding one aimed to highlight the overarching

[1]The OECD has a standard rule for equalisation, where the first adult in the household counts as 1, any additional adults count as 0.5 and children count as 0.3. This formula takes into account the fact that there are economies of scale of having more than one person sharing a household. That is, two people consume less than twice as much as one single person.

complexity associated with classifying energy use and defining various customer groups, for instance those who are fuel poor.

The next section will turn to the methodological debate. Given the pressing issues presented by the field of energy research, the next question to ask is how smart meter data can fill the gaps in the current research and give both governments and energy company analysts more tools to analyse energy use with more certainty. As will be shown, much of past research has dedicated attention to load profiling and customer segmentation as a stepping stone for suggesting how various individual circumstances may be explanatory of diversity in energy use.

2.5 Methodology for Smart Meter Data

'Life is really simple, but we insist on making it complicated.'

- Confucius,

The past decade has seen a notable expansion of smart meters across the Western world. This has led many researchers and industry practitioners to develop and survey a vast number of analytical tools that could help in segmenting smart meter big data; use it for forecasting of the energy load as well as for designing strategies to compress the data so that larger amounts of real time data can be analysed. The main target behind these methodological advances is certainly oriented around leveraging as much value as possible from the available data, thus aiding demand side management practices and assisting achievement of energy efficiency targets, which include support for energy vulnerable customers. Smart meter data, when available in large volumes, can be described as complex data, yet this does not necessarily imply that the most complex and advanced methodology is required to analyse it. This section will review the methods and techniques that are thought to be useful or were applied in the past to generate insights from smart meters. It will be argued that a compromise between advanced methodology and interpretability needs to be taken into account when choosing the right approach to study this data [2].

To draw on widespread ambitions to analyse and produce value from smart meter data, a summary of smart meter focused analytics initiatives from across the world are presented in the Table 2.2, partially adapted from Wang et al. (2018).

[2]This part of the chapter mostly sets the background for Chapters 5,6 and 7. The overview of applications directly related to the methodology used in the thesis will be presented in the corresponding chapters

What	Where
National Science Foundation(NSF)	USA
CITIES Innovation Center	Denmark
The Bits to Energy Lab	ETH Zurich, University of Bamburg, University of St. Gallen
The Siebel Energy Institute	Global
National Science Foundation of China (NSFC)	China
Energy Institute	UK
ESSnet Big Data	Europe (Austria, Denmark, Estonia, Sweden, Italy, Portugal)

Table 2.2: Smart Meter Data Analytics Initiatives *Adapted from Wang et al. (2018)*

The 'elephant in the room with smart meter data research is the lack of attention paid to processing the big data arriving from smart meters. Most of the studies, as will be shown, base their work on small samples. In addition to that, the data remains unintegrated with other time series data that could enrich the analysis of spatiotemporal energy patterns (Wang et al., 2018).

Given the above, the analytical techniques that are developed to study smart meter data as big data can be divided into the following streams: (a) outlier/missing data detection ; (b) load profiling ; (c) load forecasting and (d) data reduction. This chapter will revise each briefly. Some other, more narrowed streams of methodological research on smart meter data consider issues such as uncertainty and variability as well as understanding the nature of the data generation process that hides behind smart meter data. More details on applications will be given in the subsequent chapters when implementation of various load forecasting, profiling and data reduction techniques will be applied to the data.

2.5.1 Clustering and load profiling

Clustering and data segmentation methods can be handy for reading energy load when there are cases of missing data and where one is interested in identifying energy theft. Other and, perhaps, more common uses of clustering methods are mostly dedicated to characterisation of the load profiles such that various energy consumption behaviour can be segmented or classified. These methods can be further di-

vided into directed and indirect clustering approaches. The latter refers to cases where data reduction or transformation techniques are applied to the data prior to clustering which can contribute to the reduction of overall complexity of the models used. Some common examples include Principal Component Analysis (PCA) as in Koivisto et al. (2013). The former considers direct clustering of the data in its raw form. This will be considered in Chapter 4, which focuses on direct clustering analysis of smart meter data. The motivation behind the use of a direct clustering approach is the importance of the ability to trace back easily to the raw data, especially if such analysis is performed on energy supplier premises. Some of the most popular methods used for direct clustering are k-means, hierarchical clustering and a Bayesian non-parametric approach, the Dirichlet Process Mixture Model. These three were applied and evaluated in Granell et al. (2015) in the context of electricity energy load.

Whilst direct clustering is considered more popular, some potential problems associated with it were outlined by Wang et al. (2018). One of the most important things to consider when deciding between a directed or indirected clustering approach is the resolution of the smart meter data. For data that has a resolution of minimum 30 minutes, a direct method approach may be appropriate. This may also be because smaller activities over such a period (i.e. boiling a kettle) will be summarised. For cases where more granular time intervals are present, say 1 minute or even 15 seconds, an indirect approach may do a better job. Using various feature transformation techniques prior to clustering such as Fourier transformations or simple PCA may help to reduce rather noisy information and focus on average uniqueness of profiles whilst reducing the complexity of the overall dataset. An alternative way of reducing data may be through splitting the data into chunks of time periods that can then be analysed separately. For instance, a researcher may look only at peak hours consumption.

Overall, while there is certainly a vast choice of methods to perform load profiling, an important distinction needs to be made between stored and streamed smart meter data. Most of the past work was performed on stored smart meter data and

this applied to both academic and industry researchers. This also applies to the research performed in this thesis. The reality is however, that for the best leveraging of the data, streaming data analysis is far more in demand. Using distributed clustering and various incremental clustering methods may address this type of problem. Clustering methods such as deep embedding clustering (Xie et al., 2016) may be tried out on larger chunks of streaming data. This method can then compared to results of other clustering methods using various indicators such as Silhouette Index, just to give an example of a few that were presented in paper by Zhang et al. (2012). According to Wang et al. (2018), there is a still a large gap in the literature on how to select the most useful features from data reduction tasks prior to any clustering being carried out. An attempt to do just that will be presented later on in the thesis, when it will be turned to the application of wavelet analysis on smart meter data.

Last but not least, an important area of methodology for smart meters is outlier data classification. Outlier and missing data detection or bad data detection is certainly one of the preliminary stages of any data analysis that involves observational data. By grouping the patterns using various measures to minimise the similarity within the group and maximise dissimilarity among the energy consumption groups, one may identify such behaviours. An example of methods for both offline and online data cleaning can be found in the work of Peppanen et al. (2016). They perform their analysis under the assumption that the smart meter load data can be characterised by the neighbourhood data points and their combination. This is similar to the Autoregressive Moving Average Process (ARIMA) commonly used for time series. Extensions of this such approach can be seen in Akouemo and Povinelli (2017), Li et al. (2010) and LUO et al. (2018).

2.5.2 Recent applications

Examples of some current application of clustering used for smart meter data can be found in Chicco (2012) who has provided a comprehensive summary and comparison of the clustering approaches that can be applied to smart meter data. These applications targeted not just a group of customers, but thought to identify further types of customers (for instance customers that are suitable for tariff modifications).

Such direct classifications may have a potential to contribute towards meeting energy efficiency goals. One of the most common suggestions is the segmentation of customers by type of activity and commercial characteristics. It could be extended to include some quantities of electricity measured by, for example, annual activity, power factor and utilisation level. Models that include weather characteristics could also be included separately. Such was also considered in Beckel et al. (2014). Their study provided an extensive analysis of smart meter data for 4,232 households in Ireland over 1.5 years. The feasibility to use combined supervised machine learning and multiple regression analysis was shown in attempt to reveal various energy customer characteristics. High prediction accuracy was achieved for more than seventy per cent of the data. Beckel et al. (2014) main suggestion is that such combined method may be transferable to all smart metering systems with similar data magnitude and structure.

In addition to this, the work of McDonald et al. (2014) and Sánchez et al. (2009) used half hourly smart meter data on electricity that have provided a good foundation for the analysis of overall trends among different customers. They use Fourier analysis, self-organising maps, and various clustering methods for overall results comparison. A very clear customer segmentation was shown based using the typologies of high usage customers, low usage customers, business customers and minimal users. This confirmed the feasibility of observing expected trends in peak hours and showed a clear differentiation in consumption patterns among segmented types. Cao et al. (2013) for example, further used peak time classification to define customers from 4,000 Irish households that were most suitable for energy campaigns. According to the author, using k-means clustering was sufficient for clear customer segmentation. Kwac et al. (2014) extended such analysis by using different feature extraction that served as a base for segmentation of customers by lifestyle and consumption behaviour. They performed an extensive analysis by using both k-means and hierarchical clustering which was further extended to multidimensional segmentation based on quantity and variability of consumption.

Silipo and Winters (2013) used electricity smart meter data, also from Irish

households, and business recordings to provide a reliable prediction model of power shortages and surpluses as well as contribute to targeting mechanisms for finding customers who could be subject to different contract offers. Clustered data was used in an autoregressive time series model to predict future consumption considering the past trend. Similar to McDonald et al. (2014) the authors showed the effects of weekly and 24 hour seasonality. Most of the clusters showed significant differences in consumption during weekdays and weekends as well as for mornings and evenings. This approach was further expanded in Oates (1999) and Liao (2005) who showed the application of various time-series clustering techniques to smart meter data. The work of Liao (2005) in fact will be crucial to set up the methodology design in this thesis given the distinctive complexity of the time series structure presented in the sample in this thesis compared to those that were seen in the past work.

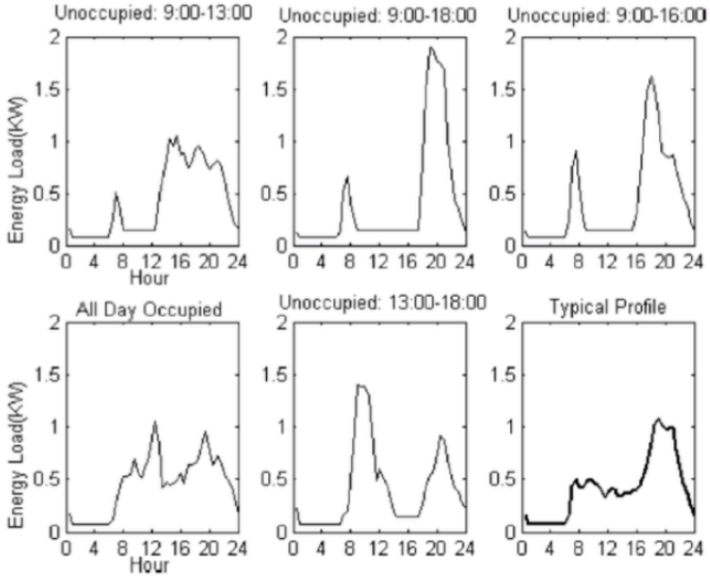


Figure 2.4: *Energy load profiles of a UK average households* Source: Yao and Steemers (2005)

Yao and Steemers (2005)have further developed energy consumption profiling by adding the dimension of UK housing typologies and household size. These are

flats, semi-detached, detached and mid-terraced properties. Controlling as well for the type of the ownership and average of energy consumption loads they profiled randomly generated data on daily energy consumption, using a hundred artificially created households and providing the results on a regional level.

The validity of the results was confirmed through a comparison with national statistical data. Six distinct profiles (Figure 2.4) were defined based on the occupancy of the household space, from which researchers could further infer the employment characteristics of the members of household, such as full-time or part-time work for example.

In addition to the above, a further contribution to clustering methods was provided in Oates (1999) and Liao (2005) , who considered surveying different time-series clustering techniques applied to smart meter data. This was particularly useful for identifying various groups of customers that could be targeted for efficiency measures and also in our case, for the provision of support to vulnerable energy customers. For energy suppliers, it is also important to understand the difference in the services they may present to different customers, which could improve their heating environment. For example, it is valuable to differentiate between the tariffs that may be suggested for elderly and disable customers vs families or young sharers.

2.5.3 Forecasting individual use and energy demand

Load prediction and forecasting are some of the most topical methodological applications being used on smart meter data, mainly in industry, and is being pushed to the very cutting edge due to suppliers interests in becoming more efficient and more competitive, whilst also meeting their regulatory requirements of accurate forecasting of pressure on national grids. The importance of forecasting energy demand is not underrated as it helps both suppliers and countries regulators manage their imports and exports of energy resources effectively. The majority of the work available to date considers high voltage forecasting (i.e. city, country or grid network levels). The preference is given to high voltage level analysis due to lower volatility and sensitivity of individual customer choice to overall load. This doesnt mean however

that individual customers decisions arent important for accurate prediction of high voltage energy load. It is the difficulty of modelling the uncertainties and volatility that is an obstacle. Each customer needs to be studied individually within any methodological framework such that their personal shifts and energy consumption behaviour are taken into account. As was seen from the first part of this chapter, each energy customer may have a unique response to factors that include but are not limited to weather change and electricity and gas price fluctuations.

Smart meter data, especially the data that is available at a resolution of half hour or less, can push the accuracy of load forecasting much further, not just at the aggregate and high voltage level but at a very granular, individual user level. This can be useful for the alteration of energy consumption at the individual level (i.e. reducing peak time load) and for alteration at national grid level (i.e. design of new batteries to meet the peak times pressure). As was noted by Wang et al. (2018) depending on which level the forecast in needed, different set of approaches may be used.

Yet, it is important to note that prediction models based on machine learning have received only limited attention in the literature on customer classification. Notable exceptions are those that focused on using energy consumption data to classify the types of appliances that are used by UK households Lines et al. (2011). Other studies, for example those by Yu et al. (2010), Lee et al. (2012) and Haghi and Toole (2013), looked instead at energy consumption point prediction using machine learning methods. As an example, Haghi and Toole (2013) looked at 6,000 smart meters in Ireland and showed the feasibility of using the Levenberg-Marquardt Neural Network algorithm with twenty hidden layers for time-series consumption prediction. Lee et al. (2012) used thirty hidden nodes with five input nodes for one output, which is the electricity consumption in the next period. With a prediction error of less then 0.2 they were able to achieve high accuracy in prediction using a neural network model. Lastly, Yu et al. (2010) compared neural networks and decision trees in application to determine energy consumption in Japanese residential data. They suggest that the decision trees may be a better predictive model from

the point of view of interpretation when compared to neural network. Yu et al. (2010) achieved around 94 and 92 per cent accuracy for the training and test subsets respectively. This is something that will be aligned with the results presented in the thesis. After surveying various predictive methods in Chapter 4, and also for classification of patterns in Chapter 6, the tree methods significantly outperform a number of other methods that were used in the experiments.

Some of the work that has been done on energy use point prediction offers a useful variety of methods. One of them could be a Bayesian approach with prior information. This was used by Hsiao et al. (1995) for a sample of 347 households, in a study that demonstrated the possibility of using prior information formed from the means and variance of past consumption to predict future energy consumption.

Some of the most recent applications that are available for energy consumption prediction using smart meter data are available in Taieb et al. (2016),

2.6 Summary and Conclusions

This chapter has provided a detailed review on the research that is available to date and that studied energy consumption from both causal and methodological perspectives. The first section has reviewed the work that studied various factors that may explain variability in energy consumption at a residential customer level. This section highlighted how challenging it is to infer in practice why one customer may differ from another in terms of their energy use. Some of the important implications for the public and social policy agenda in the UK, such as understanding how fuel poor customers may possibly be identified from energy consumption, was also reviewed. The second part of the chapter was dedicated to methodological challenges and recent work that considered development of methods for identifying outlier behaviours in smart meter data, load profiling and load forecasting as well as various data reduction methods. As was shown, the past ten years have been marked with the increasing availability of smart meter data to researchers in engineering, computer science, statistics and operational research fields. This has led to more than 200 academic papers in methodology and data analytics for smart meters

being produced from 2010 to 2017 (Wang et al., 2018).

While this certainly shows a richness of work that has been done with smart meter data, one of the immediate limitations and gaps remaining is the lack of studies that consider big data: most of the research available to date tends to be performed on small samples. This creates limits on anticipation of variability in consumption across populations as well as issues with missing data and outliers that can be more variable once a larger dataset is taken into account.

2.6.1 Variability of energy use

The preceding discussion on factors that affect variability in residential consumption has demonstrated that energy consumption may be affected by quite a few determinants. Some of these can be easily quantified, such as weather, household size, life stage or income. However, some of them, while they have the potential to explain high heterogeneity in energy dynamics, are quite hard to present in numerical form no matter how advanced the methodology used for energy data analysis is. These are the cultural and social environments, lifestyles and habits of the customers under consideration.

To complement the description provided in the chapter, one could also consider a summary of the main variables involved in the energy consumption dynamics presented in Steemers and Yun (2009) (Figure 2.5) that usefully describes the interactions and also potential channels for hidden causal mechanisms. . It also shows us that it is the interactions between the various factors that result in unique temporal profiles; household characteristics in interaction with property types, appliances used and weather tend to affect energy consumption dynamics.

From a methodological point of view, this chapter has shown that despite only the recent and yet limited availability of smart meter data, the number of methods in place to generate insightful findings appears to be quite overwhelming. Mostly developed on rather smaller data samples or data of similar structure these methods can offer both industry and academic analysts tools that can help classify the energy load, forecast future energy use as well as provide a solid framework for finding bad, outlier data as well as identifying energy theft. Some further developments are still

required to be considered for big smart meter data and streaming data.

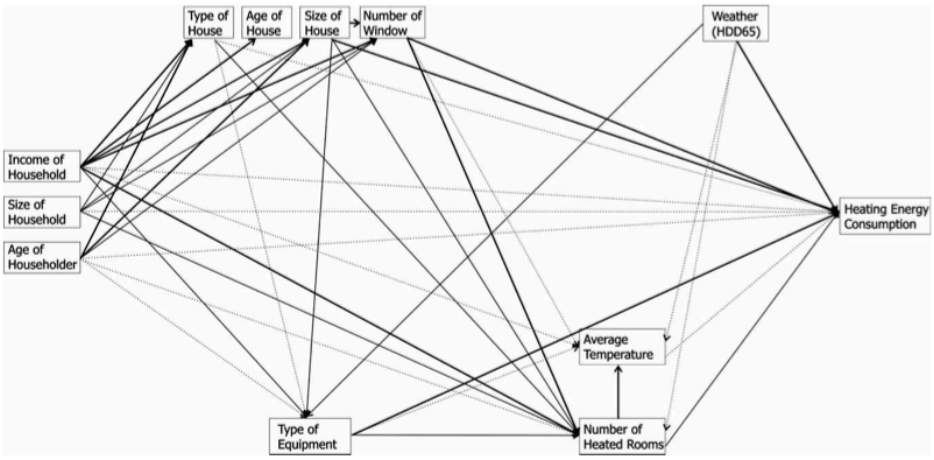


Figure 2.5: Path diagram and variables interactions in affecting energy consumption (Steemers and Yun, 2009)

As was seen in this chapter, it became more and more apparent that the consumption environment and the infrastructure for the energy services provision play a very important role. This may also be particularly useful when one is attempting to define the factors that may contribute to fuel poverty on an individual or household level. Although, while this may sound straight forward at first, a number of methodological developments need to be considered for incorporation of multiple factors simultaneously. For instance, one may consider graphical models. Due to limitations with the data available for the work in this thesis, this can only be considered as a direction for further work, hopefully to be picked up by other researchers.

In this thesis, the main contribution to the above gaps and challenges is given through understanding how much the *time* can explain ones energy use. Smart meter data, when available at high granularity (half hour) for a long enough time period and with greater spatial diversity, may tell a story about energy consumers, their unique habits, activities and even lifestyle. This thesis aims to investigate how

much smart meter data can offer in cases where there is no additional information about customers, something which can be quite common on both energy company premises and within governmental bodies.

Chapter 3

Data

3.1 Introduction

Energy consumption data recorded by smart meters is an example of highly variable spatiotemporal data that inherits not just the micro dynamics of an individual consumer but also the macro affects that may correspond to patterns affecting the countrys population at large, such as climate and weather as well as socio- economic macro variables like the average income or proximity to large cities, to name a few. Smart meter data can thus be treated as an important ingredient of any complex system designed to manage micro and macro environments of not just energy consumption but also those systems that support population dynamics as a whole. Whether it is the design of more efficient batteries driven by an understanding of the pressure on the national grid at peak hours, or an investigation of the inequalities among individuals mirrored in their energy consumption load, smart meter data offers tremendous opportunities to study individual variability of consumption in depth on both a national and individual scale. There is a widespread optimism in regards to the potential insights smart meter generated energy data may provide among the research community and industry practitioners alike (Anderson et al., 2017; Newing et al., 2016; McKenna et al., 2012; Hargreaves et al., 2010).However, one of the reasons why these advantages are still in the process of being operationalised is the complexity of the smart meter generated data, or big data, given its nature and more obviously, its size. This chapter will look at the data and various

ways in which to describe it, given its volume and its associated issues. Overall, as it will be observed, the complexity that a researcher may be willing to accept or sacrifice will largely depend on the research questions one has in hand and on the desired level of generalisability an investigator would like to consider.

3.1.1 Structure of the chapter

Prior to proceeding to generation of any insights from the data, an important note must be made on the distinction between data and information. It is useful to use the Witten et al. (2016) definition which refers to data as being in large part the information in its raw form, represented primarily by a set of facts. Information in turn, can be characterised as a system of patterns, relationships and expectations derived from the data. Accepting these definitions, the process of turning the raw data of energy consumption into useful insights about energy consumers may not necessarily produce one single answer. This chapter introduces simple data descriptions and addresses their associated limitations. Both data samples analysed in the thesis are presented along with descriptive statistic measures. The unit of analysis and sample selection for the study will be discussed in more detail, mainly to explain the motivations for choosing various sub samples that can be used for further exploration and the experiments in the remainder of the thesis.

The thesis is based at large on the analysis of two main datasets that are analysed more closely in this chapter and one supplementary one which will be described in more detail in Chapter 6. The first dataset consists of aggregated readings at the national level using postcode sector geography, while the second presents the recordings at the individual level but uses only the City of Bristol region and Census Output Area as a geographical reference. Time resolutions remain the same for both datasets. There are 48 daily half hour readings that span across the years 2014 and 2015. In this chapter, a description of the national dataset is followed by contextual motivations for sampling of big data and a discussion on how a smaller sample of data from Bristol, for instance, may be used as a representative case of the wider population. Various approaches to visualising smart meter data will also be presented. The main motivation of this thesis section is in fact to explore and exhaust

as much as possible the opportunities that are available to describe the smart meter data before moving on onto more advanced techniques that are designed primarily to generate very narrow and specific insights from this data.

The chapter begins with a basic analysis using simple measures of variability in the data such as average, standard deviation and range. The main reason for such specific choices is that they can be mapped and compared on a national scale. In fact, for the sake of interpretability, when using mapping as a tool for data description, simplicity may be preferred as there is a large amount of data and it would be best to avoid overwhelming the reader with over complicated graphics. This chapter will focus partly on how temporal data description using descriptive tables may be complemented with the visualisation of the spatial distribution of the data points.

The remainder of this chapter is organised as follows. Some simple visualisation of smart meter data patterns will be presented in the next section where examples of various approaches that can be taken to analyse smart meter data will be provided. The national sample will then be discussed using spatial and temporal dimensions in Section 3. Similar analysis will be performed for the Bristol data case which is available at slightly greater geographical resolution although the temporal granularity remains the same. In Section 4, each of the samples will be further accessed for heterogeneity within and across the units of analysis. It may appear that the analysis in this chapter is rather basic, yet it is important to remind the reader that the purpose of this section is to determine how much insight the conventional methods of data exploration may offer at this stage. A bottom up approach will be used to assess the usability of descriptive measures of data such as mean, standard deviation or geographical location. The aggregation of the data points here is inevitable, either at geographical or temporal scale or both. This will also be considered in more detail with additional attention given to the limitations and advantages associated with different scales of aggregation. Opportunities for linkage of smart meter data with administrative data sources will follow in Section 3.7. This will be complemented with a discussion of ethical and legal constraints that significantly affect the resolution of datasets available for this research. As a

reminder to the reader, that simplicity and interpretability are the ultimate goals of any data analysis: the last section of this chapter will be dedicated to the evaluation of the presented descriptive analysis , its usefulness, associated limitations and motivation for using slightly more advanced techniques that will be employed later in the thesis.

3.2 Smart Meter Data

As was seen from the literature review, smart meter data has received attention from various fields such as engineering, mathematics and statistics, geography and political science. What makes this data a distinct and rich source for social science is the fact that is available at half hour temporal resolution, always presenting the complete consumer records pertaining to fixed locations. Such data can be valuable for various applications that include analysis of social behaviours, activities, economic well-being, effectiveness of policy intervention, the evaluation of effectiveness of energy efficiency measures and many more. The granularity of data and the completeness are quite unique features as generally consumer data is not available at such a continuous scale as consumers and individuals tend to move from one supplier to another. With energy, unless the smart meter user has switched energy provider, the records of consumption will represent their journey through time without any interruptions or temporal gaps.

Overall, smart meter data presents a fine temporal granularity. In this thesis, half hourly readings are considered, yet smart meter data can also be available at 1 minute intervals or even a second. In terms of spatial granularity, data is in the acceptable format to be linked to other administrative sources, be it the postcode level or Census Output Area, depending on the granularity at the source there are possibilities for linkage to study how representative smart meter data is in terms of general population data.

Where readings are available , the researcher can be certain that those directly represent the smart meter user and not anyone else. This is contrary to loyalty card data for instance as those may be borrowed/lent. This thus makes smart meter data

a major source for spatio-temporal data mining that may reveal valuable insights into population dynamics and activities.

3.2.1 Unit of analysis

Change is always taking place - unnoticeable, rapid, large or small (Gibson, 1979). However, most processes tend to exhibit controlled, rather than unstable, variation across space and time. Understanding what is stable in terms of energy consumption habits, for example, may enable us to see what is changing with respect to those static elements. Smart meter data may offer the potential to track changes and infer potential factors contributing to them. Unlike the positivist theories of spatiality, such as those of Johnston et al. (1996) , where time is held constant and rational individual behaviour assumed, smart meter data allows us to look more closely at anomalies and incorporate more dimensions that may arise from various consumption environments, which emerge due to unique characteristics of individuals living in different places, different climates, housing conditions or having distinct lifestyles. All of these will contribute to daily, weekly, monthly and seasonal variations in consumption. And it is with these other factors in mind, that the experiment with samples of different temporal and spatial granularity are presented in the thesis.

As a consequence, in the case of smart meter data, the question of unit of analysis may be posed in both temporal and spatial contexts. For instance, the time interval chosen as a temporal unit may play as important a role in the final analysis. One may consider to analyse separately daily and nightly consumption or focus only on peaks in consumption. Readings of energy consumption aggregated spatially, for instance at postcode level, will also imply a modification to the unit of analysis. This is quite handy if the data needs to be reduced in size and more targeted insights can be obtained. One way of addressing this, is to create chunks of time intervals that reduce say, 48 dimensions to five by splitting the day into five components, one of them being late night (after midnight) and the other four representing morning peak hours, midday, evening peak hours and midnight.

Generally speaking, the choice of aggregation is largely driven by the use of the output we are interested in. When looking at national patterns, units of analysis

are usually aggregated at the postcode sector level and thus only associated with the geographical reference of that sector. Similarly, when analysing differences in consumption among Census Output Areas, Output Area is taken as the unit of analysis. However, in order to understand individual dynamics and the uniqueness of energy consumers behaviours on average, the individual smart meter data is used as the unit under study. With the sample analysed in this thesis there is no available data on the size of household or building type associated with smart meter users, there is an associated uncertainty regarding whether the smart meter and its data correspond to a single household or a number of households that reside within the same building (i.e. student halls of residence). Furthermore, with obstacles arising from the sample size and spatial resolution, there are limitations with how this smart meter data compares with recent census data. Thus, an additional degree of uncertainty about the number of individuals living in a household exists when it comes to smart meter data. This is what partly motivates the final definition of the unit being smart meter user with all the uncertainties associated around this measure. Some assumptions thus may need to be made when referring back to the consumer about the insights generated from the data. To take an alternative view, the unit of analysis may also be represented by a set of attributes of consumption. For instance, high consumption or low consumption users defined by a specific numerical threshold can be grouped together and studied as a unified object. Temporal dimensions can also be used as a unit of analysis, for instance the data can be grouped by seasons. A spatial approach can also group the patterns for instance using threshold of urban/rural regions. This thesis will be looking at various ways to both group and desegregate smart meter readings. It will be observed that depending on the choice of unit of analysis, very different insights can be generated from the data.

3.2.2 Ecological fallacies

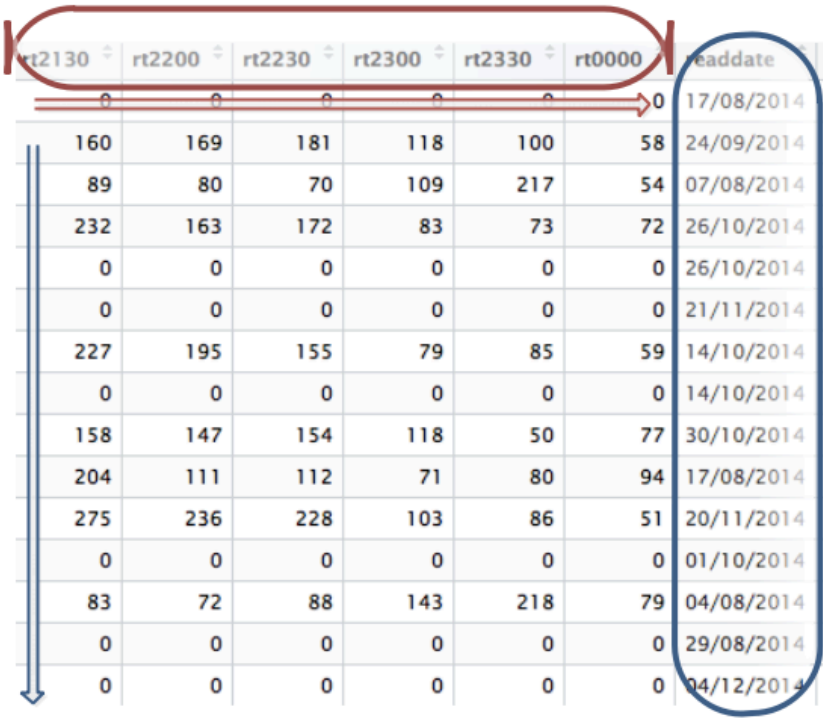
It is vital, at least briefly, to discuss the potential threat to validity of any inferences that are made using spatio-temporal datasets such as smart meter readings. One of the most common cases of ecological fallacy is to assume that a population average holds information about the likelihood of an individual in that respect. For instance:

if it is found that the average income in the UK is 20 thousand a year, it is wrong to assume that each individual in the country has an equal chance of earning this salary. This can similarly be observed when studying correlation on aggregated and individual levels. More formally, it can be shown that correlation on an aggregated level is different from the correlation observed on an individual level (Piantadosi et al., 1988). The above is formally relevant when smart meter data is being studied at an aggregated level, for example in postcode sectors or various regions, or at an individual smart meter user level. The inferences should be made accordingly. Studying the population at large tells us little about unique variation of energy consumption and, vice versa, a study based on individual smart meter users may not necessarily inform us of general population trends.

3.3 Simple Smart Meter Data Visualisations

There are a number of ways in which smart meter data may be visualised and treated with respect to its spatial and temporal dimensions. For instance, taking the example of one single user, we may see that consumption may be either analysed using one specific unit of time, say 6am consumption load across the whole year or, alternatively, by looking at subsequent readings instead. The motivation for which time point to choose will be discussed in more detail when the methodology of this thesis is considered. Examples of different lenses with which we can look at the data of energy consumption records are presented in Figure 3.1. The figure illustrates the two to represent the time series sequence for energy data: (a) **the intra-day consumption**(i.e. how does energy usage depend on hour of day) ; (b)**the inter-day consumption** (how does usage vary across days) .

If we take the intra-day consumption (red colour), it would tell us how energy consumption varies across the day as times moves forward. This example is presented in the figures below. This is a conventional method to present energy use using smart meter readings as it is highly intuitive and gives an overall idea of how unique the consumption profile is. The example of profile that may be available at half hourly resolution is presented in the Figure 3.2.



rt2130	rt2200	rt2230	rt2300	rt2330	rt0000	readdate
0	0	0	0	0	0	17/08/2014
160	169	181	118	100	58	24/09/2014
89	80	70	109	217	54	07/08/2014
232	163	172	83	73	72	26/10/2014
0	0	0	0	0	0	26/10/2014
0	0	0	0	0	0	21/11/2014
227	195	155	79	85	59	14/10/2014
0	0	0	0	0	0	14/10/2014
158	147	154	118	50	77	30/10/2014
204	111	112	71	80	94	17/08/2014
275	236	228	103	86	51	20/11/2014
0	0	0	0	0	0	01/10/2014
83	72	88	143	218	79	04/08/2014
0	0	0	0	0	0	29/08/2014
0	0	0	0	0	0	04/12/2014

Figure 3.1: *Two ways to represent the time series sequence for energy data: (a) red colour; (b) blue colour*

This is an example of a rather typical profiles, double peaked at morning and evening time intervals. Such a profile is highly likely for a smart user that can be described using full time employment characteristics (i.e. someone who leaves home in the morning and comes back around 5-6pm). To illustrate an alternative example, with activities being present throughout the day, please see the Figure 3.3

The explanation for the difference using full time employment is suggested only as an assumption. Most of the potential reasons that profiles may differ remain inconclusive till validation can be performed using other data attached to the smart meter. While this is something which may be achieved at energy supplier premises, the ethical considerations of academic research ensure that such details cannot be obtained for individual users. This certainly raises discussion about privacy and confidentiality of smart meter users. More details on this will be given in the final section of this chapter.

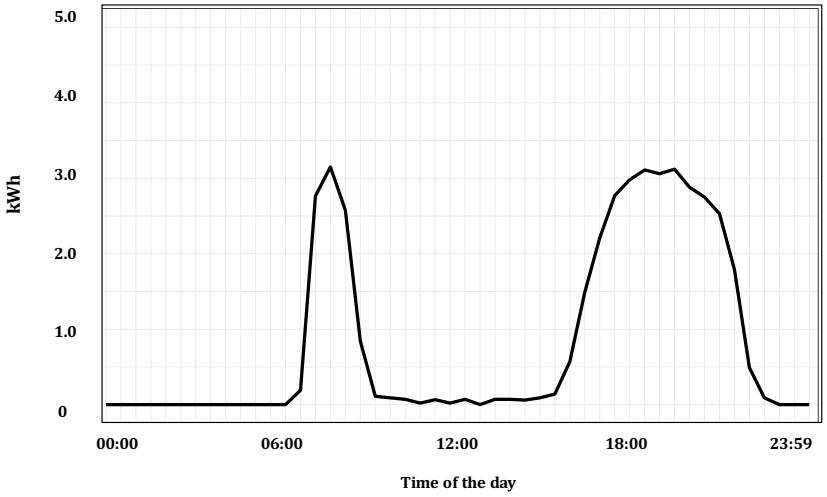


Figure 3.2: Example 1: 48 half hourly profile of energy use

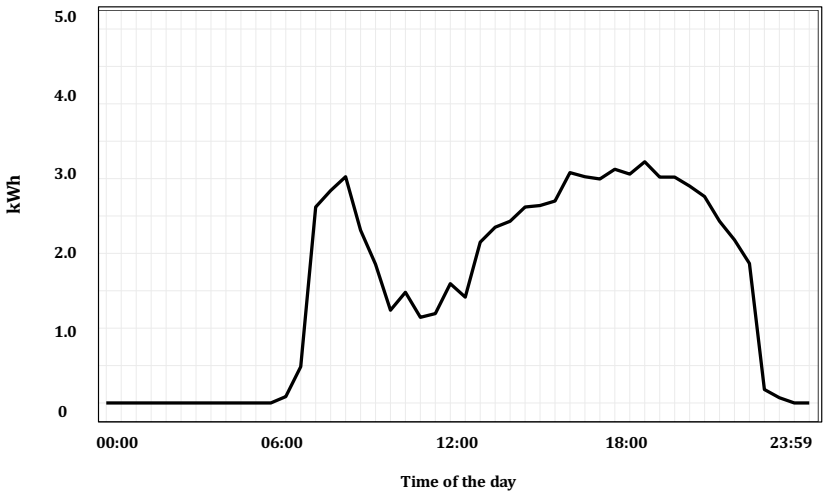


Figure 3.3: Example 2: 48 half hourly profile of energy use

3.3.1 Importance of 'good' visualisation

As was greatly acknowledged by Goodwin (2015); Jarrah Nezhad et al. (2014), data visualisation tools hold significant potential for energy consumption analysis for both energy providers and consumer households. Examples of such use could include providing the consumer with a simple visualisation of daily consumption on

the smart meter screen, known as consumption feedback, as well as offering energy provider analysts a large aggregated dataset of regional energy consumption. The power of visualisations of smart meter data should not be underestimated as they can offer an efficient and quick way of presenting data as well as communicating the information extracted to research communities in various fields without getting into much technical detail. In terms of spatial data and public policy research, the way in which data is visualised is vital as often the consumers under investigation and the ways in which they are geographically clustered is key. Maps and geography based visualisations of the smart meter data then provide policymakers with a clearer idea of the areas that need to be targeted and their boundaries As was suggested by Goodwin (2015), while visual technological improvements are evident for individual users, for example in the case of the smart meter and various web applications that help consumers evaluate data on their energy usage, there is a lack of visualisation tools that can be used by energy suppliers. This thesis seeks to fill this gap and provides an overview of various techniques that may be useful in organising and grouping the data from an energy provider perspective. As will be presented in this chapter, more conventional methods such as deriving mean or median consumption of energy consumers contain little representative power on a larger scale and tell us only a fraction of the information collected on various energy consumer groups. Smart meter data are highly variable and the coverage includes both data of a high temporal resolution (half hour for each day of the year) and spatial resolution (postcode sector). There are several ways in which the descriptive statistics of these data can be presented. Depending on the question one aims to answer, and the granularity of the outcomes under consideration, different angles can be used to visualise the central tendency and variability measures in the dataset. Thanks to developments in software such as QGIS and R, it is easy to provide quick and reproducible research tools to visualise and analyse smart meter data. Such visualisations will begin with a very simplified approach that allows one to look at annual half hourly average consumption per postcode sector. Such an approach captures the overall variability in the dataset and the range of the values one is dealing with,

as well as considering any spatially dependent regions that share similar values. Such a method, in general, is highly useful for initially, defining outliers as well as motivating the detailed analysis of a particular case study sample. An example of descriptive statistics for national scale aggregates are presented below. The mean, standard deviation and range of data variation in each half hour are computed for each postcode sector using data from the smart meters from households with a complete annual record. Once the bounds of the data sample are defined, the variation at each half hour at each postcode sector was aggregated and then the variation among half hour aggregates is analysed. Both gas and electricity variability will be measured in the remainder of the chapter. Overall, the goal is to devise methodology that can be easily applied to data arriving from both sources. Please note that mainly different hues of red will be used to describe the magnitude of gas consumption and hues of green to map the spread and intensity of electricity consumption. Neutral maps, that describe the counts of smart meters for instance will be presented using yellow-blue (national sample) and blue (Bristol sample) hue palettes.

3.4 National Sample

The first sample which will be considered for the analysis in this thesis is the national sample of smart meters available at the Postcode Sector level, year 2015. This section will describe the dataset from both temporal and spatial perspectives. It is important to remind the reader that this research is the first attempt to study smart meter readings at such magnitude. Limitations and challenges associated with temporal and spatial resolution in the dataset are still to be observed and found. Some of them will be unlocked in the following sections [1].

3.4.1 Overview of the dataset

The national dataset of smart meter data that is used for the analysis in this chapter is held by the Consumer Data Research Centre (CDRC) and was sourced from one of the UK Big Six energy suppliers. The data contain details of around 1,080,000

[1] Some of the descriptive analysis presented in this Chapter have appeared in the book chapter co authored with Roberto Murcio in Longley et al. (2018).

added up electricity and gas domestic smart meters for the year 2015, which represents 43% of the 2.3 million smart meters installed by the end of December 2015 in the UK. The spatial granularity is at postcode sector. The broader figures are shown in Table 3.1 It is important to note that throughout this section, individual figures on numbers of smart meters and measures are rounded to the nearest hundred. The number of energy users per month is constantly growing as the rollout of smart meters is increasing from one month to another. For example, in the case of electricity, 75% of the users were already present in the first quarter of 2015 meaning that these will be the customers records with the full year coverage (Figure 3.4). Between April and September, less than 5% of the total were enrolled. Finally, in December around 50,000 users were added bringing the total to 600,000 users with a smart meter by the end of 2015. We may conclude that the roll out of the electricity meters is gathering momentum. This was also confirmed by BEIS (2017). A breakdown for the roll out by quarter of the year is shown below.

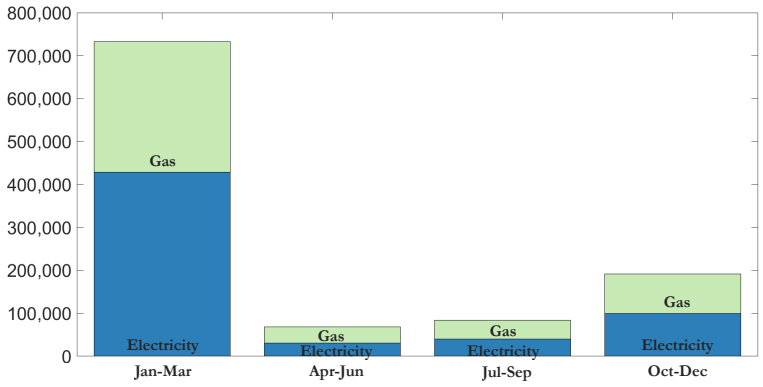


Figure 3.4: *Number of smart meters that we added at during Q2 to Q4 to baseline in Q1, 2015*

Type	Number of meters	Number of Postcode sectors with at least 10 meters installed	Meters per Postcode sector	
			Mean	Median
Electricity	600,000	8,000	70	60
Gas	480,000	7,500	60	50

Table 3.1: *Gas and electricity counts* the number of postcode sectors with at least 10 smart gas or electricity meters in Great Britain as of December 2015.

Descriptive Statistic	Electricity	Gas
Average	2,130 kWh	8,480 kWh
Median	1,820 kWh	7,105 kWh
Standard Deviation	1,680 kWh	6,510 kWh
Average (BEIS 2015)	3, 894 kWh	11,707kWh
Median (BEIS 2015)	3,148 kWh	13,202kWh

Table 3.2: *Central tendency description* The average annual household energy consumption estimated using the national sample compared to BEIS 2015 national estimates.

Region	Electricity meters (thousands)	% of all meters in the region in 2015	Gas meters (thousands)	% of all meters in the region in 2015
East Midlands	48.6	2.00%	40.3	2.00%
East Midlands	48.6	2.00%	40.3	2.00%
East of England	47	2.00%	38	2.00%
London	65.9	2.00%	54.6	2.00%
North East	24.7	2.00%	22.6	2.00%
North West	96.1	3.00%	76	3.00%
South East	57.1	1.30%	47.7	1.70%
South West	41.1	1.50%	31.6	2.20%
West Midlands	79.2	3.00%	64.8	4.00%
Yorkshire-Humber	58	2.00%	45.7	3.00%
Wales	28.1	1.80%	18.3	2.50%
Scotland	53.3	1.70%	40.4	2.60%
Total	600		480	
% of total smart meters installed in GB by all suppliers in Q4, 2015	69.00%		75.00%	
Percentage of all domestic meters in 2017[2]	2.0%		2.0%	

Table 3.3: *Breakdown of smart gas and electricity meters by region.*

The average annual consumption for gas and eletricity observed among the smart meter users in the national sample are reported in the Table 3.2. These central

tendency measures are compared to BEIS 2015 estimates. As may be note the average user in the national sample consumes slightly less compared to the BEIS mean and median.

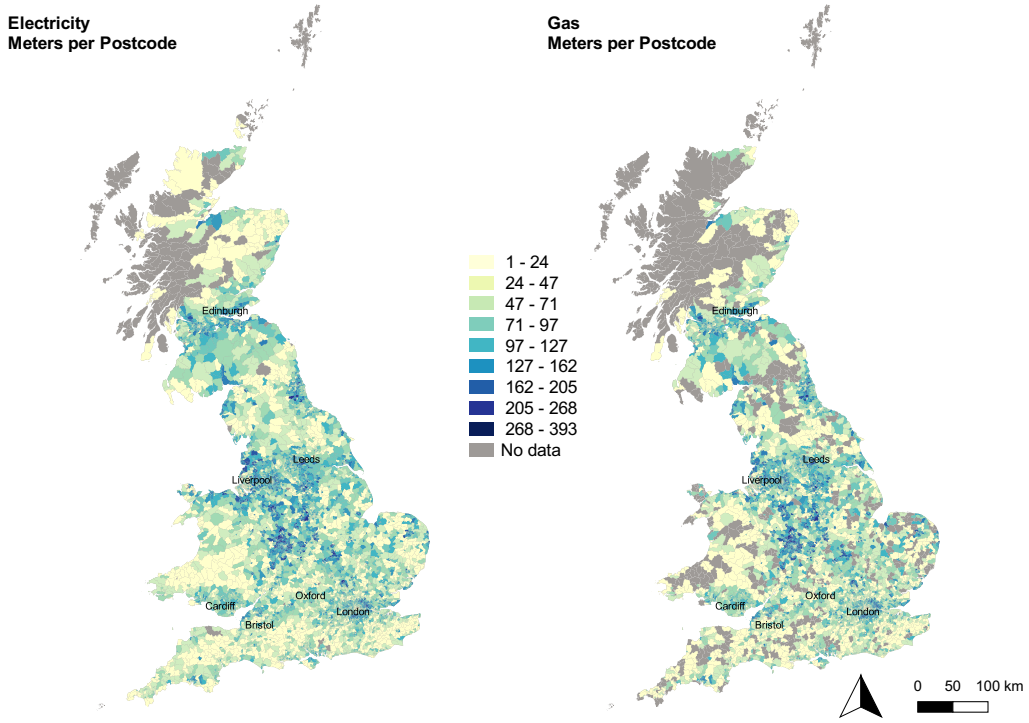


Figure 3.5: *Smart electricity and gas meters by postcode sector at the end of December, 2015.* These maps show the distribution of smart meters across Great Britain with the West Midlands and North West regions have the highest frequencies of meters per postcode sector.

BEIS (2017) reports that despite an acceleration of smart meter roll out, most domestic properties nevertheless still have traditional meters. It is unlikely to be the case that roll out by any energy company thus far has been to a random selection of addresses. For instance, some domestic properties can be unsuitable for meter installation while the needs of disabled customers may pose challenges. The perceived wisdom is that there is a bias in successful installations towards elderly people or families. This is driven by the fact that when local installation campaigns are mounted, representatives are more likely to find households from these groups at home during normal working hours. It is also important to note that nationally, around 70% of households will have electricity and gas supplied by the same company, with 17% having duel supplier, meaning they will have a separate supplier for

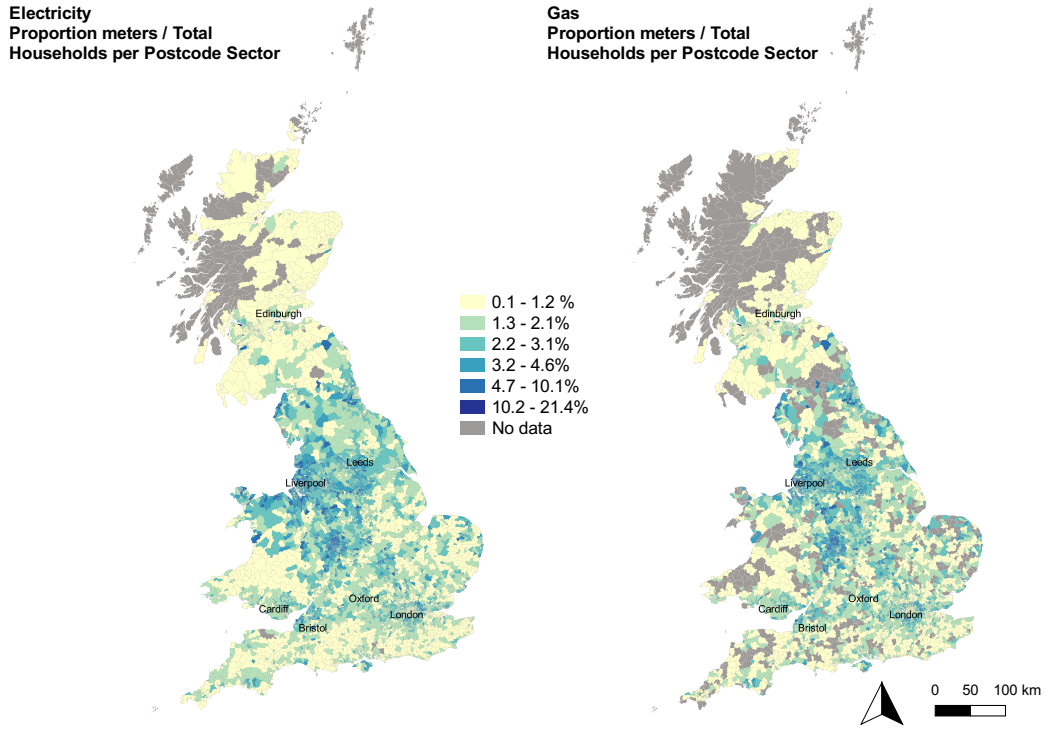


Figure 3.6: *Proportion of electricity and gas meters relative to the total number of households by postcode sector.* These maps show the distribution of smart meters proportionally to the total number of households that reside in the regions across Great Britain

gas and electricity. The remaining 12 % of households will be connected to only the electricity network (OFGEM, 2015). The geographical distribution of meters (Figure 3.5 is slightly skewed towards the North West and West Midlands regions, for both electricity and gas, where almost 30% of the smart meters are installed. In contrast, Wales and North East regions are underrepresented, accounting for only 8% of the total of available smart meters (Table 3.3). Once the number of meters in each region are related to the total number of domestic meters, we observe that the national sample represents only about 2 percent of all domestic meters in the UK. This is not unsurprising results, given that the years of 2015-2016 are still associated with the emergent yet continuing roll out of the meters across Great Britain.

According to the Figure 3.5 and Table 3.3, in more than 80% of the Postcode Sectors, smart meters were installed at between 1% and the 4.8% of the total number of households. The higher percentages can be found at West Midlands, North West and the North of Wales. North West, in fact, is second largest region by a number of

all type meters present, while northern Wales is rather unusually overrepresented in our sample. As in Figure 3.5, the grey areas represent the sectors with no available data.

Please note that in these visualisations all the available smart meters are presented, regardless of the fact that some may not have an annual coverage of records. For more in-depth analysis in this thesis only the meters with full coverage will be used. This will slightly reduce the sample yet will ensure that the temporal granularity is uniform across the units of analysis.

3.4.2 Descriptive analysis

The preliminary steps of data description in the previous section have shown how powerful mapping can be when one is interested in presenting a descriptive summary of a large dataset such as smart meter data. Spatial dimension of energy consumption and the ability of the researcher to map the descriptive statistics instead of presenting a table alone, provide an intuitive way to access the variability in the data to then guide the further choice of a case study that may be based on a specific region. The illustration of the variation in the consumption using half hourly mean, median and total range values associated with each source of energy are presented below (Figures 3.7 and 3.8). The descriptive statistics are calculated by taking the average across half hourly readings available at annual coverage. By doing so, it is assumed that the effects of extreme weather in both summer and winter, difference between day and night, as well as effects of weekends and holidays can be cancelled out once added together. Thus, the average estimate and the variation around it measured by the standard deviation should represent an approximate true central tendency measure for half hourly consumption given the size of the sample. It is observed that electricity consumption exhibits much greater variability which is recorded by standard deviation and mean values. However, in terms of the range of values (the interval between the minimum and maximum consumption), gas consumption is associated with greater overall magnitudes measured by kWh. This is somewhat expected.

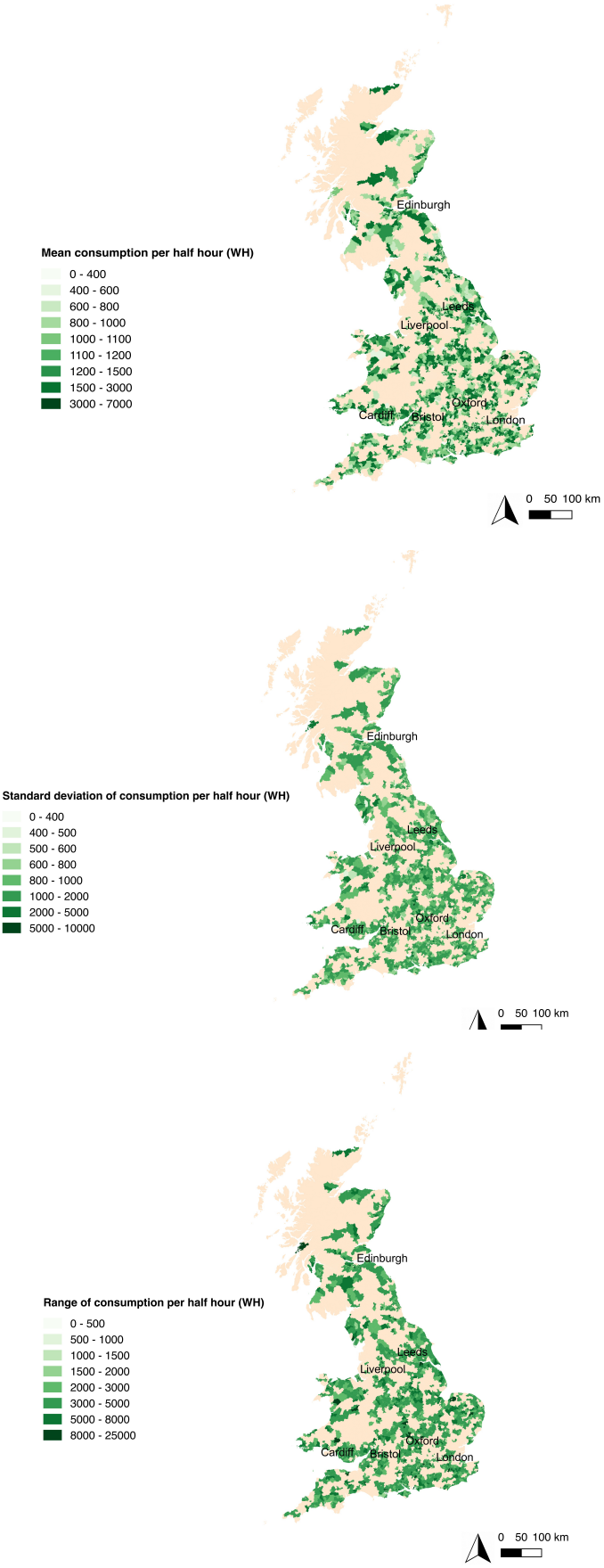


Figure 3.7: *Descriptive statistics for electricity half-hour measures*

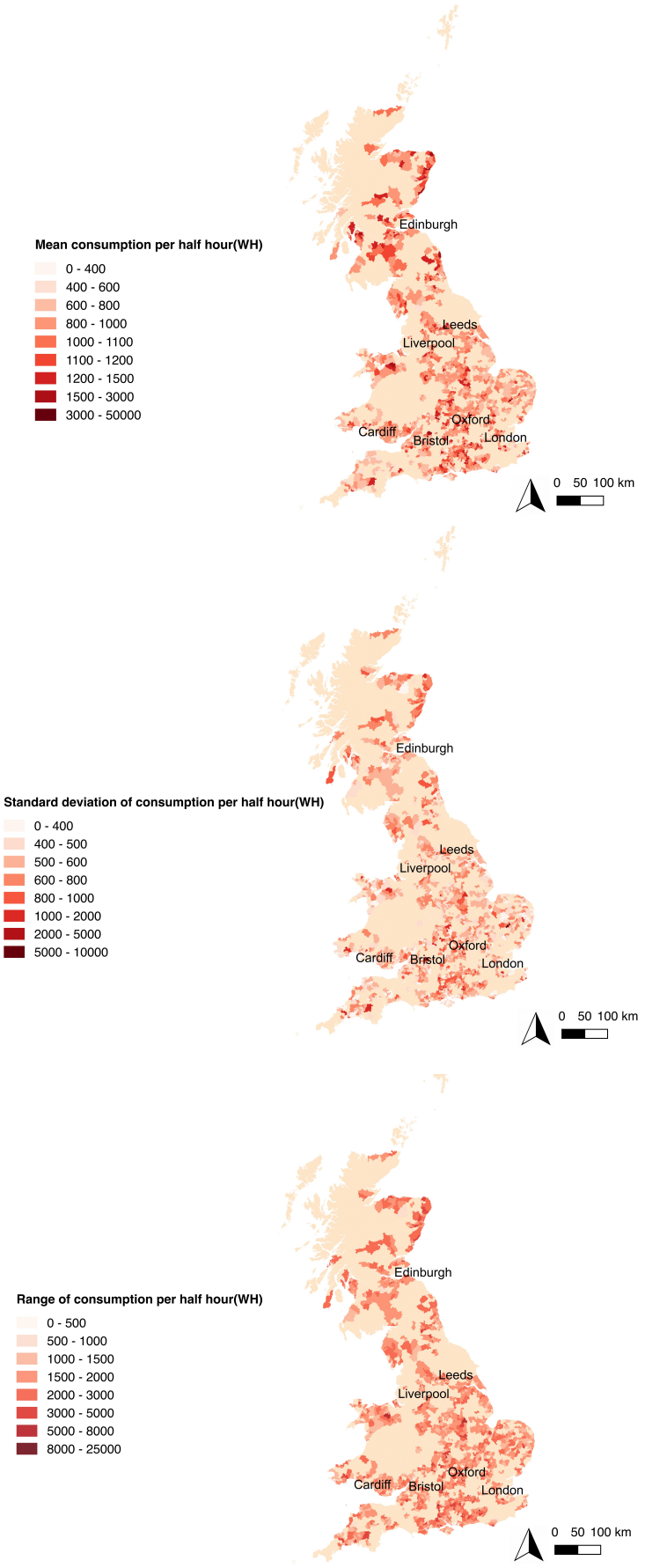


Figure 3.8: *Descriptive statistics for gas half-hour measures*

As was seen from the government figures and the comparison of the average consumption for gas and electricity, on average there are higher levels of gas consumed across the year due to excessive heating during winter. The other important feature to note is that the range of consumption and central tendency measures do not tend to cluster geographically. The figures **??** report a general description of the dataset. There is no specific focus on individual regions at this stage. The main aim is to gather how representative is the dataset of the UK population as whole. As was observed, there is scope to explore further the temporal variation in the data. The differences in the variation can be explored using the spatial locations, yet given the possible bias and low representativeness of the data regarding the general population, caution needs to be taken about any inference produced. This makes most of these findings inconclusive and in need of further validation using other datasets.

3.5 Bristol Sample

This section will present and review the sample from Bristol that is used in the thesis to complement the national sample. The Bristol sample is available at greater geographical resolution (Census OA area) yet has much smaller quantity of smart meters that can be analysed. Bristol enjoys characteristics of a cosmopolitan city that embraces quite a diverse population. This feature allows for the generalisation of results to larger populations of the UK. Compared, for instance, to a place like London, which is shaped by constant change to inflow/outflow of people, Bristol tends to have more stable overall characteristics over time.

3.5.1 Overview of the dataset

This sample will be used in this thesis mostly for temporal analysis. Given the fact that data corresponds to a single region it becomes possible to generalise any results obtained using this sample at least to the level of Bristol. These results can then be connected to the national sample to see if there is any correspondence/similarity. As can be seen from Figure 3.9 , each of the Census OA areas has very low representation (between 1-4 users per area). Such small representation make any inferences based on the OA characteristics potentially invalid due to the threats to inference

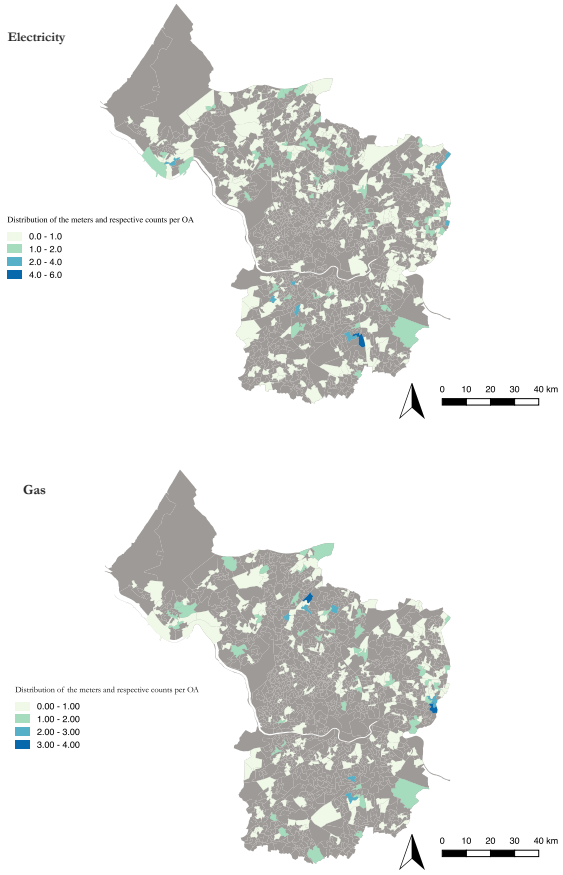


Figure 3.9: *Distribution of the meters by OA, Bristol*

from ecological fallacies (i.e. the smart meter users in the sample may not be representative of the average characteristics reflected in Census data aggregated to OA level). The data grouped to the level of Bristol may serve as a better alternative. The Table 3.4reports the total counts of smart meter users for Bristol area. Having about 1200-1400 smart meter users per city can be a fairly good and representative of the general population in the region.

	Electricity	Gas
Count of unique users	1214	1415
Count of unique Census OA in the sample	948	790

Table 3.4: *Counts of smart meter users and unique OA identifiers in Bristol sample*

3.5.2 Descriptive analysis

The descriptive statistics for the Bristol sample (Table 3.5) appear quite representative of trends in average domestic consumption when compared to government figures. The mean and median consumptions are in fact closer to the national averages reported by BEIS than the ones observed for the national sample in Table 3.2. Nevertheless, what is evident is that the distribution of consumption for both electricity and gas are skewed towards lower levels of consumption compared to measures reported by BEIS. Gas consumption is associated with significant variation (standard deviation of about 22,000 kWH a year) which may indicate the presence of outliers or possible inclusion of non-residential type smart meter users [3].

Descriptive Statistic	Electricity	Gas
Average	3,372 kWh	13,641 kWh
Median	2,879 kWh	8, 825 kWh
Standard Deviation	2,230 kWh	21,563 kWh
Average (BEIS 2015)	3, 894 kWh	11,707kWh
Median (BEIS 2015)	3,148 kWh	13,202kWh

Table 3.5: *Central tendency description* The average annual household energy consumption estimated using the Bristol sample compared to BEIS 2015 national estimates.

As an extension of the descriptive analysis, Census 2011 Geo Demographic Classification data was attached and used to study how representative the Bristol sample is of geo demographic groups that are dominant in the region. The average total consumption per day was selected as a measure of variability. Six subgroups were then attached to the data, given the area of Bristol where smart meter user reside. As may be seen from Figure 3.10 the relationship may appear as vague and not definite as this stage. This may be due to low proportion of the meters compared to total number of people reside in each of the OA (i.e. less then one percent representation). Given the fact, that Census 2011 Classification is obtained using the average characteristics of the residents in the area, having a representativeness

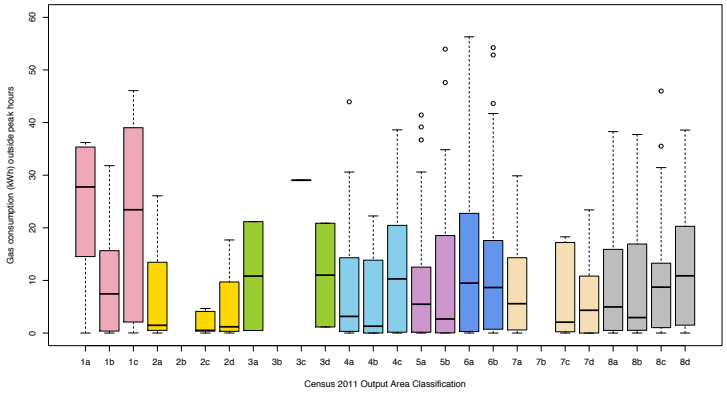
[3] Please note that BEIS values are obtained for 2015 while Bristol sample data correspond to the year of 2014. The reason why the BEIS 2015 values are chosen is due to the fact the the measure of domestic energy use was updated by OFGEM (2015) to reflect a more inclusive average trend in the consumption. This is used of course under the assumption that no significant difference in gas and electricity can be observed when 2014 and 2015 are compared.

of less then 1 percent may be associated with greater uncertainty and should not be used for the inference of why the energy consumption may vary in the sample. Interestingly, other research attempted to link profiled energy consumption patterns to socio-demographic classification and found little correspondence between temporal profiles and socio-economic groups (Haben et al., 2013) . This suggested that studying actual energy consumption at greater temporal breakdown may further inform us about behavioural patterns. For instance, variability in half hourly consumption can be used as an indicator of distinct consumption behaviours or lifestyles that can also challenge current geo demographics classifications completeness.

3.6 Sample Selection

This section looks closely at how best to select samples for more detailed research and analysis. As the data magnitude is large, various samples will be used for the study in this thesis. From descriptive analysis above, one could conclude that the data may be segmented using simple measures of variation as a way to define similarity across consumption records. Such a method was used in the past when only annual energy consumption measures were available to the researchers and the government (DECC, 2013; DCLG, 2015). Having data of greater temporal granularity may extend such analysis: the descriptive statistics such as mean, median and range may be useful to illustrate the overall magnitude of energy consumption. However, such an approach totally hides all the essential information about the dynamics of the energy consumption. The customers with similar mean may have a very different behavioural pattern. Likewise, as the total per day consumption value may be similar overall, it is hard to make conclusions about the similarity of the consumption profile hidden behind the overall measures. To make this point clearer, some illustrations below are presented. Random daily readings were selected from the sample based on either similar average half hourly consumption load, total per day consumption or the variation in consumption across time using the measures of standard deviation.

To study the heterogeneity of the consumption profiles a bottom-up approach



1a - Farming Communities	0.5%
1b - Rural Tenants	1.8%
1c - Ageing Rural Dwellers	1.0%
2a - Students Around Campus	1.4%
2b - Inner-City Students	0.0%
2c - Comfortable Cosmopolitans	0.5%
2d - Aspiring and Affluent	0.4%
3a - Ethnic Family Life	0.2%
3b - Endeavouring Ethnic Mix	0.0%
3c - Ethnic Dynamics	0.0%
3d - Aspirational Techies	0.0%
4a - Rented Family Living	6.7%
4b - Challenged Asian Terraces	0.8%
4c - Asian Traits	1.5%
5a - Urban Professionals and Families	20.5%
5b - Ageing Urban Living	8.6%
6a - Suburban Achievers	8.5%
6b - Semi-Detached Suburbia	25.2%
7a - Challenged Diversity	3.1%
7b - Constrained Flat Dwellers	0.0%
7c - White Communities	0.9%
7d - Ageing City Dwellers	0.5%
8a - Industrious Communities	5.6%
8b - Challenged Terraced Workers	2.2%
8c - Hard-Pressed Ageing Workers	5.1%
8d - Migration and Churn	4.9%

Figure 3.10: Correspondence between the average total consumption of Gas (Wh) per day with Census 2011 Geo demographic Classification

is used. Various descriptive measures of the patterns (i.e. mean half hourly consumption, geographical reference) are fixed at a time and a few patterns are selected to study how similar/different the dynamics in consumption behaviour are. Here, the variation at the individual user level is taken for the analysis. This is contrary to a top-down approach where the analysis begins with the variation at the aggregated national level and then is narrowed down to regions, geographical areas and only then to individuals users (Swan and Ugursal, 2009).

First illustration considers the case where the mean half hourly consumption is the same for each of the readings (Figure 3.11). The second one presents two

readings that have a similar total per day consumption (Figure 3.12).

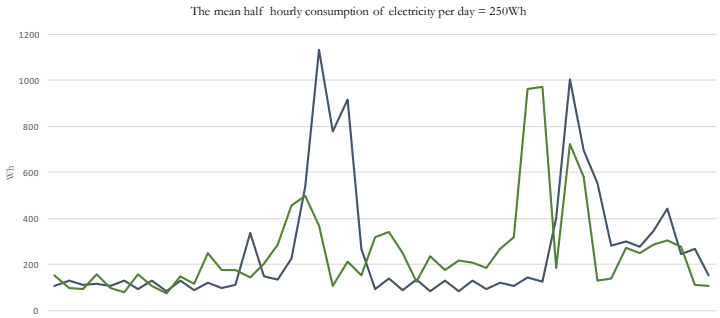


Figure 3.11: *Two randomly selected electricity consumption patterns that can be described by the same mean of half hourly consumption Wh.*

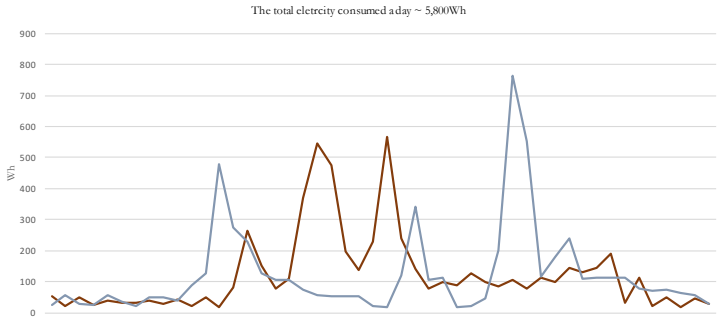


Figure 3.12: *Two randomly selected electricity consumption patterns that can be described by the same value of total per day consumption in Wh.*

As seen above, while numerically the random patterns that were selected can be described as similar, the dynamics of energy consumption are very distinct, no matter whether the mean of half hour energy consumption or total value of consumption per day is chosen. Intuitively, it may be expected that standard deviation measures may be able to capture similarity in the pattern. Below (Figure 3.13) presents the case where mean half hourly consumption, total per day and standard deviation are roughly the same. As can be observed, it cannot obviously be concluded that these consumption behaviours are similar and can be grouped together as identical profiles. The next chapter will look more closely at other methods that can address this issue, such as clustering of the dynamic structure instead of the numerical range of the values occurring during the day.

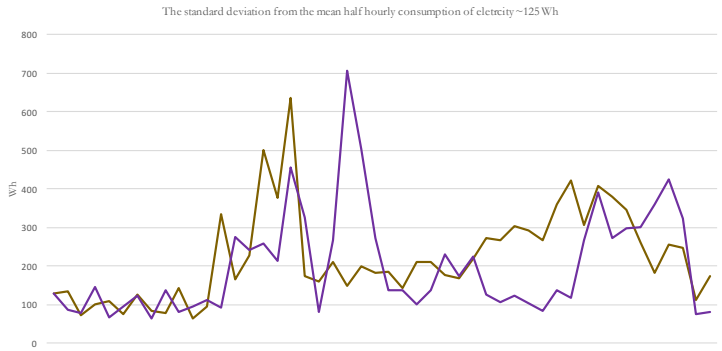


Figure 3.13: *Two randomly selected electricity consumption patterns that can be described by the same value standard deviation from mean half hourly consumption in Wh.*

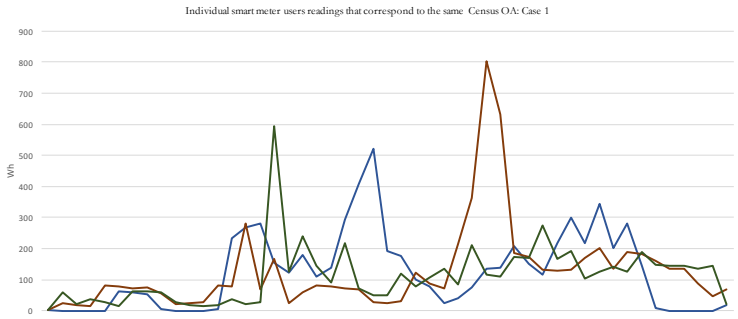


Figure 3.14: *Individual smart meter users readings that correspond to the same Census OA: Case 1*

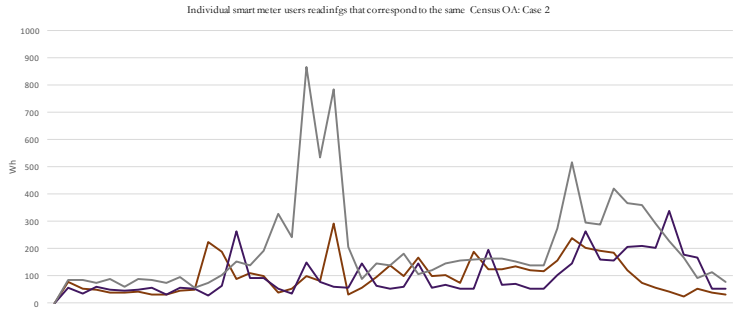


Figure 3.15: *Individual smart meter users readings that correspond to the same Census OA: Case 2*

Similar analysis was performed using the geographical reference as identification. The Bristol sample was selected due to its greater geographical resolution. Three electricity consumption profiles were selected at random from the same OA. The temporal resolution was fixed at the weekday of October. As can be seen, it cant

be assumed that homogeneity within the geographical area is necessarily a property of energy consumption behaviour. Heterogeneity of the consumption behaviour is present both within and across the geographical units.

3.6.1 Aggregation

Different levels of aggregation may be taken at any point of the analysis. For instance, if we are interested in describing the whole national dataset, aggregation may be useful for both data reduction and for the speed of calculations that one would like to perform. The researcher thus may experiment with the unit of analysis being a geographical point of the resident as well as looking primarily at smart meter users as individual units under consideration.

If it is more important to give attention to the context where the unit of analysis exists, then there will certainly be a number of limitations when dealing with smart meter data. A registered individual user of a smart meter can equally be represented by just a single person as well as being associated with an entity of properties being owned. For the national dataset in this thesis, the challenge is the uncertainty over what comprises a user, which requires some assumptions to be made.

3.6.2 Defining outliers

There are number of ways through which outliers can be observed and defined for analysis of smart meter data. For cases where additional data on individuals is not available, the definition of outliers becomes a rather subjective task. Outliers can be defined using both temporal profiles as well as the aggregated values of energy consumption (i.e. annual total gas or electricity consumed). The measures of variability, such as standard deviation from the average energy use per smart meter user can also be a useful indicator of patterns that the researchers may want to exclude or separate from the overall analysis. This is particularly useful for identifying a non-residential smart users profiles (i.e. for instance, under the assumption that low standard deviation is associated with the absence of peaks and presents a continuous behaviour throughout the day). Even more simply, by looking at total per day consumption, the values which are significantly different from the average total per day

consumption reported in government reports for instance may be taken into consideration as outliers. While overall, this thesis will take into account all the data, including outlier or unusual behaviours, it is important to be able to identify and remove if necessary the various cases that may affect the descriptive analysis. Once again, this can be driven by the research question in hand. When one is interested in an average consumption behaviour (i.e. double peaked consumption patterns), the outliers may be removed. Similarly, to understand anomalous or unexpected behaviour the reverse approach can be used. Later in the thesis, it will be shown that using various clustering approaches for instance may help in identifying both types of behaviours. A more simple way could be visualisation of the patterns which then can be judged for its appropriateness for inclusion as usual/average behaviour. Please note that this is possible when the dataset is fairly small (i.e. less then a hundred individual smart meter users with yearly coverage).

3.7 Smart Meter Data and Administrative Datasets

Before proceeding with the discussion of which datasets can aid in the analysis of the smart meter data available for this research, it is important to define the term administrative data. Administrative data is data that is recorded as a result of the administrative systems operations, such as collection of data by various government bodies as well as the records of various transactions (Connelly et al., 2016). While in this thesis smart meter data is treated as an example of retail/consumer data it is vital to acknowledge that there remains ambiguity around the definition of smart meter data as it can equally be characterised as an administrative dataset due to its relevance for both government and energy industry bodies and because of the nature of information it provides about the population. Energy consumption is an essential resource for the countrys residents wellbeing and can be treated similarly as the consumption of education or health resources. The privacy concerns associated with smart meter data can also be considered as similar to those of administrative data. A more clearly defined administrative data source that will be accessed for possible linkage is the 2011 Census (Please see Appendix for some other potential sources

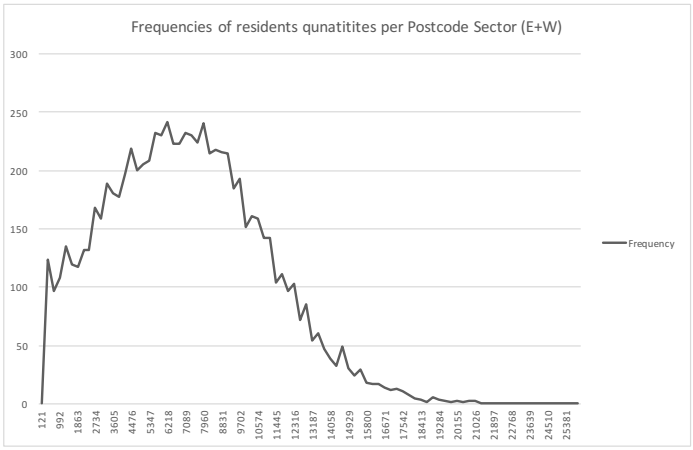


Figure 3.16: *Frequencies of residents quantities per Postcode Sector with mean of 6979. median of 6799 and standard deviation of 3777.* Source: *ONS, 2017*

Postcode Sector (source)	Postcode Sector in MSOA	MSOA
EC2Y8	1	E02000001
EC1A7	1	E02000001
EC1Y0	0.6988818	E02000001
EC1Y0	0.24201278	E02000576
EC1Y0	0.05910543	E02000575
EC4V5	1	E02000001
EC4Y7	1	E02000001
EC1A9	1	E02000001
EC4A1	1	E02000001

Figure 3.17: *Postcode sector and the corresponding output areas*.The table reports the numerical proportion of MSOA that falls into postcode sector. Source: *ONS, 2017*

that can be considered). This data is available at various geographical resolutions and provides a fairly complete picture of the population of the UK in terms of both socio-economic characteristics of residents as well as the description of the housing conditions in which they reside. While this data is certainly rich and useful for the analysis in the thesis, the issue that needs to be assessed is how feasible it is to connect Census 2011 data to smart meter data at the postcode sector level.

In regard to spatial granularity, national data available for this research is at postcode sector level. The average quantity of smart meters for a given postcode sector level is 59 meters with 1 being the minimum and around 393 and 349 as the maximum number of meters for electricity and gas respectively. Postcode sector can be considered as a fairly large geographical unit of analysis in terms of the residents it may be embracing at one time, yet there is less uniformity in how postcode area

borders are defined compared to more socio-economic oriented geographical units such as Census Output Area. Figure 3.16 illustrates the average number of residents that can represent a postcode sector. As can be observed, on average around 7 thousand residents are expected to be in the postcode sector. Given the average of 59 meters per sector in the national sample, the Great Britain sample may represent only about 1-2% of the geographical area. Both visual and numerical assessments of possibilities for joining smart meter data to other datasets at MSOA level is given in Tables 3.17 and 3.19 and Figure 3.18.

Figure 3.18: *Postcode sector and the corresponding output areas for Bristol* The bold lines show the postcode sector boundaries that are mapped on top of Census OA boundaries. As can be seen number of OAs falling into postcode sectors varies significantly. Within the city centre for instance we can see very high density of OAs while less is observed within the rural areas. Source: *ONS, 2017*

In terms of corresponding MSOA areas to Postcode Sector it is observed that there are about 26741 areas for 8000 sectors. While there are some postcode sector which fully nest in MSOA, on average only 30% of postcode sector can be represented by MSOA. This is certainly limiting also due to the fact that nesting properties are not uniform across postcode sectors and MSOAs.

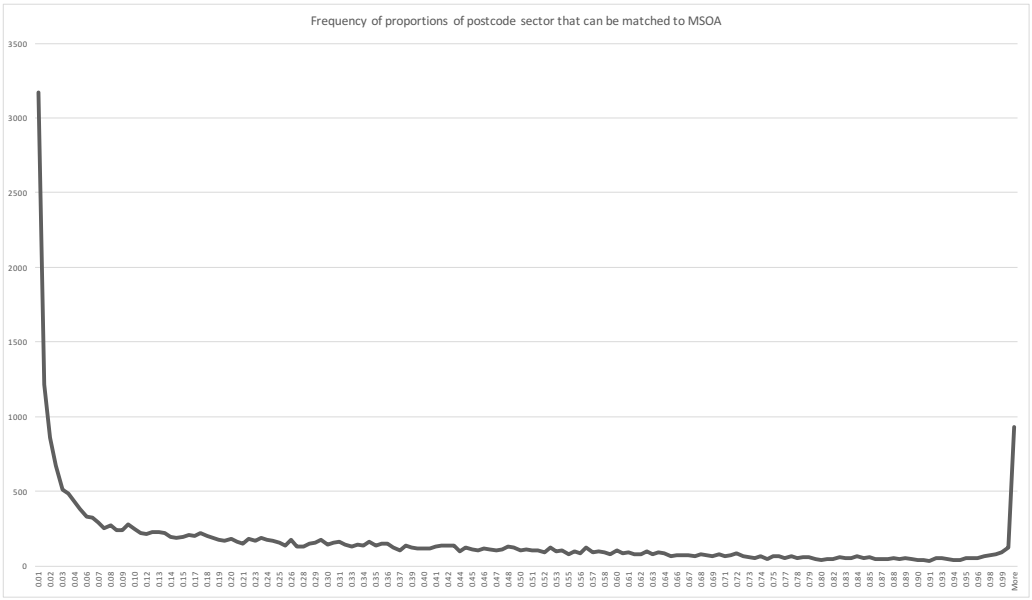


Figure 3.19: *Frequency of proportions of postcode sector that can be matched to MSOA for the whole United Kingdom* It can be notes that most of the postcode sectors can be matched to only up to ten percent of corresponding MSOA area. Source: *UKDS, 2017*

A possible alternative solution is to use static characteristics of the output areas. These are housing conditions and areas that can be differentiated as urban vs rural, student vs working population areas, etc. A very useful dataset to achieve this goal is a geo demographic classification of UK areas that controls for both areas and individuals socio-economic characteristics and presents an aggregate measure to differentiate regions on both large and small scales. The challenge is to find a way to aggregate characteristics for each area that can be derived from a Census. As can be seen, the proportion of postcodes in MSOA areas vary significantly, with more than a half being under 50 percent. It can be feasible to first focus on areas that have more complete matching and then interpolate to the ones with smaller proportion. Being able to test for heterogeneity across the areas that are being matched together with will be crucial for such kind of the analysis. This thesis will not be considering linkage on such level but further work may consider transforming postcode sector areas into smaller or larger geographical units that can be more easily connected to socio demographic data on the UK population.

One of the alternative datasets and, perhaps, one of the few large datasets avail-

able at postcode level geography that can serve as a useful contribution to energy consumption data is provided by recently released Domestic Energy Performance Certificates. The data that is available is at the postcode level and holds various information such as type of property, floor size, number of occupied/heated rooms and most interestingly, estimated fuel costs. Energy Consumption data at postcode level can be easily linked to the certificates as it will directly nest geographically. This would solve a couple of limitations outlined above that are related to linking the smart meter data to census output areas as well as allowing for clustering of the data with additional features that may be explanatory for the differences in variations. However, while theoretically this idea sounds very promising, the limitations and issues inherited in EPC data need to be studied. As the data is fairly new and novel to academic research, the amount of investigation required may be equivalent to that performed in this thesis on smart meter data. Certainly, a more common source to consider would be Census. However, linking to datasets such as Census may be a slightly more challenging task in comparison with EPC simply because of geographical granularity. However, a useful technique such as estimation of local heterogeneity may help one to define areas that can compose overall postcode sector characteristics based on similarity at lower scale geography. In other words, if the areas which are nested in postcode sector have similar attributes or are homogeneous, it may be assumed that postcode sector area on average will also be associated with those attributes either in terms of housing conditions or household characteristics.

This section looked at the possible linkage only briefly as to provide a stepping stone on how these data may be taken for further analysis. The final chapter of the thesis will return to the discussion and investigation of these in more detail once the capabilities of sole smart meter data in revealing information about UK population have been shown.

3.8 Ethics

'Just because it is accessible does not make it ethical'

- Boyd and Crawford (2012),

This section considers the ethical issues that may be associated with research involving smart meter data. Ethical constraints are mostly evident from the restrictions on the data resolution available to researchers. As was seen from this chapter, while smart meter data may have greater temporal granularity, there is certainly a lack of geographical resolution, given that few smart meter users fall into fairly large geographical areas in the datasets studied in this thesis. There are a number of legal and ethical issues that are associated with the access to individual smart meter data. The summary of some of the most commonly identified concerns are adapted from McKenna et al. (2012) and presented in the Table 3.6.

To protect the consumers from possible issues that are determined by current research, the available data faces a compromise of either greater spatial or temporal granularity. Having both may potentially threaten the individual smart meter users privacy.

While the table above may present some of the very threatening consequences of access to smart meter data by third parties, not all of them are certainly observed in the real world. Commercial uses can be most common as leveraging insights from smart meter can aid the effective operationalisation of energy provision, product advertisement and marketing campaigns. As a preventative measure for more serious and sever consequence such as use of smart meters as legal evidence of the person being at home or used for spying on each other by household members, the restrictions posed on smart meter data sharing are generally high. In the UK, after installation of the smart meters, energy customers agree by default that their data may be accessed by the energy supplier. However, they do have a right to opt out from their data being shared with energy providers or the government if they wish to do so. Smart meter users can also select the aggregation of their energy consumption readings. For example, share only weekly total consumption, monthly or annually [4]. One of the obvious observations of the work presented in the thesis so far

[4] This was true for the time of the study, 2015. Due to various General Data Protection Regulation (GDPR) considerations, which came to effect in May, 2018 and given the fact that smart meter data can be treated as personal data, the arrangement have changed by giving customers more extensive

Application	Example	References
Illegal uses	Burglar finding when homes are unoccupied; stalkers tracking the movements of their victims	Lisovich et al. (2010); Quinn (2009); Cavoukian et al. (2010); McDaniel and McLaughlin (2009); Lerner and Mulligan (2008); Subrahmanyam et al. (2005)
Commercial uses	Targeted advertising: use of individual or aggregated household smart meter data to target advertising at a specific household or individual	Lisovich et al. (2010); Quinn (2009); Cavoukian et al. (2010); McDaniel and McLaughlin (2009); Anderson and Fuloria (2010); Bohli et al. (2010)
Use by law enforcement agencies	Detection of Illegal activities (i.e. sweatshops, unlicensed commercial activities, drug production); verifying defendant's claims (i.e. that they were 'at home all evening')	Lisovich et al. (2010)
Uses by other parties for legal purposes	In a custody battle: do you leave your child home alone?	Quinn (2009)
Use by family members and other co-inhabitants	One householder 'spying' on another (i.e. parents checking if their children are sleeping or staying up late playing video games); partners investigating each other's behavior	Hargreaves et al. (2010)

Table 3.6: *Privacy concerns related to smart meters (Adapted from McKenna et al. (2012))*

is the limitations imposed by the geographical resolution available for this research. While data is rich temporally, given its continuous nature (i.e. readings are associated directly with the activities within the entity of the smart meter user), trade-off between temporal and spatial granularity is inevitable. This trade-off however allows for generating general insights about individuals without compromising their

rights with respect to the data they share. They can opt in into data sharing and they also have a right to be *forgotten* under GDPR by requiring the energy supplier to erase their past data. The debate on smart meter data being too sensitive for energy customers privacy and how this should be addressed under current regulations remains open. For an excellent review of the privacy issues in smart metered future please see Véliz and Grunewald (2018)

anonymity. This is vital for making sure that this research is performed in line with ethical standards.

3.9 Conclusion

In this chapter the data that will be used for the analysis in the thesis was presented. The basic descriptive analysis such as the counts of smart meter users, average dispersion around the mean energy consumption for both gas and electricity and a brief comparison of the figures to overall population estimates reported in government re- ports were presented. As observed, both the national and Bristol sample are associated with a high temporal granularity yet suffer from rather low geographical resolution. This, as identified above, may be driven primarily by ethical reasons. Due to the potential threats that can be posed to privacy of the smart meter users in the cases where data is of high temporal and spatial resolution, compromise between one and another is inevitable. This is by no means a limitation. The work in the thesis will therefore be focused largely on the temporal granularity and allow for generation of rather general insights about energy consumption in the UK without compromising the anonymity of the users. It was shown that the national sample may have a potential to be analysed from a geographical point of view (using postcode sector level). Given how few smart meter users are presented in each of the Census OA in Bristol, Bristol was chosen as a unit of geographical analysis, with the representativeness of the results then being assessed using the national sample. To study the national variation in energy consumption on a geographical level, it was shown that even using MSOA may be a challenging choice as there is no uniform correspondence between postcode sector and the output areas. Dealing with datasets of large scale inevitably invites thoughts about possible data reduction measures that can be applied to specific samples to be analysed at various geographical or temporal levels. The choice of sample may be driven largely by the question in hand. For instance, for understanding consumption in urban areas and its trends, it may be necessary to restrict the analysis to areas within cities. Choosing areas where certainty about dwelling stock and household socio-economic characteris-

tics is greater, may be useful where smart meter data on energy consumption at a household level is missing. Complementing the smart meter data with information on dwelling stock and socioeconomic characteristics allows for a more contextual analysis of energy usage and unlocks possibilities for inference on why consumption may differ according to geographic areas. Given the sample that was illustrated in the chapter, this may however be challenging. On the other hand, having a large enough sample that is aggregated at lower geographical resolutions also has the benefit of retaining anonymity so there is no need to be concerned about privacy issues or unlocking individual data.

Chapter 4

Methodology and Results: Preliminaries and Clustering

> *The goals in statistics are to use data to predict and to get information about the underlying data mechanism. Nowhere is it written on a stone tablet what kind of model should be used to solve problems involving data. To make my position clear, I am not against data models per se. In some situations, they are the most appropriate way to solve the problem. But the emphasis needs to be on the problem and on the data*
>
> - Breiman (2001b),

4.1 Introduction

Conventional and traditional statistical methods available to social science researchers tend to be employed in research areas that use small, yet very clearly defined samples of data (Breiman, 2001b). These samples are often easy to collect by the researchers themselves or can be available via open source repositories due to the simplicity associated with these data' storage and management. Some of the most detailed data that can be available on population is that which is collected using targeted surveys. While this manual collection of data allows for more precise and specific answers to research questions, one has less flexibility in reusing these data to answer other research questions that may arise during the research process,

or which are incidental.

In the case of big data, the issue is rather the opposite. There is a greater flexibility about which kind of insights that can be gathered from the data, yet this flexibility can be problematic if there is a very specific question one wishes to answer. As a consequence, it may be more appropriate to let the data tell us what may be discoverable and then operate within those limitations (Kitchin, 2014; Anderson, 2008). Smart meter data is no exception. Before even looking at the smart meter data, one may think of various research questions that can be possibly answered with these data. For instance, the goal may be to assess whether it is possible to infer differences in energy consumption based on the characteristics of the regions where energy consumers reside; or to estimate the effect of living in rural or urban areas on energy consumption dynamics. While these questions may sound straightforward, when actually looking at the smart meter data directly one realises that unless there is knowledge of key variables such as the household characteristics, or the conditions of the property they reside in, these data may not be as informative as first thought.

In a survey, for instance, it is possible that a great deal of contextual information about the person and their house is collected during the personal interview. On the other hand, one may fail capture the dynamics in their activities - which can be for instance present more clearly in smart meter data. Relying on smart meter data over information collected through methods such as surveys also has the benefit of a certain reliability. It can not easily be faked, mistakenly answered or based on a misunderstanding. The point here is that having both smart meter data and survey data is an ideal methodological scenario that minimises the uncertainty or unreliability of either data source.

This section will look more closely into how valuable insights can be gained from smart meter data alone by applying various statistical methods to first identify, and then evaluate the patterns in the data to provide descriptive statistics. The reasons this methodological approach utilises statistical learning (and by extension machine learning) is firstly, due to the magnitude of data and the subsequent inabil-

ity to manually describe or identify patterns within it. Secondly, recent progress with machine learning in other domains has demonstrated the ability to generate valuable insights from large datasets, where manual exploration is difficult or impossible (Sebastiani, 2002; Fan and Bifet, 2013; Chen and Zhang, 2014; Witten et al., 2016). The variety of these methods allows for greater flexibility as well as these approaches reduce the need to make unrealistic assumptions which do not necessarily represent the scenario studying. It also allows for data of a broad range of shape, magnitude, and process to be considered for analysis.

Before looking at these models more closely, there is an interesting paradox to consider. Most social scientists would say that greater variability of data is preferable for research as it allows us to capture the casual mechanisms and relationships with more certainty. However, this argument only holds where the inference can be generalised across all or many samples in the data-set. In the case of energy data, and particularly smart-meters, one has to contend that there will be vast differences in the data-generating process across consumers. Such heterogeneity in the population makes causal analysis challenging, as patterns averaged either across consumers, or even time, may not converge to the population expectations. These challenges largely motivate the work in this chapter, which aims to dissect this variability and characterise the data-generating process.

4.1.1 Structure outline

In this chapter, the methodological tools that were implemented on the smart meter data will be discussed in details. One of the goals of this research project is to find the optimal ways to describe smart meter data, classify the patterns in an informative way and finally, devise a design that can help in solving the task of prediction of energy consumption using both pattern predictions and exact point prediction tools. All of these achieved under conditions where no other data is available. This may be particularly relevant for energy companies analysts and governments who are interested in the automated process of data analysis in real time. Anticipating the fact that the smart meter data collection is expanding each day with more readings being stored, solutions that can turn these data into insight using their raw form may

be highly valuable. Lastly, the choice of methodology is driven by generalisability of the results whilst ensuring flexibility of the assumptions on which models are based. This is crucial as it will be shown that data and its structure varies significantly from user to user as well as there is a huge variability within individual users consumption patterns.

This chapter will be focusing primarily on the immediate description of the patterns and classification using clustering. The description of the data from the point of view of the associated generating process (i.e. spatial and temporal stationarity) will also be discussed. The remainder of the chapter is structured as follows. The first part discusses various approaches that can be used to analyse datasets of varying complexity by paying specific attention to the trade off in the complexity of the models and their interpretations. This is followed by a preliminary analysis which examines the casual mechanisms that are based on probability; a time series analysis; as well as an introduction to spatial heterogeneity and correlation.

A run through of the research design, that aims at segmentation of energy consumption patterns, is followed by a discussion on the suitability of the various clustering techniques in the section 5. The results of clustering will be presented for each of the samples: national case and Bristol sample. Further assessment of how aggregation affect clustering results will be discussed. This will be followed by the test of results predictability as to evaluate how well the approach taken for clustering can classify the new data into the obtained groups.(Section 6-8). This will be followed some possible extensions to the clustering analysis such as 'use out of peak hours' classification in Section 9. The final section, that will discuss the results and limitations of the experiments, concludes the chapter.

4.2 Statistics and Machine Learning for Smart Meter Data

As was seen in the previous chapter, a very simple approach to describe the data (i.e. mean, standard deviation) may serve as a useful tool to illustrate variation in the data on a small scale, yet it may fail to describe the unique dynamics that may be

associated with consumption patterns at the level of individual users. As a solution to this problem, in this chapter, applied machine learning methods will be surveyed on samples at both a national and individual city (Bristol) level. Prior to this, this section will discuss the very fundamental and even, philosophical, consideration of what constitutes a good empirical model. These are considerations on how to find an appropriate statistical model/process which can describe the smart meter data, and fundamentals of pattern recognition and uncertainty. These discussions will serve useful not just for this particular chapter but for the rest of thesis.

4.2.1 Distinction between natural and statistical approaches to study the data

This section provides a brief overview of how one may look at the data prior to the analysis. Focusing more on a problem in hand, the thesis is partially influenced by the work of Leo Breiman, a statistician who after spending a significant amount of time in industry has developed a very applied and pragmatic way of looking at applied statistical science.

There are clear distinctions between the ways in which one considers data when it comes to natural, statistical and machine learning approaches. Each approach has its place in smart meter data analysis and the appropriateness will depend solely on the research question in hand. Breiman (2001b) separates statistical modelling into two cultures. One of them always assumes the data was generated by some specific process or in other words a stochastic model. Another one, and perhaps favoured by Breiman, is where a researcher uses various algorithmic models yet assumes that the data generation process is unknown. The former process, as Breiman suggests, often leads to rather irrelevant theories which as a consequence alienate the performed statistical analysis from more complex and real world problems as the assumptions behind these data models are almost impossible to meet.

A natural approach to the data generation process is to assume that x and y can be related to each other in the following fashion: the predictor variable x and outcome variable y. Imagine, that this could be consumption readings x and the clusters to which the algorithm aims to assign these readings to y. What is important to un-

Figure 4.1: *Two ways to represent the relationship between input (predictor) and output (outcome) variable* Regression analysis represents an interpretative approach while 'unknown' is referred to black box solution, where process that connect x to y exists but cannot be described using modelling language

derstand the nature of the mechanism that describes the associations between the two best so it can be reproduced and replicated. In Breiman (2001b), the author assumes that there is a black box connecting the two and the distinction among the approaches lies primarily on the contents of this box. For instance, one may have a parametric regression analysis that will hold some knowledge about the mechanism and will have some interpretative power while preserving the simplicity of the relationship between x and y; the alternative way is the unknown, here represented as a black box solution. Example of black box methods, such as a neural network models, are usually associated with good accuracy of y prediction based on values of x but have high complexity and low interpretability.

The work of Breiman further highlights a number of rules or perceptions which are critical for the performance of simple yet reliable statistical analysis. These are: (1) the primary objective should be to find a good solution that will presumably hold for a long period of time; (2) prior to modelling, a significant amount for time should be spent working with the data to get a sense of it and its inherent dynamics; (3) a model with a solution is to be preferred; (4) the error on the test set always a provides some measure of the suitability/efficiency/success of the model. While there can be disparities among statisticians on whether Breiman's set of rules is even feasible, this thesis takes the side of Brieman and aims not just to provide methodological solutions that may withstand the time, but also learn a great deal

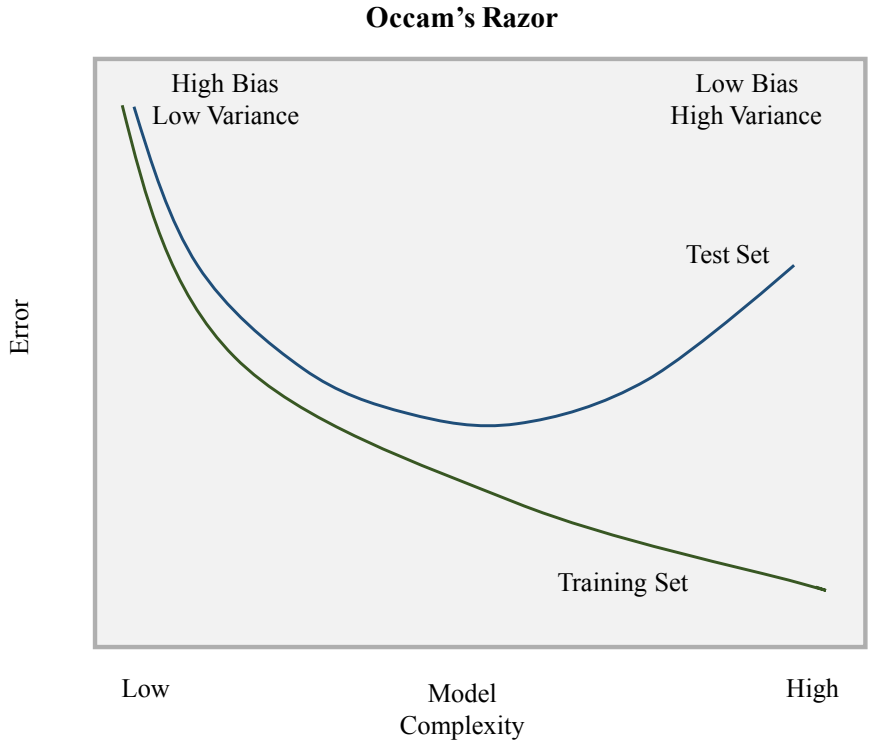


Figure 4.2: Occam's Razor representing the trade off between bias and variance and its effect on error of prediction as the complexity of the model increases. *As can be seen Test set error will be more sensitive as training error gets smaller and smaller implying that the data is over-fitted by the model and while model is highly complex and prediction error is low on a training set, the model will not perform well on a new/unseen sample.*

about the data, its dynamics and the statistical processes which may underpin the uniqueness of temporal profiles.

One of the formal ways to access the model performance critically is known as Occam's Razor. The concept is discussed in more details in the next section.

4.2.2 What Constitutes good Modelling?

This section complements the discussion above and moves on from conceptual discussion on what constitutes a good modelling framework, to more practical aspects. One of the common issues that statisticians look out for when choosing an appropriate model is the trade-off between bias and variance. Bias occurs where a model may be too general such that it misses important and distinct features required to model relationships in the data. Variance occurs in the opposite extreme scenario

where a model overfits the relationships in the data due to being too specific to each unique feature in the data, yet lacking the ability to generalise to other samples.

One solution to the bias and variance trade off is through methods referred to as cross validation. Additionally, this may be complemented by regularisation that will allow for introduction of penalties that restrict overly complex models. Both of these methods will be discussed in more detail later in the thesis, as their specific use will vary depending on the methodology under consideration. For broader illustration, one may consider the trade-off illustrated in Figure 4.2, where this phenomenon is described by Occam's Razor. Occam's Razor; the idea that all else being equal (i.e. predictive power) the simplest model is preferable, can be used to illustrate why lower model complexity may sometimes be preferred. In practice, this means that results should be able to be generalised to new/unseen data, represented here by the test set.

A final goal of any modelling, is to develop an objective way to help mimic and reproduce data and relationships in the data, as closely to the way it is produced in the real world. For instance, if one learns that students who have fully attended classes and have received more details on the assessment's aims and structure during the seminars; are these students more likely to score better in the final exam? To study that, one may simply design a model using the data on attendance and final scores and model this relationship to assess whether attendance and being fully informed are sufficient predictors of final scores.

In the case of energy data, to start with, it is known that time plays an important role for energy consumption variability and that depending on the time of the day, different behavioural responses can be observed in energy consumption levels. While the relationship between time of the day, and variation in energy consumption is rather self explanatory, one may further consider regional effects, or how habitual the consumer is in their energy use as monitored by the smart meter; for instance, does consumption of five weeks ago have anything to do with the consumption of the current week. By doing so, it may be possible to design a simple model in terms of interpretation, yet some mathematical developments may be needed to reproduce

the dynamics that generated consumption differences.

4.3 Smart Meter Data as Time Series Data

After looking more broadly at how various data can be treated from statistical point of view, we may now narrow down to the discussion of how smart meter data can be described using statistical processes. Most attention will be given to understanding what gives this data some of its unique characteristics. This is mainly due to the time component inherited in the nature of smart meter data.

This section looks at the smart meter data treated as a time series process. It is important to define the data in terms of its statistical features/characteristics prior to any analysis as this is what essentially motivates the choice of appropriate model. Stationarity in terms of both space and time will be discussed with a particular focus given to heterogeneity issues. While statistical tests and diagnostics for stationarity are widely researched and established, assessing the heterogeneity of time series patterns is still a topic of interest.

Time series data can be characterised as a sequence indexed by time and values of of the variables of interest. For example, this may be denoted $x_1, \ldots, x_T$ where variables are indexed for time-points $t = 1, \ldots, T$. For smart meter data there are various ways to represent time series sequence. For instance, one of the sequence could be a total consumption per day, while another sequence example could be half hourly energy loads. Depending on the way one views energy consumption readings, different data generation process may be attributed. In this section, the challenges presented by smart meter data are discussed when looked at using traditional time series assumptions and possible alterations. The aggregation level of smart meter data matters significantly for the results observed. This is valid for both segmentation and prediction and was discussed previously in the context of different time series data analysis by Shellman (2004). Lastly, in case of smart meter data and meter users one may analyse the data in both univariate (a single data-stream) and multivariate (multiple data-streams) setting. If one is interested in the prediction of average energy demand that is composed or consumption by

individual users it may imply a pooling of a large amount of individual series that may be correlated or be totally independent of each other. On the other hand, in the univariate case each customer's time series is taken separately for the attempts to predict their consumption using only their own history of behaviour.

4.3.1 Stationary time series

A very simple model for energy consumption at the granularity of smart meter is given by a sequential chain model (Figure 4.3). If one considers the arrows in the figure to represent conditional dependency relationships between random variables, then this Figure only depicts single step dependencies, i.e. the current value only depends on that directly before it. Such models are often referred to as Markov chains, and higher order chains may also be specified such that x_t may depend on more than just x_{t-1}. For instance, a second order chain would mean x_t could depend on x_{t-1} and x_{t-2}.

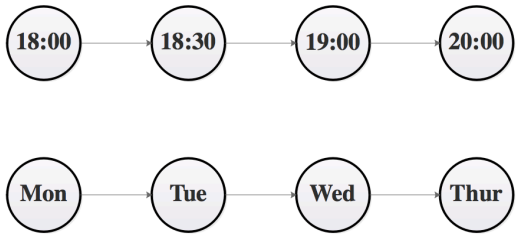


Figure 4.3: Chains based on half hourly readings and on total per day readings

The illustration above is rather very simplistic yet useful for setting the scene for initial analysis of the data. One thing that is immediately noticeable from the sketch in Figure 4.3 is that the dependence pattern stays constant as a function of time, i.e. there is always only a one-step dependency. Typically, when one fits statistical models, it is assumed that the value of the parameters stays constant as a function of time, a setting which closely relates to the ideas of heterogeneity and stationarity.

There are multiple definitions of stationarity. A *strictly* stationary process is one in which the joint distribution of the variables $\{X_t, X_{t-1}, \ldots, X_{t-s}\}$ remains the

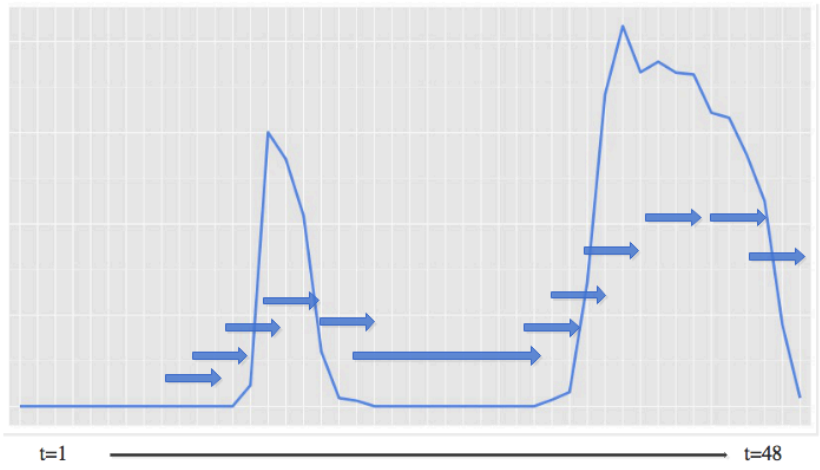


Figure 4.4: *Decisions are made at each* ***t***

same for any time shift k, i.e.

$$p(x_t, x_{t-1} \ldots, x_{t-\tau}) = p(x_{t+k}, x_{t-1+k} \ldots, x_{t-\tau+k})$$

for all lags τ. This is a very strong requirement, so frequently it is assumed a process is a weakly, or covariance, stationary process. In this case, it is assumed that the mean and variance remain constant as a function of time (see Wooldridge (2015) for more on definitions). Such processes are very convenient for providing statistical guarantees on the estimation of parameters, as they effectively mean that more can be learnt about the process as one collect more and more data. That is, as long as the complexity of our model grows slower than the rate of data increase, one should be able to get better and better estimates of the population parameters.

However, such stationarity assumptions pose a large challenge when working with smart-meters. For instance consider the decision making process of a consumer throughout the day, c.f. Fig. 4.4. The user will typically use energy when required, they make decisions which are impacted or motivated by a vast array of scenarios in their daily life, from drying their hair, to turning the heating up when its cold. Since, these all vary on a daily, weekly, and monthly level, if all one gets to observe about this consumer is their energy usage, it might be expected that this series be highly non-stationary.

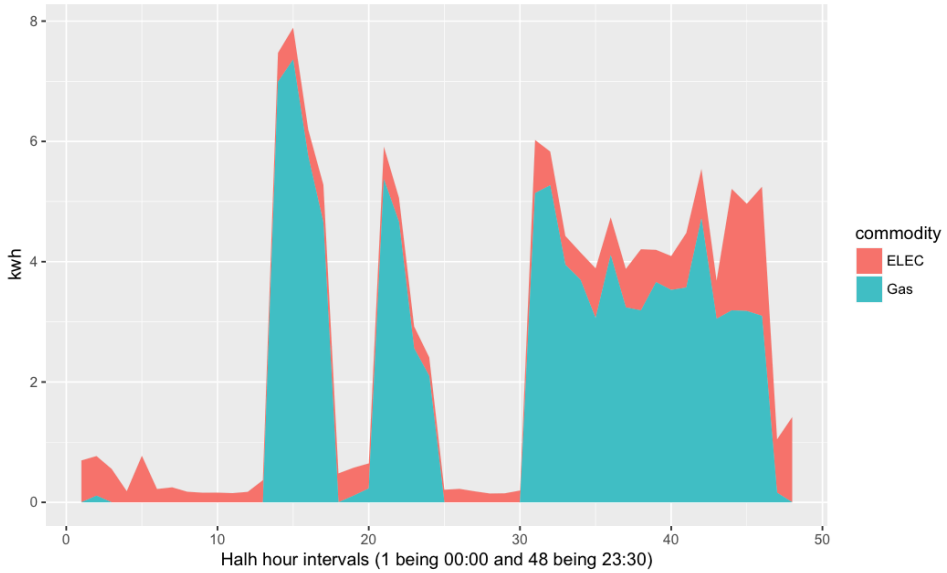


Figure 4.5: *Combined electricity and gas consumption for a random individual.* As can be seen electricity vary at much lower scale in comparison to gas.

The illustration may be further expanded to the integration of two series simultaneously (gas and electricity). The smart meter time series sequence may potentially be analysed simultaneously with other time series such as the weather for example. The most simple co-integration is in fact adding two energy sources together: electricity and gas. While expected to be highly correlated for some consumers, one may see different type of relationships where one of the sources may be more static and follow stable trends through the year while another source may be more variable. In time series statistics, the term co-integration is used for modelling relationship between two or more time series. For instance, rather than attempting to fit the model to predict a single point of the time series, one is interested in modelling the combination.

Failure to diagnose correctly stationary or non-stationary process may lead to false inferences about the phenomena under the study as well as reliability of the statistical results. In this chapter it will be investigated how the sampling frequency (or equivalently aggregation level) of data affects the output of analysis. It will further be shown that it may be safer for researchers sometimes to focus on methods that do not rely on stationary behaviour as this may affect prediction results.

Sometimes, while a process itself may not be stationary, its differenced pro-

cess can be. That is, the process $z_t = x_t - x_{t-1}$ for all t may be stationary, even if x_t is not. One example of such a process is the Autoregressive Integrated Moving Average (ARIMA) model (Saboia, 1977). A quick examination of the smart-meter data however, shows that such a process may not be useful in our case. For instance, consider Figure 4.6 which plots the differenced smart-meter data over the course of the day. Whilst it can be observed that many of the peaks are removed, there are still larger periods of volatility surrounding the peak regions, these correspond to the associated increase/decrease in consumption during these periods. A simple calculation of the empirical auto-correlation function (that is an estimate of the auto-correlation) is given to the left, and demonstrates that even in the differenced sequence significant autocorrelation (and thus dependency) still exists. While one could fit an ARIMA model to this process, through for instance the Box-Cox methodology of iterative model-building-testing (Box and Cox, 1964, 1981), the appropriateness of this model class would still be in doubt as the non-stationarity still persists even after differencing. Alternatives to the simple autoregressive classes of model, with less strict assumptions are considered later in this chapter.

4.3.2 Spatial stationary process

As with temporal stationarity, heteroskedasticity, or in other words, violation of the assumption of the constant variance of the series, poses a threat to the inference. In the case of daily energy consumption pattern, one may observe differences in variances around peak hours of consumption such has morning and evening where more activity maybe observed. One may also observe that there is increased variability throughout weekend days for instance.

In the case of spatial stationarity, the definitions are similar. The observations of the temporal processes will depend directly to how far are they from each other in terms of respective distance Cressie (1988). For more formal definition see below where L represents a set of possible spatial locations and $l_1 - l_2$ is the respective distance between two locations:

$$\text{mean}[x(l)] = \mu \quad \forall l \in L$$

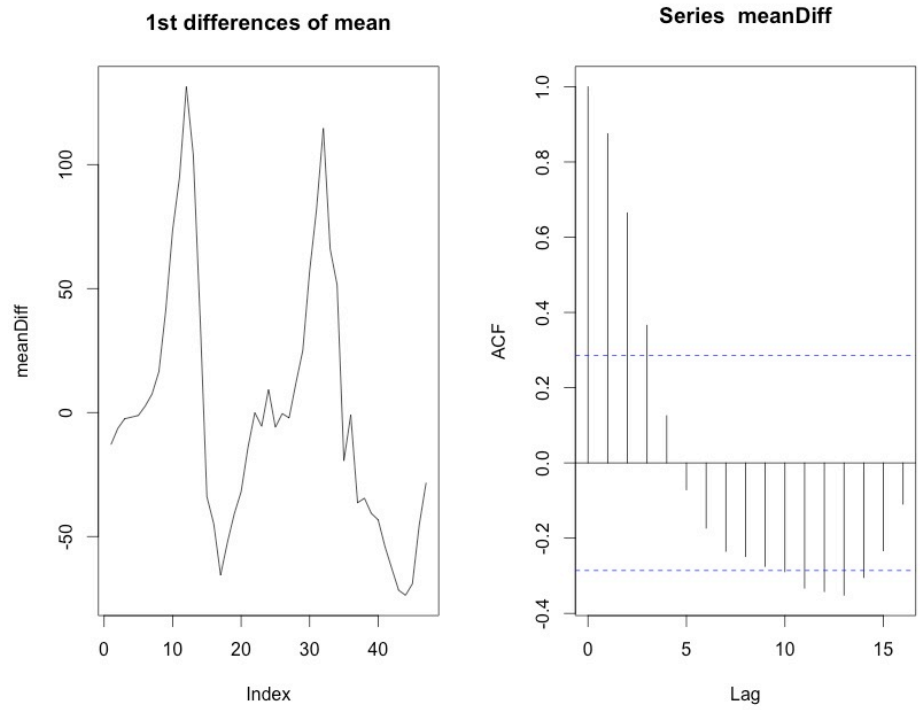


Figure 4.6: *First difference of the smart meter data time series*

and

$$\mathrm{Cov}[x(l_1),x(l_2)] = C(l_1 - l_2) \quad \text{where} \quad \forall l_1, l_1 \in L\,.$$

Similarly, to temporal stationarity, constant mean and variance of the series are assumed but now across the space. The only difference is that the dimensions in spatial case may be extended to more than two. As is commonly seen with temporal auto-correlations, that points in time are more dependent, near objects in space are also more like to share some common information. Spatial stationarity can be violated in nature when the dependence properties at different distances $l_1 - l_2$, depend on the absolute locations of l_1 and l_2, not just their relative positions. Generally, the dependence structures expected are in line with the Tobler rule of geography:

'Everything is related to everything else, but near things are more related than distant things' Tobler (1970).

To measure the degree of spatial autocorrelation, some of the most common tools are Morans I (Moran, 1950) and Gearys C (Geary, 1954) with latter being

used most widely (Haworth, 2014).

4.3.3 A Pragmatic solution for assessing stationarity

At the end of the day, there are various ways to diagnose whether the processes one observes are stationary or non-stationary. For instance, one could use a test (i.e. Dickey-Fuller test (Dickey and Fuller, 1981) for stationarity on each consumers time-series independently. In general, it may be expected that different users behaviour be more or less stationary. This makes formally modelling the observed data using traditional time-series models, for instance the ARIMA model, challenging and time consuming if we were to apply the analysis on each individual time series.

As an alternative to performing such extensive model construction and testing for each user, this work first attempts to cluster users according to similar energy use dynamics. We would like to move away slightly from the strict assumption of stationarity to be able to include customers with very heterogenous behaviour that may not fit standard parametric model structures. While such structures may be easy to identify in the small samples, it may be more challenging to identify those in large samples of data were manual inspection may be nearly impossible.

Pragmatic solution, as the name of the section suggests, could be to look at the patterns as a combination of data points that can be studied from data generation process point of view. By operating this way, no strict requirements on time and energy consumption relationship are imposed. As was shown in the previous chapter (Data), smart meter data can be described as highly heterogeneous and challenging process to generalise about when available in very large quantities. To address this, clustering methods will be applied to data. These are presented by unsupervised machine learning techniques, often used as a way to group the data so it is more intuitive for visualisation and allocation of sub samples that can be used for more detailed research.

In the next section, Gaussian Mixture Models are presented as perhaps the most optimal among available clustering methods to be applied to smart meter data. The best features of Gaussian process based models is the ability of the mechanism

to change the model parameters as more data arrives. Gaussian process clustering is a probabilistic approach that accounts for the uncertainty in the decision of certain class to be assigned. Segmentation in its nature may be obtained using generative approach (modelling conditional on other classes distribution) or discriminative approach (modelling the probability of the class directly).

One may say that if these models sound so appropriate for large and complex datasets, what may appear counterintuitive why they are not used that widely. The issue in hand is perhaps computational as the requirement of the process is the ability to handle inversions of very large matrices and for instances of very big datasets we may need extra operational capacities. We will attempt to see the consequences of reducing the sample to aggregated energy use as a way to address this and compare the results with diss-aggregated data results.

Several and very common methods for clustering the time-series are assessed from a computational and empirical view very briefly, to convince the reader that Gaussian Model may serve as best compromise between very simplistic approaches and very precise time series analysis that was discussed earlier with respect to stationarity. The methods that will be discussed are known as k-means clustering. They are often one of the easily available approaches that applied data scientists tend to look at. Given the very thorough discussion of the nature of smart meter data it is suggested that extra caution need to be taken when those methods are considered. A connection to the traditional statistical analysis of time-series, as discussed in this section, is made via the use of Gaussian Mixture Models in Section 4.4.2.

Next section will look at clustering more broadly as a method to group patterns and characterise data in a systematic way.

4.4 Clustering

The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true

believers who have experience and great courage.

- Jain and Dubes (1988),

This chapter asks the question whether clustering is, interestingly, even possible for energy consumption data. As seen from the literature review, one of the main tasks in smart-meter data research is the segmentation of energy behaviour and customers characteristics using clustering methods. However, the samples tend to differ as well as representations and additional features that are used in conjunction with the energy data. The clustering techniques in this chapter were applied to analyse samples both at a national and city (Bristol) scale level. The main task of this chapter is the creation of artificial labels that can characterise similar patterns of energy consumption.

There are number of methods that are available for the segmentation of time series data using large datasets. These are primarily designed using various statistical distances that can help the statistical algorithm to group the data points based on their similarity. In this section these methods and associated limitations in the context of energy data analysis are presented . Firstly, the intuition behind the clustering methods; K-means and Gaussian Mixture modelling, is discussed and the techniques are compared.

Clustering is an unsupervised machine learning method that is used primarily to discover the underlying structure of the data that have no labels a priori. For example, having solely energy consumption recordings, one knows little about whether the consumption patterns may be aggregated into groups based on similarity. For instance, are people who work full time grouped together, while those who stay at home throughout the day may also be identified in the similar group. The objective is thus, to find an algorithm which would ensure that similarity between individuals within each cluster is maximised, while similarity between clusters is minimised.

To date, a number of methods were developed for clustering the data. The majority of these have been shown to give reliable performance on static data Liao (2005), however, often disregard the dynamic of structural groupings, posing chal-

lenges when considering spatial and temporal dimensions in the analysis. Fortunately, there are number of pattern transformations techniques that may help in preparation of the data for the algorithms that are developed for the static data. One of the immediate solutions could be to transform dynamic data into the static format. For example, one may calculate the mean for each of the individuals and create numerical indicator that represents an estimate of average consumption for the individuals in our sample. This also can be done for geographical references reducing the dimensionality of the data and allowing for greater generalisation, but at the expense of precision (either in time, or at to an individual consumer). As was seen from Chapter 3, descriptive measures such as average energy use or total use per day tell us little about the dynamics of energy consumption, thus serving as poor measure to group patterns together. It was also shown that arriving from the same spatially referenced group does also not necessarily imply that consumption patterns can be meaningfully grouped together. According to Liao (2005), the decision on which clustering method to use for time series further depends on the type of the series. The characteristics can include: discrete vs real valued, uniformity of the sample, univariate vs multivariate series as well as lengths of time series considered for the analysis.

Most clustering algorithms consider maximising dissimilarity among the group as the objective using various distance measures (k-means, hierarchical). Alternatives, may consider data generation process replication, i.e. Gaussian Mixture Models or Bayesian clustering by dynamics. If the data is highly variable, caution need to be taken about which one to choose. A further issue, is how to restrict the algorithm to select only similar groups and leave the remaining data-points as outliers, rather than trying to assign every user/data-point to a group.

While clustering is widely used by researchers, it is important acknowledge the difficulty evaluating clustering results, both in terms of quality as well as computational expense. The latter is especially important when scaling up is considered for the analysis to large datasets, not to mention when adding additional features for already highly dimensional data structures. Besides, if individuals are subject

of grouping, additional data that characterise this individual may help to validate the clustering results and evaluate the magnitude of error caused by generalisations. One should note, that in practice this may not always be possible, and as in the case of this thesis additional information only constitutes some broad geographical knowledge of where the smart meter user lives. In the next sections we present one of the most popular clustering methods seen in the literature, k-means, and discuss this method's suitability with respect to smart meter data characteristics. Some limitations of the k-means methods will eventually lead to the choice of Gaussian Mixture Models, which will be given the central attention in the chapter.

4.4.1 K-means

K-means is one of the simplest and fastest methods to minimise the similarity among the objects within each class centres. Often used in consumer research for problems such as classification of consumer baskets Jain (2010) and for various geo-demographic classification, yet mostly on static features, i.e. data from Census 2011. This method is chosen for description here as without prior knowledge of clustering methods it is the one which is chosen often by interdisciplinary researchers and industry practitioners.

The steps of algorithm are as follows:

K-means steps:

1. Choose the centre to start with and a number of centres (k) to consider. (i.e. $k = 5$).

2. For each of the centres assign the data points that are closest compared to the rest of the data.

3. Call the set of point N. Revise and update the clusters by using the mean of the points that were assigned to each of the centres.

The way the closest points are chosen in k-means is usually via the Eucledian distance, and is one of the most common distance metrics. It can be represented with the equation below where x_i and x_j are vectors of dimension P. The squared

difference among data points is taken to calculate likeliness of them belonging to the same cluster.

$$d_e = \sqrt{\sum_{k=1}^{P} (x_{ik} - x_{jk})^2} \tag{4.1}$$

For the survey of some other distances that can be used please see Liao (2005). For time series similarity measures, one among many quite useful way to look at similarity of the patterns could be by treating time series process as piecewise linear function (Möller-Levet et al., 2003) in Liao (2005). The similarity may be taken by taking the squared differences of the slopes for each of the time-series. The problem in application to the energy data is an evident non-linear behaviour of energy use patterns and how these can be incorporated into this standardisation.

One of the additional problems with k-means is the randomisation of the centres at each iteration that can lead to slightly different results each time algorithm will be run. Furthermore, k-means is unable to deal with outliers, an issue which may limit its usefulness for smart-meter data.

While this is certainly one of the most popular methods used in social science data application. One of the limitations associated with the use of k-means is dimensionality and this will become more evidence once smart meter data is considered. Generally, issues of dimensionality may occur when when one is dealing with too many variables that correspond to each of the individual units of the analysis. In our case, each smart meter user has at least 48 features, corresponding to half hourly readings. Adding more features in, may slow down the computation. Some other issues are associated with the Euclidean distance that forms the similarity measure. Temporal profiles imply the occurrence of numerical peaks in the data that can be of different sizes. These are also may be correlated. Each individual fundamentally would have low and high periods of consumption. The similarity measure thus needs to be able to handle such variation, and be able to group both high and low dynamics simultaneously as this is an essential feature of energy dynamics (for more extensive review of similarity measures please see in Liao (2005))

As an alternative to k-means Gaussian Mixture Models are proposed. Given the considerations for the data which were outlined in the first half of the chapter, the next subsection aims to convince the reader that this method may be an appropriate for smart meter data segmentation K means algorithm was applied to the same datasets that are presented in this chapter. In the case of the aggregated sample, only one cluster was observed, meaning that algorithm may have failed to differentiate between the patterns using half hourly energy use as an independent features.

4.4.2 Gaussian Mixture Models

Gaussian Mixture Models constitute a probabilistic framework in which to perform clustering. Unlike k-means, this method also has added consideration of the data generation process that underpins the data, and uses a model for this process to perform segmentation. It is based on a probabilistic method for clustering that handles diverse types of data, including dealing with missing data and hierarchical structures. The probabilities for each data point to be in a particular cluster are first assigned and then a cluster is allocated to each point using those probabilistic measures. The mixture is formed using the probabilities obtained from the standard Gaussian representation:

$$p(x|\mu,\Sigma) = \frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp\left\{-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right\},$$

with μ representing the mean vector and Σ being a covariance matrix. A mixture of Gaussians is then represented as the following:

$$P(x) = \sum_{i=1}^{H} P(x|\mu_i,\Sigma_i)P(x \in \text{cluster } i)\ .$$

As an example, Figure 4.7 demonstrates how consumption variability can be represented as a mixture of densities. As can be seen, these data can be described with a mixture of Gaussian shaped distribution, although they may differ in size or shape. The GMM algorithm is implemented in R in 'mclust' package (Scrucca et al., 2016). For the mixture models, a likelihood based estimation procedure is

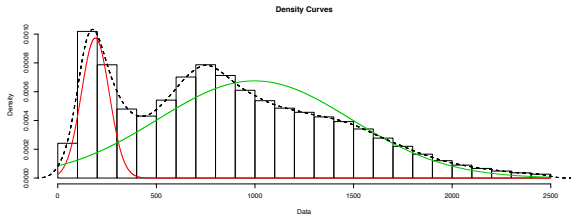


Figure 4.7: *An example of how the energy consumption density can be represented with the mixture of Gaussian distributions*

utilised.

4.5 Experimental Data and Results

A summary of the data sub samples that used in this section is given in Table 6.1. Two sub-samples are taken. One of these is at an aggregated level which include all the customer data, yet reduces the magnitude of the whole sample by taking the average consumption per half hour across individuals in each geographical reference, postcode sector, thus creating the aggregation based on both spatial and temporal averaging. The disaggregated sample corresponds to a random selection of patterns at individual level which were not compressed at any level and can be regarded as raw data. Intuitively, aggregated data may be expected to be more appropriate as it represents a general picture of energy use. However, what will be shown from the clustering experiments, segmenting such data may not always tell as much about diversity of energy use in the country that can be captured if individual readings were taken for analysis. Through suppression of diversity of consumption, customers may look more alike in cases where actually they are not being similar to each other. The results of GMM clustering analysis applied on the experimental samples are presented in Figure 6.6 and Table 4.2.

As may be observed from Figure 4.8 and Table 4.2, while the GMM algorithm is dealing with different samples in terms of size and diversity, interestingly, the same number of clustered groups are obtained. However, the key differentiator between the two cluster models is the shape of the Gaussian models used to fit the patterns. While the aggregated sample presents smoother shapes, more variation can be seen in the disaggregated case, meaning that there is more dissimilarities

Data	Overall	Aggregated Sample	Disaggregated Sample
Unique identifiers	489,000	8,171	15100
Days	365	365	365
Daily readings)	48	48	48
Total observations	8,567,280,000	143,155,920	19,272,000

Table 4.1: Data structure. *The structure of the samples. Please note that aggregated sample is obtained by taking the average consumption among individual users at each of geographical reference, postcode sector level*

Segment	% of total sample (Aggregated patterns)	% of total sample (Disaggregated patterns)
1	24.0%	15.7%
2	10.6%	14.2%
3	5.3%	1.4%
4	0.9%	5.9%
5	1.9%	20.0%
6	21.9%	3.4%
7	15.5%	13.6%
8	14.4%	22.5%
9	5.5%	3.4%

Table 4.2: Results of consumption pattern segmentation using GMM.

in the data. This is intuitive, as after taking the averages of consumption unique variation will inevitably become hidden.

Figures 4.9 and 4.10 present the shapes and variation within the resulting clusters using the GMM model. Note, this differs from Figure 6.6 which projects the clusters and points onto the top two principle components. This visualisation provides a useful way of assessing and attempting to interpret the clusters. As can be seen the number of clusters was identical in both aggregated and raw data, however, the shape of aggregated clusters is far more smoother in the aggregated when compared to those of the disaggregated sample. This fundamental difference may have had a direct implication for the predictability of aggregated clusters as the differentiation on the aggregated level may be more challenging as essential dynamics that distinguish the patterns were collapsed during the averaging of energy consumption.

In terms of spatial allocation to each of the clusters, there is an unbalanced allocation. This is caused by the fact that on average, as was seen in Figure 4.4, energy customers may be alike in their temporal behaviour, particularly characterised by morning and evening peaks. In the case of clustering, the less represented groups

Figure 4.8: *Resulting clusters in high dimensional space*

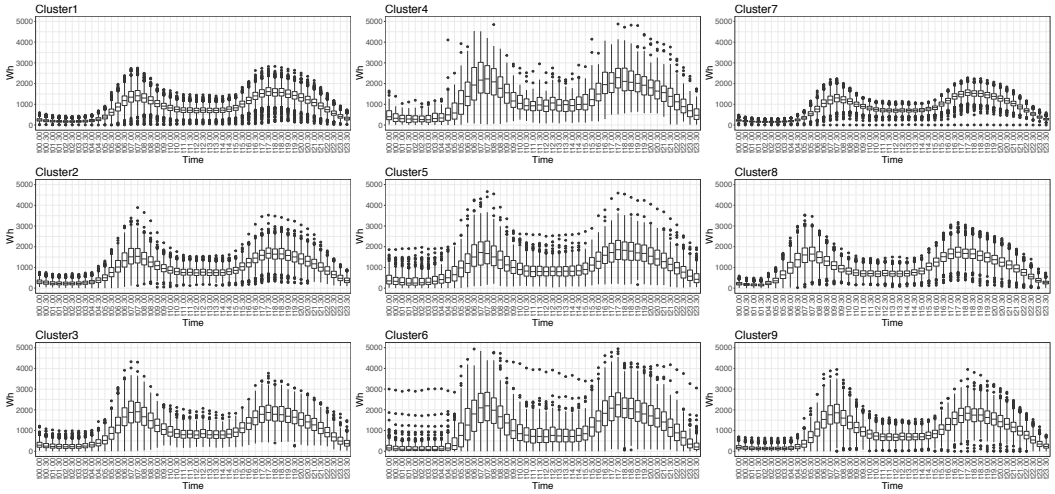


Figure 4.9: *Clusters observed on aggregated sample*

of patterns are indeed those with lower expected energy consumption, profiles that vary from very low to very high, persistent usage during the day.

To conclude this section, it is useful to recall the clustering results observed in Yao and Steemers (2005) which were seen earlier in Chapter 2. As may be noted, some patterns obtained seem to share similarities with those obtained by other researchers. On average, the energy consumption can be distinguished using

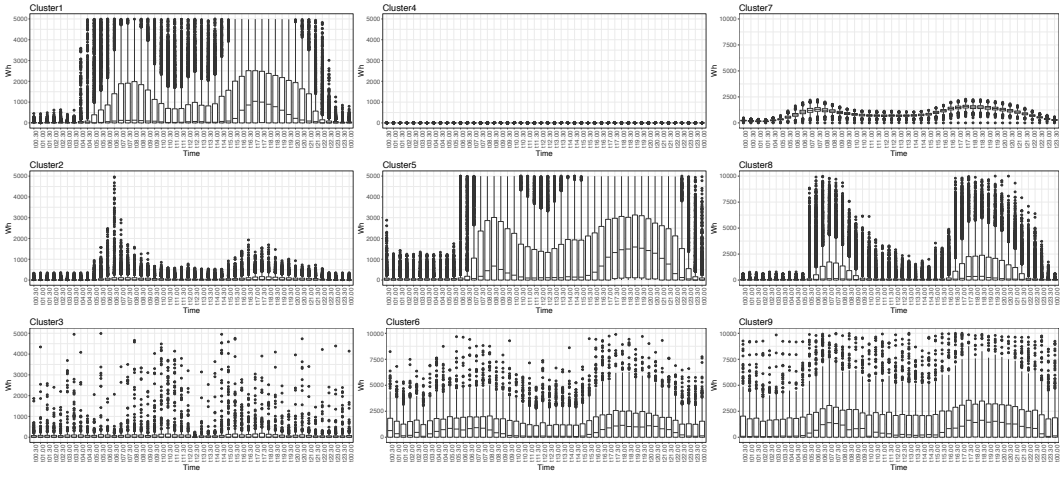


Figure 4.10: *Clusters observed on disaggregated sample*

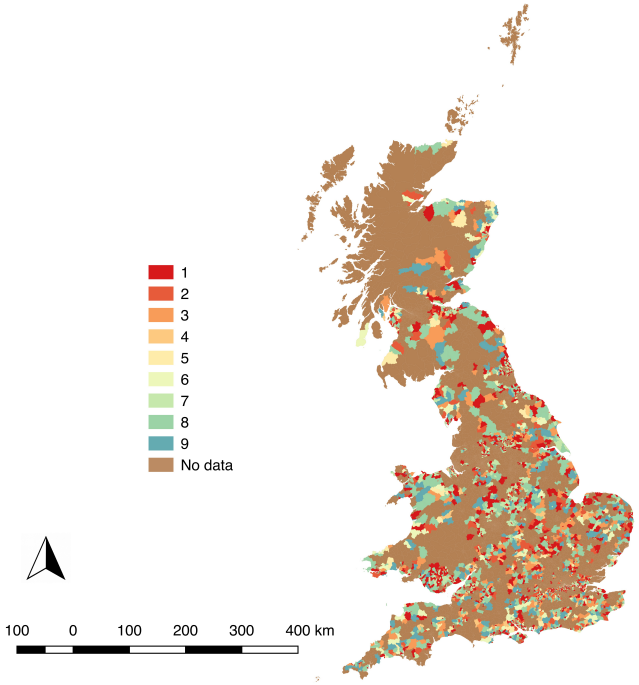


Figure 4.11: *Spatial distribution of the clusters observed using the aggregated sample*

behaviour in and out of peak hours. Further interpretation of the clusters essentially requires further information about the consumers, i.e. demographic data. However, in our case this was not available. Nevertheless, what can still be tested, is how well the clustering method applied in this section can help in allocation of new smart meter readings to clusters. This step can be crucial to conclude on the reliability of the clustering results seen above. The next section considers this in more details.

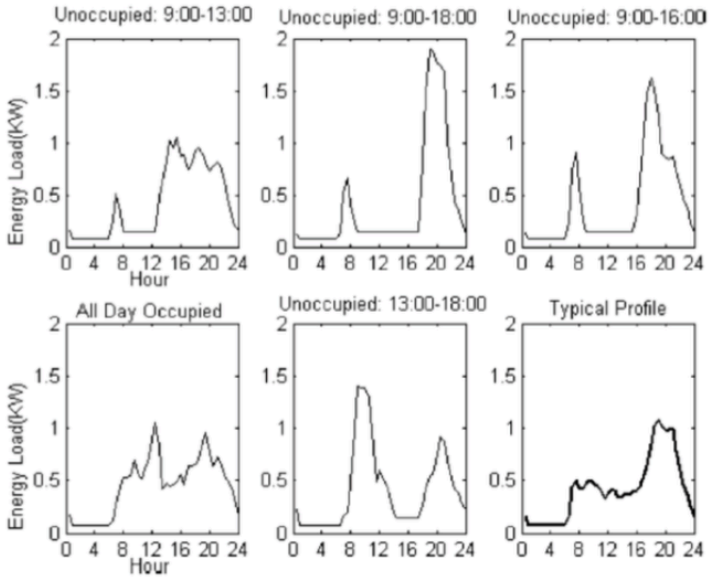


Figure 4.12: *Energy load profiles of a UK average households*
Source: Yao and Steemers (2005)

.

4.6 Predictability of the Clustering Results

To study how well the clusters are allocated to smart meter readings, the predictability of cluster allocation is now assessed under the case where new patterns are introduced into the study. Various tree methods were chosen to segment new/unseen data into clusters that were obtained using GMM models. This was performed for both aggregated and disaggregated samples. What will be observed immediately from the results is that performance on aggregated sample is poorer compared to that disaggregated sample. This is driven by the fact that through aggregation important unique dynamics of customer behaviours that are crucial for segmentation into the distinct clusters may be lost.

In this case study, labels from GMM segmentation of the data are predicted for test sets of data, comparing performance on aggregated and disaggregated samples. Three algorithms are assessed: K-Nearest Neighbours, Random Forest, and Gradient Boosting Trees. K-Nearest Neighbours (KNN) is one of the simplest clas-

sification methods for both binary and multi-class problems. It is particularly useful for problems where the conditional distribution of the outcome variable on the independent variables is unknown. Random Forest (RF) and Gradient Boosting Trees (GBM), are based on decision tree mechanisms. They are differentiated by the approach used to select the best combination of trees and how samples of data are incorporated in the learning process. These methods are especially valuable due to their simplicity in interpretation compared to other machine learning algorithms. They can easily be used for regression and classification type problems and, additionally, to model non-linear relationships. The choice of models here is driven by their popularity in past research, specifically in the multi-class setting. It that has been shown that random forest for instance tends to work particularly well with smart meter data in prediction analysis (Weiss et al., 2012). The model simplifies the analysis and does not require narrowing of the sample through exclusion of certain days as it automatically incorporates patterns in the data corresponding to distinctive individuals. Previously, researchers tended to omit weekends or holiday periods from the analysis as they are often associated with greater heterogeneity among customers (Flath et al., 2012).

4.6.1 K-Nearest Neighbour

K-Nearest Neighbour (KNN) is considered one of the simplest classification methods for both binary and multi-class problems. It is particularly useful for problems where the conditional distribution of the outcome variable on the independent variables is unknown (James et al., 2013). KNN works by taking an input point, x, and K points that are in some sense close to it, i.e. in its neighbourhood. The points nearby in the feature space can then be used to select an appropriate label. The estimator can be written mathematically as

$$\widehat{Y}(x) = \frac{1}{K} \sum_{x_i \in N_K(x)} y_i \, ,$$

where y_i represents the labels of the points in the neighbourhood $N_K(x)$ of input point x.

4.6.2 Tree-based methods

The other methods assessed here include Random Forest (RF) and Gradient Boosting Trees (GBM), are based solely on decision tree mechanisms. They are differentiated by the approach used to select the best combination of trees and how samples of data are incorporated in the learning process. These methods are especially valuable due to their simplicity in interpretation compared to other machine learning algorithms. They can easily be used for regression and classification type problems and, additionally, to model non-linear relationships. For a complete introduction to RF and GBM please see Friedman et al. (2001b).

The Random Forest algorithm is based on building decision tress on bootstrapped (randomly sub-sampled) data with a smaller subset of randomly sampled predictors at each decision node. A large number of trees is grown until a stopping rule is achieved (e.g. minimum 5 observations in the terminal nodes) and then aggregated for final prediction. An example of the successful use of Random Forest in civil war onset prediction can be found in Muchlinski et al. (2015) and Strobl et al. (2008).

Our implementation of the model is as follows. The input variables are represented by the sequence $\{b...B\}$ which is a combination of half-hourly readings. The model draws bootstrap samples, Z, from the training set, and random forest trees are built using a combination of predictors that are responsible for the split of these trees. Once a number of tree classifiers have been generated, the average is taken among all and form a single classifier. Output is represented by $\{T_b\}_1^B$ The class is then predicted for the unseen data (test set) through the majority vote that selects the best performing trees:

$$\widehat{C_{\mathrm{RF}}^B}(x) = \text{majority vote} \left\{ \widehat{C_b(x)} \right\}_{b=1}^{B} ,$$

where $\widehat{C_b(x)}$ is the classification given by a single tree.

An alternative tree algorithm known as Gradient Boosting was first used to tackle classification problems, however, is now widely used for regression as well (Friedman et al., 2001b). Like Random Forest, the gradient boosting algorithm

takes advantage of both weak and strong classifiers. The reference "weak" refers to the fact that on average classifiers may bring a prediction which is slightly better or just the same as a random guess. Unlike Random Forest where at each iteration a new solution is being trained to then find the average best among many, in the Gradient Boosting model the solution of the already trained model is updated as more samples are taken. The trees are therefore updated at each iteration to obtain more powerful classifiers.

In boosting models, one first assigns weights $w_i = \frac{1}{N}$ to each of our training observations that include both input and output variables, with N being the total number of observations (Friedman et al., 2001b). The process is then iterated F times during which the classifier $G_f(x)$ is fit using the observation weights. The observations which were misclassified at the previous stage are assigned greater weights, so at each iteration more importance is given to the observations that were harder to initially classify. The error associated with each model fit calculated as:

$$e_f = \frac{\sum_{i \varepsilon N_i} w_i I(y_i \neq G_f(x_i))}{\sum_{i \varepsilon N_i} w_i},$$

where $I(x) = 1$ if $x = 0$ and $I(x) = 0$ otherwise, and is known as the indicator function.

Those with the highest error are assigned an increase to their weights using the factor of $\exp\gamma_f$, where γ_f . The final output $G(x)$ is based on continuous iterations of model fit using re-weighted observations until the error rate of penalised observations is minimised

4.7 Testing the segmentation mechanism

Table 6.2 reports overall accuracy and kappa values for each of the models used to predict the data segment. Entries for 'Accuracy' report the overall prediction power of the model including both true positives and true negatives over total of true and false positives and negatives. The Kappa statistic is used for the evaluation of classifiers by comparing the observed accuracy of prediction with that of a random chance. The optimal parameters were obtained using ten-fold cross-validation that

allows for using all the dataset in the training process. The data is being split into train and test set ten times and each chunk is used for training and testing of the model.

Model	Accuracy	Kappa
Aggregated sample		
K-Nearest Neighbor	23%	0.14
Gradient Boosting Trees	37%	0.29
Random Forest	40%	0.29
Disaggregated sample		
K-Nearest Neighbor	65%	0.58
Gradient Boosting Trees	80%	0.73
Random Forest	79%	0.75

Table 4.3: *Results of multi-class prediction.*

KNN		1	2	3	4	5	6	7	8	9
	1	**13.79%**	3.42%	1.67%	0.00%	0.00%	0.00%	26.85%	5.08%	2.04%
	2	15.17%	**23.29%**	0.00%	0.00%	0.00%	0.00%	9.34%	11.02%	0.00%
	3	11.03%	21.23%	**6.67%**	9.52%	3.45%	10.53%	6.23%	13.56%	10.20%
	4	4.83%	13.01%	40.00%	**47.62%**	34.48%	21.05%	0.00%	7.63%	16.33%
	5	11.03%	14.38%	11.67%	23.81%	**41.38%**	21.05%	5.06%	8.47%	10.20%
	6	6.21%	7.53%	26.67%	19.05%	20.69%	**47.37%**	1.56%	11.02%	42.86%
	7	8.28%	1.37%	0.00%	0.00%	0.00%	0.00%	**30.35%**	0.85%	0.00%
	8	16.55%	5.48%	1.67%	0.00%	0.00%	0.00%	14.79%	**17.80%**	2.04%
	9	13.10%	10.27%	11.67%	0.00%	0.00%	0.00%	5.84%	24.58%	**16.33%**
GBM		1	2	3	4	5	6	7	8	9
	1	**27.03%**	11.54%	1.52%	1.39%	1.32%	1.19%	24.79%	15.38%	2.15%
	2	**12.61%**	**36.54%**	21.21%	4.17%	5.26%	0.00%	5.98%	8.65%	4.30%
	3	**9.01%**	21.15%	**30.30%**	6.94%	10.53%	3.57%	3.42%	9.62%	11.83%
	4	**0.90%**	2.88%	12.12%	**52.78%**	22.37%	19.05%	0.00%	2.88%	7.53%
	5	**1.80%**	10.58%	1.52%	16.67%	**44.74%**	13.10%	2.56%	5.77%	3.23%
	6	**0.00%**	0.00%	13.64%	18.06%	10.53%	**41.67%**	0.00%	2.88%	26.88%
	7	**23.42%**	1.92%	0.00%	0.00%	0.00%	0.00%	**48.72%**	5.77%	2.15%
	8	**18.92%**	6.73%	6.06%	0.00%	0.00%	2.38%	13.68%	**29.81%**	12.90%
	9	**6.31%**	8.65%	13.64%	0.00%	5.26%	19.05%	0.85%	19.23%	**29.03%**
RF		1	2	3	4	5	6	7	8	9
	1	**26.13%**	10.53%	3.17%	0.00%	0.00%	1.33%	27.73%	13.33%	4.12%
	2	11.71%	**43.42%**	22.22%	1.41%	0.06%	0.00%	9.24%	9.17%	5.15%
	3	9.91%	26.32%	**22.22%**	8.45%	12.64%	0.00%	0.84%	12.50%	15.46%
	4	0.90%	2.63%	14.29%	**46.48%**	27.59%	17.33%	0.00%	1.67%	9.28%
	5	4.50%	13.16%	14.29%	21.13%	**44.83%**	0.07%	0.84%	4.17%	4.12%
	6	0.00%	5.26%	7.94%	19.72%	9.20%	**50.67%**	0.00%	5.83%	17.53%
	7	21.62%	1.32%	1.59%	0.00%	0.00%	0.00%	**49.58%**	5.83%	1.03%
	8	18.02%	7.89%	3.17%	0.00%	0.00%	2.67%	10.92%	**30.83%**	13.40%
	9	7.21%	13.16%	11.11%	0.00%	2.30%	21.33%	0.84%	16.67%	**29.90%**

Figure 4.13: *Confusion matrix reporting the correspondence between observed vs predicted class*

One of the immediate observations is the difference in performance when considering aggregated versus disaggregated analysis. Aggregated models are associated with higher misclassification rates, suggesting that by aggregating essential dynamics that contribute to identifiable patterns indeed have been lost.

KNN	1	2	3	4	5	6	7	8	9
1	**73.00%**	2.00%	15.00%	0.00%	3.00%	1.00%	6.00%	10.00%	0.00%
2	0.00%	**67.00%**	0.00%	3.00%	0.00%	0.00%	1.00%	0.00%	0.00%
3	0.00%	2.00%	**60.00%**	0.00%	0.00%	0.00%	2.00%	0.00%	2.00%
4	0.00%	0.00%	0.00%	**97.00%**	0.00%	0.00%	0.00%	0.00%	0.00%
5	20.00%	0.00%	42.00%	0.00%	**87.00%**	4.00%	4.00%	15.00%	1.00%
6	2.00%	0.00%	13.00%	0.00%	2.00%	**64.00%**	2.00%	2.00%	8.00%
7	0.00%	18.00%	17.00%	0.00%	0.00%	0.00%	**67.00%**	1.00%	0.00%
8	1.00%	10.00%	29.00%	0.00%	1.00%	0.00%	19.00%	**70.00%**	1.00%
9	3.00%	0.00%	29.00%	0.00%	6.00%	31.00%	0.00%	2.00%	**88.00%**

GBM	1	2	3	4	5	6	7	8	9
1	**88.49%**	1.23%	2.24%	1.22%	4.36%	0.00%	3.43%	4.14%	0.24%
2	0.59%	**84.06%**	0.00%	6.97%	0.04%	0.00%	6.63%	1.56%	0.00%
3	0.32%	1.28%	**81.34%**	1.51%	0.21%	0.00%	1.11%	0.28%	0.96%
4	0.00%	0.09%	0.00%	**83.52%**	0.00%	0.00%	0.00%	0.00%	0.00%
5	6.82%	0.32%	5.22%	0.75%	**90.38%**	4.00%	2.18%	5.28%	0.00%
6	1.26%	0.00%	5.97%	0.00%	0.88%	**92.22%**	0.05%	0.51%	3.37%
7	0.09%	8.70%	0.75%	4.05%	0.18%	0.00%	**78.35%**	3.15%	0.00%
8	1.22%	4.33%	2.24%	1.98%	2.14%	0.22%	8.25%	**84.82%**	0.00%
9	1.22%	0.00%	2.24%	0.00%	1.83%	3.56%	0.00%	0.26%	**95.42%**

RF	1	2	3	4	5	6	7	8	9
1	**82.73%**	0.96%	9.62%	0.00%	8.27%	0.00%	4.29%	5.00%	0.35%
2	0.60%	**84.50%**	0.00%	2.64%	0.00%	0.00%	6.92%	2.24%	0.00%
3	0.60%	1.33%	**75.00%**	0.00%	0.31%	0.27%	2.54%	1.34%	0.70%
4	0.00%	0.23%	0.00%	**97.14%**	0.00%	0.00%	0.00%	0.00%	0.00%
5	8.01%	0.23%	3.85%	0.00%	**83.67%**	6.04%	3.66%	7.67%	1.40%
6	2.26%	0.23%	5.77%	0.00%	2.27%	**70.60%**	0.79%	1.22%	23.16%
7	0.64%	9.77%	1.92%	0.22%	0.21%	0.00%	**67.97%**	2.76%	0.00%
8	3.22%	2.75%	3.85%	0.00%	0.31%	0.27%	13.63%	**79.23%**	0.70%
9	1.93%	0.00%	0.00%	0.00%	4.96%	22.80%	0.20%	0.54%	**73.68%**

Figure 4.14: *Confusion matrix reporting the correspondence between observed vs predicted class*

More detailed information can be observed from the confusion tables that highlight differential performance of the prediction methods across clusters. While RF and GBM tend to perform better on average, KNN showed higher accuracy in some classes. This is possibly related to different 'bias-variance' trade off for each of the tree models. While boosting aims to reduce the bias by taking the average of predictive performance among the estimated models, Random Forest fundamentally searches for a solution that reduces variance by imposing a strict structure of reducing the number of predictors at each split of the tree.

As observed from the confusion tables (Figures 4.13 and 4.14), the prediction methods show differential performance across clusters. One of the immediate observations is the difference in performance when considering aggregated versus disaggregated analysis (Table 6.2). Aggregated models are associated with higher misclassification rates, suggesting that by aggregating, information on essential dynamics that contribute to identifiable patterns are lost.

While RF and GBM tend to perform better on average, KNN showed higher accuracy in some classes. This is possibly related to different 'bias-variance' trade off for each of the tree models. While boosting aims to reduce the bias by taking

the average of predictive performance among the estimated models, Random Forest fundamentally searches for a solution that reduces variance by imposing a strict structure of reducing the number of predictors at each split of the tree.

Often, the classes that are better represented in the data may be associated with better performance as there is more data available for the training. In our case, this had no implication on performance. Classes with smaller number of observations were more easily differentiated, while the bigger ones showed higher levels of misclassification. Later in the thesis, it will be shown that such trade off persistent also in applications that consider regression analysis of smart meter data.

4.8 Some further extensions

This section presents possible extensions to the clustering analysis above. The extensions are aimed at narrowing down the analysis both spatially and temporally. This is based on the assumption that by increasing resolution of space and time one may be more confident about the relationships in the data. The clustering within the sample of Bristol energy consumption is presented. It follows by an attempt to use clustering on a chunk of time (out of peak hours). The results in this section are inconclusive, yet are presented to give the reader a flavour of the kind of approaches that may be taken to answer more specific questions about where, when and how smart meter users consume energy.

4.8.1 Narrowing the space

The aggregated data for Bristol was created in similar to national sample faction. Average half hourly measures across the whole year were taken at OA level. The temporal profiles were then clustered. The only difference of this sample is much finer OA level geography compared to that of postcode sector. The resulting clusters are presented in Table 4.4 and Figure 4.15.

The number of distinct clusters is smaller than that of the national dataset. Nevertheless, some immediate correspondence with the clusters that were defined previously can be noted for clusters 1 and 2. The consumption in Bristol is observed to be differentiated by both peak hours and throughout the day patterns. Most of

Segment	% of total sample (Aggregated patterns)
1	27%
2	38%
3	35%

Table 4.4: *Results of consumption pattern segmentation at OA level in Bristol using GMM.*

the output area aggregates are associated with very low or no consumption during the night time. This may suggest that variability of energy consumption at a finer geographical level, i.e. over the city of Bristol, could be representative of wider UK energy use. Further to this, it may help us in filling the gaps where data is missing by defining some common energy behavioural patterns that are more frequent in each of the areas in Great Britain, or as one may call it, typical profiles.

4.8.2 Narrowing the Time Resolution

A further extension to the temporal analysis of energy data may involve segmentation of the patterns in terms of peak hours as they have shown to be important for the definition of the clusters above. Only one day was selected for the analysis, an average weekday in the end of January. Figure 4.9 suggests that regardless of segmentation, there are quite similar patterns around morning and evening peak hours which vary in magnitude but are evident for each of the clusters. Examining peak and outside peak hours separately may tell us slightly more on households presence at home as well as particular habits or routine (i.e. waking up early for work, late nighters). It can be shown that for defining the interactions between characteristics of people living in the area and energy, concentrating on specific time and location may reveal more information about energy consumption rather than when both time and space are aggregated. As may be observed in clusters 3 and 4 in Figure 4.16, consumption happening throughout the day is more frequent around the coastal regions and less occurring in central England. Overall, while the results seem fuzzy at this stage due to the lack of additional data, the approach itself can be taken to study a very specific and narrow questions such as “where people are likely to be at home during the day?”.

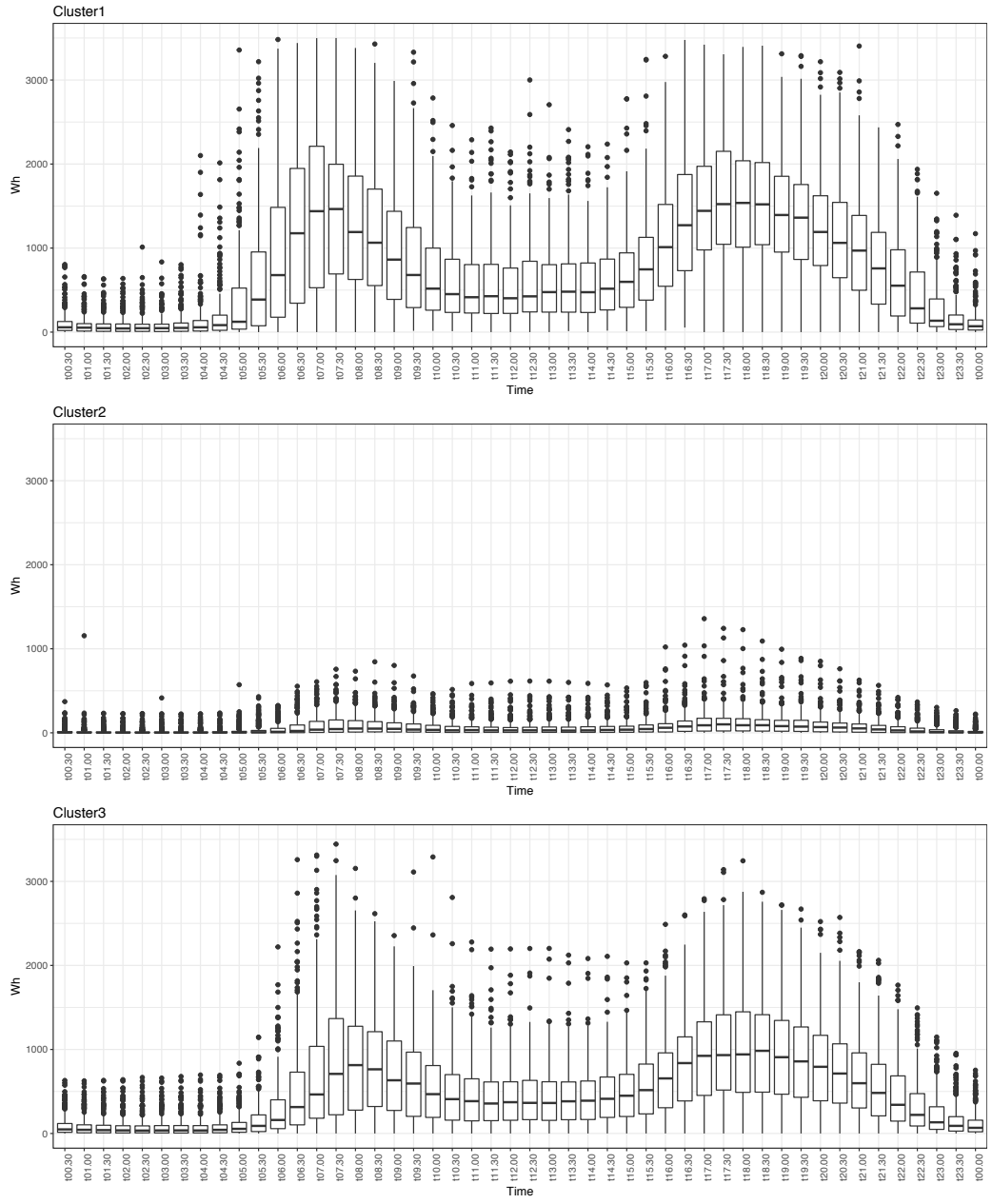


Figure 4.15: *Clusters derived from annual aggregates at OA level for Bristol.*

4.9 Conclusions

This chapter has provided an introduction to the smart meter data as time series process. The concepts of stationarity and uncertainty were discussed in the context of smart meter data. Some unique characteristics of temporal and spatial dimensions of the data were presented to undermine the complexity of the data and motivate the choice of methodologies to be used for data segmentation. Clustering models

Figure 4.16: *Clustering of off-peak hours data. The peak hours consumption levels are characterised by the the times between 11.30am and 3pm*

that are available to group the energy consumption patterns together were briefly introduced. It was demonstrated that the data can be meaningfully clustered using Gaussian Mixture Models. The chapter suggested a possible strategy for prediction and characterisation of temporal profiles. It was shown that both segmentation and predicability of segments groups tend to work differently depending on whether the aggregated or disaggregate samples are under study. For prediction in particular, it was shown that using aggregated data records leads to much higher rates of misclassification, while the most granular data can be classified and predicted with more certainty. Compared to Random Forest, in practice some classifications may be better performed using Gradient Boosting trees. However, the performance may be at the cost of over-fitting the data (Friedman et al., 2001b) Nevertheless, what is observed is rather a mixture of performances with each method winning or losing for different prediction class. This may be related to the essential 'bias-variance' trade-off that is handled differently by each model. While boosting aims to reduce the

bias by taking the average of predictive performance among the estimated models, Random Forest fundamentally searches for the solution that reduces the variance by imposing a strict structure of reducing the number of predictors at each split of the tree.

On the interpretative side, an approach to narrow down either the spatial or temporal dimensions was presented briefly. Given the limitations of the datasets accessed in this thesis it is challenging to move towards any inference why profiles of energy consumption may differ. Looking at various regions one by one or at distinct chunk of times may be handy if the researcher has a specific question in mind. For instance, in the Section 8 of this chapter it was shown that by using clustering at various temporal and spatial scales one can access:

- 'How representative the variation in energy consumption in Bristol of the wider population of the UK?'
- 'Where people are likely to be at home during the day?"

The next chapter will look in more details at various approaches to characterise smart meter data, this time in a regression context. The aim, is to predict the patterns of energy consumption and understand what influences the variability of energy usage across time, one can learn more precisely about the process that generates the energy consumption pattern. Some further applications arising from this chapter could be an attempt to build the hierarchical structure of energy consumption based on seasons, month, days of the weeks. It may also be constructed with the addition of a spatial dimension for data where greater spatial granularity is available to study; this would allow us to assess if there are any spatial pre-determinants of the variation in the energy use. To improve even further, a possible suggestion of using the co-integrated time series where both gas and electricity are combined may perhaps presents a more complete picture of the segmented energy use for the UK using this or similar data.

Chapter 5

Methodology and Results: Regression Analysis of Time Series

5.1 Introduction

The following two chapters are concerned with the tasks of forecasting and prediction with smart-meter data. Once the data is meaningfully grouped, the next step would be to attempt to predict energy use of either an individual user or to look at the energy use of customers as a group that share similar patterns.

There are two types of predictions that can be considered: point or sequence prediction, and classification. Point prediction refers to the scenario where one may want to predict either next half hourly energy consumption or the whole sequence of consumption, say for the next 24 hours. There are number of approaches that are developed for point prediction. Primarily, they assume stationary or weakly stationary processes embedded in the time series. As discussed in the previous chapter, the assumption of stationarity may not be appropriate and the researcher should also consider alternative methods that may relax the assumption of stationarity.

The importance of prediction is mainly centred on the opportunities provided via consumption feedback, which can be offered to both the energy supplier and the smart meter user. For example, this could be a personalised saving suggestion at smart meter user level based on past use. Alternatively, it can be more inclusive and aggregated information such as helping to understand the overall pressure

on the grid by looking at aggregated consumption at different time periods. One could use this to anticipate if any interventions are required given predicted future usage. Lastly, a detailed analysis of predictability of certain patterns may tell us about overall periodicity in the consumption behaviour. In other words, if the customer consumption can be easily predicted using the model one may conclude the customer is to some level being periodic in their behaviour so it is easier for model to learn what they will be doing next.

Classification refers to the scenario where one is interested in using smart meter data to predict a specific class/category that characterises specific energy patterns. The previous chapter demonstrated how using classification may assist one in studying how stable clustering allocations are performed for grouping smart meter data patterns when new data is subsequently introduced. In the next chapter, we will investigate a further type of classification problem, this time where the label of smart meter data has a meaning. For instance, smart meter data can be used to classify different type of customers (i.e. family vs single occupant) or classification of property types (i.e. terraced house vs detached). For this thesis, a very specific example of the label is chosen: energy vulnerability of the customer based on affordability of the energy bills. Such an application is highly topical for both applied and methodological reasons. On the application/industrial side, the UK energy market and policies require companies to attempt to identify the fuel poor (Rosenow et al., 2013). From a technical and academic point of view, it is interesting to discuss such classification, as interpretation of the label is usually subjective and qualitative rather then quantitative.

Application of predictive analysis for both forecasting and classification are yet quite limited within smart-meter industry research. For instance, a number of software companies such as IBM are aiming to offer tools to analyse smart meter data, however, if there is no data infrastructure put in place such that real time analysis is possible, these solutions may remain infeasible for quite some time. The experiments presented in this thesis may also suffer from the the same issue, as while feasible on the cases of smaller samples, they certainly may fail if data of

larger magnitudes are considered. A number of additional preliminary steps for data analysis such as clustering prior to forecasting and feature transformations prior to prediction may need to be thought of. Some possible directions for such feature extraction are discussed in Chapter 7.

The energy research community focuses largely on addressing the above implementation barriers by targeting the complete roll out of smart meters by 2020, whilst also exerting a sustained drive to develop the tools and techniques that will be in place once data will start arriving on automated basis (Wang et al., 2018; OFGEM, 2015). To give an example, one application of predictions for smart meter data may include the design of a feedback loop for customers, which allows them to see the predicted costs of the energy usage if they change their behavioural patterns. Past research have shown that usage of home in displays by customers to analyse their own energy use have a potential for immediate savings and behaviour change (Faruqui et al., 2010b). However, the study of how periodic or habitual the individuals are in their energy consumption still remains a relatively untapped subject due to unavailability of long term consumption smart meter data. The uniqueness and individuality in energy consumption patterns makes it challenging to design a standard and generalised approach to study their patterns.

5.1.1 Structure overview

One of the first contributions of this chapter is to highlight the challenges and opportunities that are associated with the variability and heterogeneity of energy consumption patterns when sample is pooled from various regions with various type of users that are unidentified. This case is characterised by the situation where data cannot be reduced based on some meaningful selection (i.e. selection of specific income group customers or certain property types). A key question which is asked in this chapter is whether the available methods that can be used to predict next day consumption are capable of handling such diverse and large datasets. Census 2011 OAC classification is used to complement the previously introduced Bristol sample which is then used to predict consumption trends in winter and summer.

Some preliminaries on time series forecasting, and the nature of linear models

are introduced before moving on to generalised linear models and smoothing. The central methodology of this chapter, Generalised Additive Models (GAM) with application to smart meter data will follow in Section 5.3. Section 5.4 will present some justification for the samples chosen for the analysis, and brief descriptive statistics for this sample will be provided. Results of using GAM to fit energy consumption for various types of smart meter users will then be presented and analysed. Some issues and problems with the analysis will also be outlined, opening up more opportunities for further research. The conclusion and discussion of potential limitations and further ideas arising from this work will round the chapter.

5.2 Preliminaries

Previous work in point prediction for energy data has largely been constrained to modelling annualised energy use Huebner et al. (2015) instead of the real consumption due to unavailability of smart meter data. With growing access to smart meter data, one may think that now the energy consumption can be easily predicted with almost no uncertainty. What was observed in a previous chapter however is that smart meter data and the dynamical structure of such data are highly complex and may require an advanced methodology to be used. This section will review some important fundamentals that need to be considered to describe the smart meter data when one considers the point or the patterns prediction.

In the previous chapter, the issues of both temporal and spatial heterogeneity were touched upon briefly. This section will focus on how these heterogeneous behaviours can be controlled for in the predictive analysis. In particular, the main explanatory cause for heterogeneity appears to be from effects of seasonality that is present both within the days, weeks, month and quarters of the year.

This chapter will be based on relatively small samples of data, a random selection of individual patterns will be used. The aim is essentially to experiment with various patterns of energy consumption to study if they can be described using regression analysis and consequently be predicted. Once it can be shown that prediction strategy is satisfactory on a small sample, larger selection of patterns may

be considered. In a broader sense, this may be relevant for the predictive tasks associated with studying the cases where the researcher may want to investigate an overall pressure on the energy network and national grid. However, as is demonstrated in this work, it can be challenging to take larger samples of smart meter data for regression analysis, as more data means more variability and complexity.

The next section will present an overview of the baseline structure of linear models. These models underpin the main method used in the chapter, Generalised Additive Models and is necessary to be mentioned here to provide a guiding intuition behind the outcome and the predictors relationship. In this set up the outcome variable would be the next half hour energy use and the predictors will be represented by the series of past energy use.

Another reason of revising a linear model (as with k-means in clustering application), this is the method which will be often initially considered by a researcher or an industry practitioner due to its popularity and simplicity of application. To move a step further and introduce a discussion for extending linear model to GAMs, both nature of the approach and nature of the data that is suitable for these models are discussed.

5.2.1 Linear model

Despite the growing availability of statistical tools available to researchers, the most common and reliable type of statistical analysis remain to be a technique that is based on a linear relationship between predictors and outcome variables. These can be used to estimate the dependence of $E(y)$ on predictors vector X where the model is linear and is of the fixed parametric form. Such parametric structure implies a very definite form of relationship (i.e. linear) which limits flexibility of the regression line in expense of generalisable and concise summary of the relationship between predictors and outcome variables.

This section revises the assumptions and the structure of simple linear model that can be used describe the relationship between the outcome and predictors variables. The intuition behind linear model sets an essential base for Generalised Additive Models (GAM) that will be applied to smart meter data in order to forecast

next day energy consumption pattern. The most basic and intuitive way to approach time series data prediction is by using a 'simple' Ordinary Least Squared (OLS) estimator to identify parameters.

OLS approaches for time series forecasting rely heavily on the assumption of independence which was discussed briefly in a previous chapter. The linear model is canonically specified by

$$E(Y|X) = \alpha + \beta \mathbf{X}\,, \tag{5.1}$$

In the energy consumption case $E(Y|X)$, is the expectation of consumption in the next half hour as a function of past consumption observed, can be expressed as the function of explanatory variables X. In a linear model, least-squared estimation is used to obtain α and β such that residual sum of the squared differences between the fitted by the model line and observed data are minimised. From the definition, it is appropriate to use linear model where relationships between explanatory variables and the outcome tend to be linear. However, non-linear cases may be still studied using the linear model construction by inclusion of quadratic terms, i.e. X^2.

Some of the important assumption that linear models rely on are those of independence, normality, homoscedasticity, and linearity. Behaviour of the error term is one of the most important components in the regression analysis, mainly its statistical distribution. Error, also know as the residual term, itself does not mean a literal 'error'. Rather, it is a measure of the variation in the data that is random and cannot be explained by the suggested relationship between the variables. Ideally, the error term is expected to be normally distributed and has no correlation with independent variables. If this holds, one may conclude that bias is minimised and the model has picked up all the relevant variation that can be explained by the model.

Linear regression is a powerful tool that can be used on its own or serve as a base for more advanced techniques. Some of the main purposes of linear regression were outlined in Rencher and Schaalje (2008). These are prediction, data description and explanation, parameters estimation, variable selection and control of output for experimentation. Where relationship may not be perfectly linear, various mod-

ifications to the linear model can be made either via transformation of one of the predictors (i.e. using polynomial) or by fitting the smoothing function that can be then used as an independent variable. The example is shown in the next section.

5.2.2 Smoothing

Smoothing is the method that similar to regression analysis that may help in describing series which have no obvious trend. A smoothing function (smoother) can be described as an approach that is used to summarise the trend in some random variable (y), in this case next hour energy consumption readings. Smoothers belong to the family of non-parametric models and are often used to either describe the data generation process or for estimation of how the mean of of the outcome variable (y) is dependent on the predictors (x). The latter is related to the type of the smoother that will be used in the structure of Generalised Additive Models. It is also widely used in GIS applications when one is interested in the intensity of the mean dependancy based on the various geographical units. Smoothing represents an essential component of non-parametric regression. While regression can be of parametric nature (i.e. linear), its components may be non parametric (i.e. smoothers).

There is a huge variety of smoothing functions ranging from very basic such as natural splines to more diverse, such as families of gaussian kernel smoothers. They mostly differ in the approaches to how the data is being averaged across neighbourhoods of X, these neighbourhoods can be described using smoothing parameters.

In this thesis, the survey of smoothing functions will be limited. As will be seen, different customers may have very unique periodicity and cyclic behaviour. Hence, there may be no general smoother that can be applied across all smart meter users equally. Choice of smoothing function may also largely depend on the frequencies of erratic or irregular consumption patterns. An example of kernel smoothing applied to the annual aggregation of daily consumption patterns is presented in Figure 5.1. A Gaussian kernel smoother is used in this scenario. As was seen from the previous chapter, a Gaussian density based approach to smart meter data may be a wise initial choice.

From the first glance at Gaussian kernel smoothing of the average annual daily

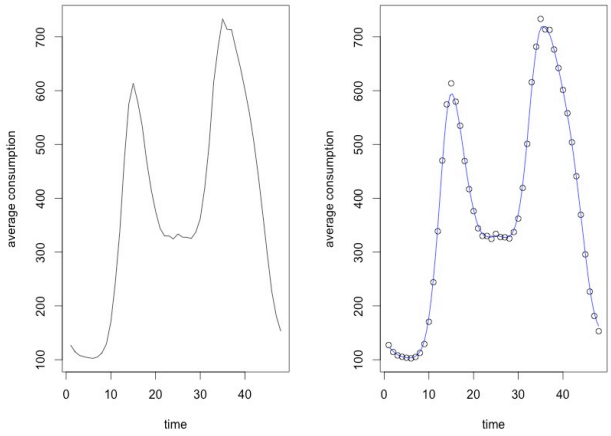


Figure 5.1: Gaussian kernel smoothing of the average annual daily pattern

pattern in Figure 5.1, one may notice that the smoother fits directly to the points of energy use, meaning that there is little or no space for generalisability left if we were to consider using the same function to a series of different patterns. A recall of bias-variance trade off is necessary here to inform the reader that depending on bias -variance trade off one may be looking at, we may choose very different complexity of smoothing parameters when setting up GAMs. The section below provides an overview why this is an important consideration when one attempts to describe energy use.

5.2.3 The Bias-Variance Trade-Off

The concept of Occam Razor introduced in the previous chapter needs recalling in this section as we are approaching the discussion on model selection. The bias-variance trade off is a trade off between the complexity of the model and variability in accuracy achieve by that model be it prediction or classification tasks. In the context of linear regression, the most obvious bias variance trade off can be described using concepts of overfitting and under-fitting. Models that have less bias are often associated with higher complexity in their parameterisation, thus making them less likely to generalise to new data, and to overfit the data. Overfitting may also happen due to noisy data points being parametrised. Thee points could be represented by rate unusual deviation in consumption which not necessarily indicates

significant shifts in user behaviour . Where a model is simple enough the bias may be much higher, however, there will be less chance for such input variables noise to be taken as important for describing the relationships between input and output variables, making the model more generalisable to other unseen data. Ideally, one would attempt to balance both bias and variance. It will be shown that thanks to smoothing component in Generalised Additive Models (GAM) presented below, one may reduce chance of overfitting or under-fitting by appropriately regulating the smoothing parameters of the model.

5.2.4 Metrics for Models Comparison

Given the discussion above, intuitively one may want to measure how much variance increase over bias decrease is gained by attempting different models structures. There are various measures in place for that. The main metrics that will be used in this study are primarily, R squared and adjusted R squared for accessing the explanatory power of the models and secondarily, Akaike Information Criterion (AIC) and Bayesian and Information Criterion (BIC) to access the complexity of the given models.

R squared is used to measure the goodness of fit of the models that can be found by calculating the ratio of the variation in the data that can be explained by the suggested model (i.e. linear model with multiple predictors). As more predictors are included in the model, R squared is expected to rise, this does not always mean however that the model is better as there may be an issue of overfitting as well as this measure will be sensitive to the total number of data points in the dataset that is used for model fitting. To correct for this, adjusted R squared is often preferred as it would adjust to higher number of predictors with respect to total data points on which model being trained.

AIC and BIC are metrics that often being used to measure complexity of the model that is based on the analysis of bias-variance trade off in the model as well as number of predictors that are used in the model. In other words, the metric help find a model that fit data well but also can be generalisable enough. Both are based on the likelihood of the estimated model to predict true values. Lower AIC and

BIC are indication of better fit (for more details on metrics for model comparison and mathematical intuition behind them please see (Judd et al., 2011; Burnham and Anderson, 2004).

5.3 Generalised Additive Models (GAM)

The techniques and analytical strategy applied in this chapter are based in parts on work by Wood (2006, 2004) and Laurinec and Lucká (2016). More specifically,Wood (2006) has developed an application of GAM to EDF energy load data, while Laurinec and Lucká (2016); Laurinec (2016) has considered data on commercial buildings electricity consumption in the United States. The work in this chapter extends their analysis by applying the model in the context of the residential customers where periodicity of consumption may be more variable. The aim is to test whether these models are suitable for big data analysis with residential gas and electricity consumption.

The Generalised Additive Models (GAM) allow for the model to be based not just on the sum of individual variables (i.e. linear regression) but the sum of smoothing functions of those explanatory variables. This implies that non linear effects can be studied as well as effect of individual smoothing functions can be examined separately (similar to the assessment of significance of explanatory variables in the linear regression model we can access the explanatory power of the functions). One of the highlights of GAMs is that we can account for seasonality within the day, week, month or even a year and incorporate those seasonalities into our prediction by allowing the model to be built using the functions of daily, weekly, monthly dependencies. The GAM belongs to the family of the non parametric model types which makes it slightly more flexible than ordinary Generalised Linear Models (GLM).

Formally, the model can be written as follows:

$$g(y_i) = \beta_0 + f_1(x_i) + ...f_k(x_i) + \varepsilon_i \tag{5.2}$$

where y_i belongs to the exponential family distribution (to see more about

exponential family distribution please look at Chatfield et al. (2010)), $i = 1,...N$ references the data-point, y is the outcome variable (next half hour prediction), and $x_1,...x_k$ are independent variables (past energy consumption). The unknown smoothing functions are represented by $f_1,...f_k$. Furthermore, we can write these functions as combinations of basis functions b_{ij} according to

$$f_i(x) = \sum_{j=1}^{q} b_{ij}(x)\beta_{ij} \, . \tag{5.3}$$

This parameterisation allows us to linearise the representation, basically, all we need to do is find the appropriate parameters of β_{ij}. This function can also be referred to as a spline with a basis function b. The splines or smoothing bases can be variable (i.e. cubic, cyclic-cubic) depending on being most suitable for the regression fit. We can now represent the overall model as a linear combination of these functions. To do this, we can consider making a big matrix $\mathbf{X} = (b_{ij}(x), \ldots, b_{kq}(x))$, the expected value is now given by:

$$g(E(y)) = \beta \mathbf{X} \, , \tag{5.4}$$

where β is now a vector of size kq (if q is constant for all $f_1, \ldots, f_k$). An estimator can be formed in a similar manner to OLS by minimisation of the following function:

$$\|y - \beta \mathbf{X}\|_2^2 + \lambda \sum_{i=1}^{k} \int_0^1 \left[f_i''(x))\right]^2 dx \, . \tag{5.5}$$

Lambda (λ) is a smoothing parameter, larger values encourage smoother functions f, the notation $f''(x)$ is used to denote the second derivative of the functions. The integral of second derivative squares can be presented as:

$$\int_0^1 \left[f_i''(x))\right]^2 dx = \beta^T \mathbf{S}_i \beta \tag{5.6}$$

The coefficients can be estimated using below:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda \sum_{i=1}^{k} \mathbf{S}_i)^{-1}\mathbf{X}^T y \tag{5.7}$$

$\hat{\beta}$ is an estimator of the regression coefficients. It can be estimated using a penalised approach called Penalised Iteratively Re-weighted leas Squares (P-IRLS) (Wood, 2006).

5.3.1 The Back-fitting algorithm and cross validation

Some of the most common issues with GAMs, due to the nature of smoothers chosen, is imprecision in model fit as well as challenges with model selection process as the number of smoothing functions that can be incorporated can vary.

Hastie and Tibshirani (1990) developed one of the early model implementations availiable as a package for the R software. In their package 'gam' to fit the GAMs to the data they used back fitting algorithm developed by Friedman (1991); Breiman (1993). In a nutshell, their algorithm attempts to solve a system of equations by attempting finding a solution to the set of functions till the model converges to a unique solution. Yet, in general, there is no strong condition on availability of the unique solution to given problem as it is highly conditional on selected smoother and initial condition of the algorithm (i.e. functional form used at the start of iterative process) so every time the back fitting algorithm is run it may provide a variety of results. This implies that there is no guarantee of a unique solution to the estimation procedure. An alternative to consider is the cross validation. Cross validation splits the given sample in the training and test data chunks that are being rotated till the whole sample has been used. The average model fit can then be taken considering the results on each data split.

New developments allowed for cross validation to be incorporated instead which makes fitting easier and perhaps, more robust. Simon Wood from Bristol developed a new package 'mgcv' in 2006 that is now used more commonly for GAM applications (Wood, 2001).

Cross validation is the process which used for most of the machine learning methods that were applied to smart meter data in this thesis. It is one of the best

ways to address the issue of bias variance trade off and ensure that the model we fit on training data can be as as generalised as possible to the new and unseen data.

5.4 Fitting GAMs: Data and Results

To study how predictable customer behaviour is in terms of gas and electricity consumption a number of cases were selected for testing the GAM approach. Using Census 2011 Output Area Classification (OAC), four random customers were extracted from the dataset that correspond to Bristol sample which is available at greater geographic resolution. These are two customers (one for gas and one for electricity) from the area that is characterised as Rural Residents by OAC and two customers from the area that is characterised as Urban Professionals and Families.

As will be seen throughout, the results relating to gas consumption remain more variable and perhaps more challenging to study. The next chapter considers this in more detail and is based solely on gas readings. All these samples were picked under the assumption that periodicity of behaviour may be slightly different for those who are based in urban area and may be employed full time compared to those who can be described as rural residents. Ageing was further chosen as a covariate to assess wether an ageing population is more like to have more periodic and continuous consumption compared to urban professional and whether ageing category itself can be further differentiated by urban and rural area of residence's characteristics (the results for these customers can be found in Appendix 8.6).

Before proceeding to the results, it is of further importance to note that there may be an uncertainty associated with the fact the OA classification is based on average demographics in a region. However, the individual consumer we pick from this area may not exhibit all of these demographic characteristics. In other words, these classifications often tend to be characterised by the characteristics of dominant socio demographic group. There is always a chance, even if small, that the customer selected from for instance urban professional, may be characterised by many other characteristics which have nothing to do with urban setting or professional life.

Nevertheless, as will be seen from the results there are some intuitive corre-

spondences between area characteristics and the consumption patterns. Previously in the thesis, it was attempted to to use the Census Output Area Classification to explain some variability of the total per day consumption in various areas. While such attempt have shown rather weak relationship, what is observed here is that when the periodicity of behaviour is studied instead, there are much clear correspondences to OA characteristics. This once again confirms the observation that aggregation may hide unique characteristics of smart users behaviour that can be explained by the socio-demographics of the areas they reside. Using gas in these scenarios may undermine the distinctive behaviours across seasons. Electricity will also be explored for the comparison later in the chapter. It will be shown that electricity is not necessarily totally different from gas when it comes to fitting the GAMs. Yet, there may be less differentiation by seasons (this will be shown in Section 5.5).

To compare the overall performance of the models, a measure of explained variation such as R squared will be used to access the quality of the overall model fit. Comparison of fitted and real values are visualised together with the 3D graphic where both daily and weekday trends can be shown simultaneously. Most of experiments consider a further comparison between summer and winter. As will be shown there are quite huge differences in the overall patterns description and their predictability when gas and electricity consumption from the different seasons are considered . To complement the overall results and visualisations, the residuals check that can aid the understanding of model fit and highlight the possible problems that may be taken for further research, especially for big samples, are presented.

Each sample of readings was trained using the variation of GAM models. Starting with a very simple model where the parameters of weekly seasonality are taken independently from daily seasonal behaviour. It then expanded to a model that consider interactions of weekly and daily dynamics, assuming that conditionally on individual weeks, there may be different effects of daily variation on the prediction of next half hour readings. When looking at the error term, it will be shown that there are persistent occurrences of heteroscedasticity which are caused by the

energy consumption being unique or heterogeneous from day to to day. This is not that concerning when looking at small samples, however, the problem becomes more apparent where more months taken into analysis. Overall, it will be shown that energy consumption is hard to be modelled at full in practice. In other words, it is challenging to find a model that can possibly explain every little component of variation in energy use across half hour intervals. However, as broad results (Section 5.5) presented here show, high precision may not be always necessary if one is simply interested in understanding the periodic components of consumption, and their weighting given the total variability of consumption.

One of the important aspects to note, is that compared to linear models, commonly used by social science researchers, it is hard to use hypothesis testing in GAM models to describe the results of the model fit. The coefficients are slightly meaningless (it will be shown that only magnitude of the coefficients can be used for comparison) due to non linearity of the behaviour the model attempt to describe and there are no confidence intervals provided by the model output. All in implies that there is no precise calculation of predicted value as it may be described by different parameters at various parts of the fitted line. As a consequence, regression tables, graphical results such as smoothing splines and residual plots serve more effectively as fit quality check in GAM context.

5.4.1 Data Samples Description

This section presents a brief description of the selected samples of smart meter users that were picked for the analysis using the characteristics of the Output Area they reside in and the seasons. It further can be differentiated by energy source: gas and electricity. The samples that were chosen had more of less full annual coverage. Nevertheless, as will be noted from the tables below, some smart meter recordings are more complete than others. Most of the users had missing data across days which was not systematic. These missing values were removed where consumption readings were coded as NA in a raw data [1]. Threadings where consumption coded

[1]The reason behind recoding missing values to NA instead of replacing the value with the average energy consumption was due to the fact that by replacing value as a mean, the pattern of energy use will be changed. Most of the missing values occurrences were spotted for time between 11.30pm-

as zero were kept as this can be a realistic measure of consumption . On the positive side, the users that were picked up for the analysis had full coverage for specific months that were chosen to describe differences between weeks in winter and summer. As can be observed below, rural residents consumption can be described as slightly smaller in the overall magnitude compared to that of urban professionals.

Tables 2 and 5.2 present the overview of the sampled users. The order of the presentation corresponds to the order of results that are presented in the rest of the section.

	Electricity		
	Mean (half hour)	St Dev (half hour)	N
Rural Resident (Annual)	125.01 Wh	144.42 Wh	15300
Rural Resident (Jan-Mar)	152.92Wh	163.72Wh	3374
Rural Resident (May-Aug)	97.20Wh	122.74Wh	3605
Urban Professional (Annual)	235.84Wh	273.28Wh	15300
Urban Professional (Jan-Mar)	216.515Wh	271.45Wh	3374
Urban Professional (May-Aug)	222.05Wh	231.55Wh	3605

Table 5.1: Descriptive Statistics for Electricity Samples, "Rural/Urban" group

	Gas		
	Mean (half hour)	St Dev (half hour)	N
Rural Resident (Annual)	901.92 wh	1686.54Wh	14833
Rural Resident (Jan-Mar)	1685.70Wh	2241.47Wh	3374
Rural Resident (May-Aug)	239.67Wh	581.65 Wh	3605
Urban Professional (Annual)	490.16 Wh	1444.85 Wh	14449
Urban Professional (Jan-Mar)	801.15Wh	1550.90 Wh	3374
Urban Professional (May-Aug)	29.49 Wh	251.28 Wh	3605

Table 5.2: Descriptive Statistics for Gas Samples, "Rural/Urban" group

5.5 Experimental Results

This section presents the results of experiments performed for this thesis using GAMs. As application of these models in social sciences is not very common, there is no standard way of regression output presentation. Consequently, the results can be assessed from various angles. Regression tables, smoothing splines, residuals

12.30am. On average, these are the time intervals where consumption tend to be lower compared to the day average

checks, and 3D illustration of the fit that incorporates both weekly and daily dimensions of the consumption that were fit by the model. These were selected to give as much detailed picture as possible of model performance on selected subsamples of data. To start with, results using samples of electricity consumption are presented from a sampled smart meter user in the rural area. It is then compared to the area characterised as 'Urban Professionals'. Winter and summer patterns are compared, a large sample over several months of readings is also taken for analysis. This is to demonstrate that extrapolation within one smart meter user may not be as straightforward, meaning that there can be differences in the model performance when applied to small versus large time span.

This section and this chapter as a whole once again highlight how complex the smart meter data are, and how difficult it is to generalise models across the population. The next subsection will present an attempt to fit GAMs to smart meter data. The rest of the chapter will further consider the strategies to report and evaluate the model results more effectively [2].

5.5.1 Electricity

5.5.1.1 Rural

The analysis here considers smart meter users arriving from the Output Area characterised as 'Rural'.

Figure 5.2 presents samples of consumption patterns for winter and summer that are presented with the GAM fit that was achieved using unrestricted GAM model with interaction of 'Daily' and 'Weekly' parameters. The unrestricted model demonstrated best performance in terms of explanatory power measured by R squared (0.6 for winter and 0.2 for summer). More detailed results are presented in the Table 5.3.

Even without considering model fit, one immediately notices how distinct the behaviour of the customer in winter compared to summer. This is evident from

[2]Most of the strategies for analysis and results presentation were borrowed from the tutorial of Peter Laurinec (Laurinec, 2016). He performed a very similar analysis of non residential smart meter data using Wood (2001) package 'mgcv'

both magnitude and consumption periodicity. This is not surprising and as will be seen from the rest of the chapter, regardless of which customer one chooses and no matter whether it is gas or electricity the patterns will differ.

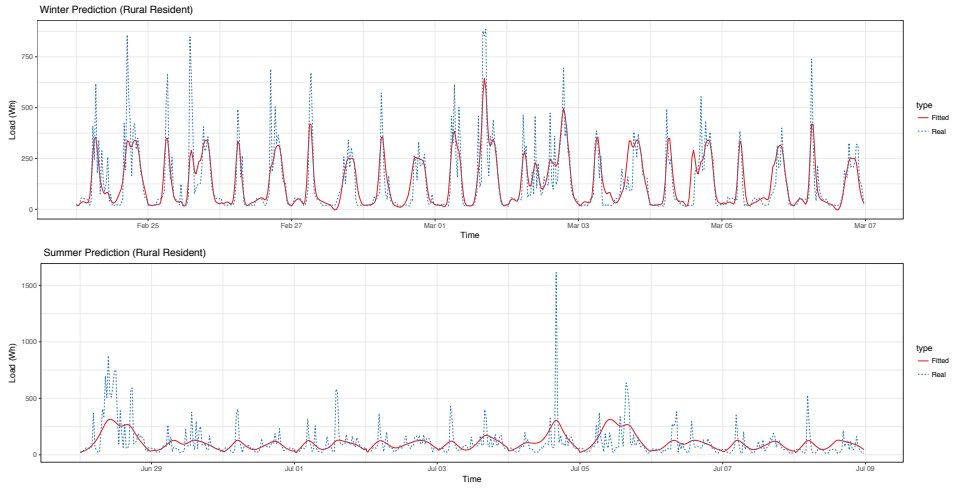


Figure 5.2: GAM fit for a customer that belongs to OA characterised as 'Rural Resident'

Both patterns can further be visualised in the form of a surface (c.f. Fig. 4) that can help visualising both daily and weekly shape of the fit. On the left axis there are 48 half hourly periods and on the right are 7 days a week. Weekends are associated with peaks of consumptions for this customer (both in winter and summer there is a strong afternoon peak on Saturday and midday peak on Sunday)

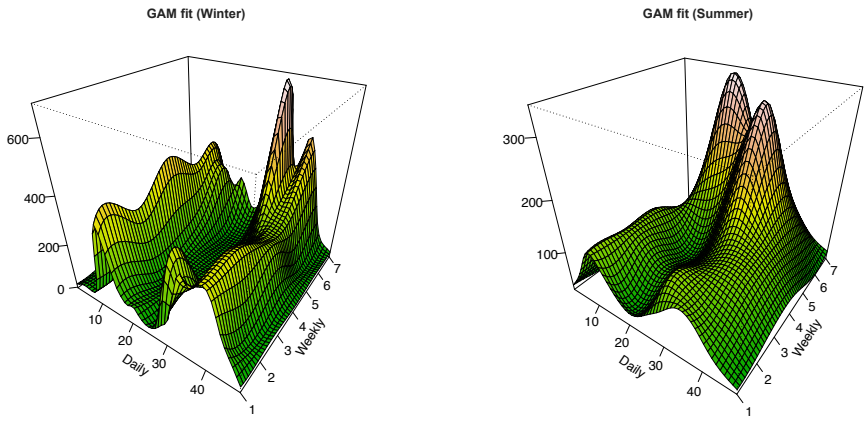


Figure 5.3: GAM fit for a customer that belongs to OA characterised as 'Rural Resident' in 3D.

So far, the results lack a metric for assessing model fit, and what the model

tells us about explanatory power of individual time variables. As a further step, the visualisations of how each of the time periods affects y variable are presented below. There are two graphics that show the relationship between x and y on a daily basis and on a weekly basis. These are further differentiated for two samples (one sample of about 11 days and one sample of about 4 months). This comparison is essential, as smaller samples are much easier to describe using GAM whilst on a larger sample too much variations may make it hard to generalise about effects of individual half hour periods.

While these samples are fairly small, a slightly bigger sample was selected to see what the smoothed behaviour may look like. Smoothing line which is passing through the readings is presented below for both daily and weekly variation which is further differentiated by the size of the sample (Figure 5.4). Both winter sample and six month sample are considered for the comparison and assessment of long term dynamics of this particular user. One can see that while variation across the day can be represented as fairly typical (defined morning and evening peaks), there is an increasing trend for consuming more by the end of the week. Overall, for this specific customer, it may be said that their sampled pattern behaviour appears representative of their behaviours over longer period of time.

In terms of specific regression results, Table 5.3 presents a detailed picture of what has happened behind the scenes of the fitted lines which were seen in the beginning of the section. The table shows results for GAMs fit with and without interaction terms, this is further split into a restricted and unrestricted model. An unrestricted model is represented by a model with unlimited estimated degrees of freedom (EDF). Having restricting on EDF may prevent the issue of overfitting as it will restrict the model of fitting a way too complex smoothing function. Unrestricted EDF thus allows for much greater complexity, making the model fit better but at the expense of overfitting. Results are further differentiated by small[3] and large samples (please see 'Numbers of Observations' row).

[3]Small sample is represented by the sample for winter. This was selected due to that fact that on average summer fit, not only in this particular case, but across the set of experiments was associated with realtively poor model fit

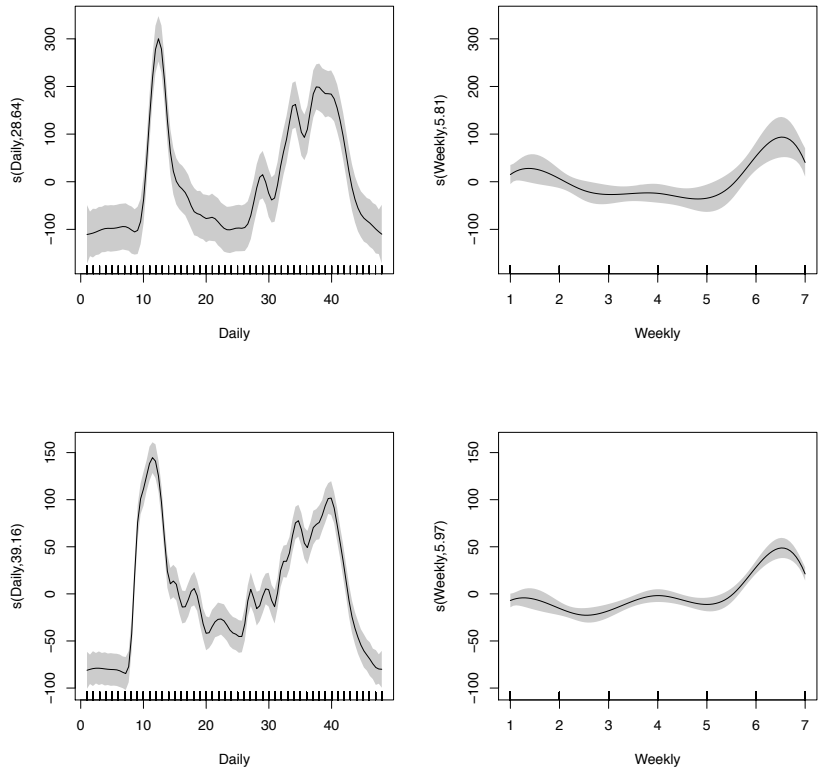


Figure 5.4: The fitted response due to each variable/covariate contribution in winter, 'Rural Resident'. Winter sample (up) and 6 month sample (bottom)

The resulting coefficients represent here the complexity of the smoothing splines. Complexity is directly related to the value of EDF (greater the EDF, more complex is the spline (Laurinec, 2016). The stars indicate how statistically significant is daily or weekly values are in explaining the outcome variable which is the next day consumption. As can be seen, both daily and weekly variation is important for this smart meter user for predicting their future use. However, the interaction of two shows better explanatory power which is measured by the ratio of explained variation in the data over the total variation, R squared. The unrestricted model here demonstrated somewhat better explanatory power (R squared of 0.69) but is associated with greater EDF, which implies greater complexity of the smoothing splines. When using larger sample that is associated with more variation of energy use, the model fit is quite poor. Again, this is not surprising due to many factors that may be affecting the customer across the year apart from time alone, implying that there

is a need for controlling for more variables that may be correlated with changes in energy use.

	GAM	GAM + Inte. (Unrestricted)	GAM + Inter.	GAM	GAM + Inter. (Unrestricted)	GAM + Inter.
(Intercept)	133.70***	133.70***	133.70***	116.93***	116.93***	116.93***
	(4.71)	(4.39)	(4.50)	(1.45)	(1.42)	(1.41)
EDF: s(Daily)	28.64***			39.16***		
	(34.90)			(44.32)		
EDF: s(Weekly)	5.81***			5.97***		
	(5.98)			(6.00)		
EDF: te(Daily,Weekly)		113.26***			127.82***	
		(146.97)			(165.80)	
EDF: t2(Daily,Weekly)			60.42***			99.88***
			(74.30)			(118.62)
AIC	6481.41	6472.15	6455.99	77519.36	77344.41	77250.72
BIC	6637.02	6964.22	6722.48	77837.32	78220.23	77938.03
Deviance explained	0.57	0.69	0.63	0.25	0.29	0.30
R^2	0.54	0.60	0.58	0.25	0.28	0.29
GCV score	12562.35	12979.29	12091.58	13226.81	12868.66	12676.24
Num. obs.	528	528	528	6288	6288	6288
Num. smooth terms	2	1	1	2	1	1

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.3: Regression Output, 'Rural Resident'

Further to the results above, residual checks (Figures 5.16, 5.15) are presented to more explicitly assess the behaviour of the error term. As was mentioned in the previous section, for the model to have valid results, one of the most important assumption to be satisfied as normality and randomness of the error term. As can be seen, when using small samples, this assumption may be satisfied with more ease than when bigger sample is taking into analysis. More variability brought in with more data creating endogeneity problem. On average, for this sample, the distribution of the error term is positively skewed. This doesn't look as problematic for small sample as for the larger one. Overall, more data is required. Not as much data on energy use but data that can serve as potential covariates of energy use. It may be useful to include various socio-demographic or property characteristics in the model to see whether the distribution of the error term may improve. The resid-

uals analysis below suggests that model doesn't not explain well everything that is happening in energy consumption variation, especially for the case of large sample. There is also no evidence for suggesting that unrestricted sample is significantly better than the restricted one.

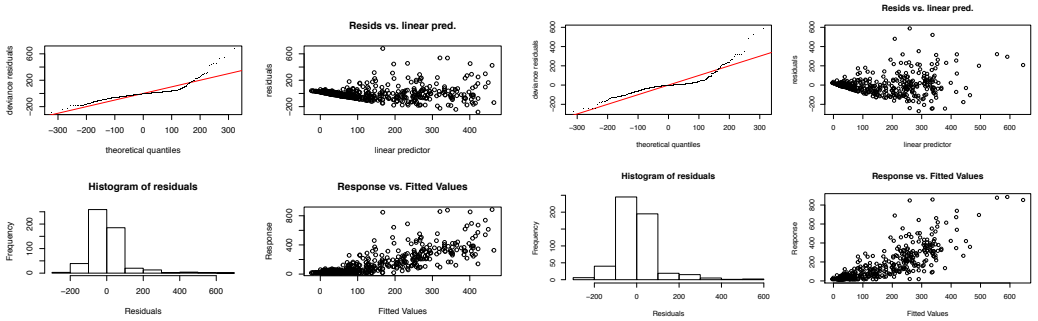


Figure 5.5: Residual Check for Winter Fit, 'Rural Resident'. Restricted (left) and unrestricted model (right).

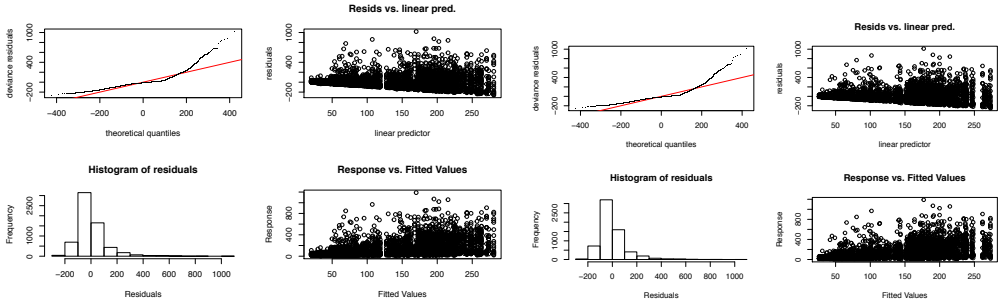


Figure 5.6: Residual Check for 6 months sample, 'Rural Resident'. Restricted (left) and unrestricted model (right).

This section has suggested a strategy on how the results of a GAM fit can be analysed and presented such that one may gauge some insights about what is statistically happening behind smart meter readings. These results do not necessarily suggest that energy consumption analysed with GAMs will always be associated with similar performance of the model, neither does it suggest that these are average consumption characterises of rural resident consuming electricity. The rest of the chapter will present three other examples to highlight the diversity of energy use from user to user and model performance. An urban customer is presented next in a similar fashion, before the chapter proceeds to the analysis of gas consumption.

Gas consumption will be shown to be more variable compared to that of electricity, something that was already seen earlier in the thesis when smart meter readings were looked at from angle of clustering analysis.

5.5.1.2 Urban

It is expected that the customer arriving from the area that is characterised by 'Urban' may have slightly different consumption behaviour compared to 'Rural Residents'. This section looks at 'Urban Professionals' area. While there is certainly seems to be more randomness in winter consumption behaviour , summer behaviour have occurrence of absence of consumption and may looks rather regular (See Figure 5.7).

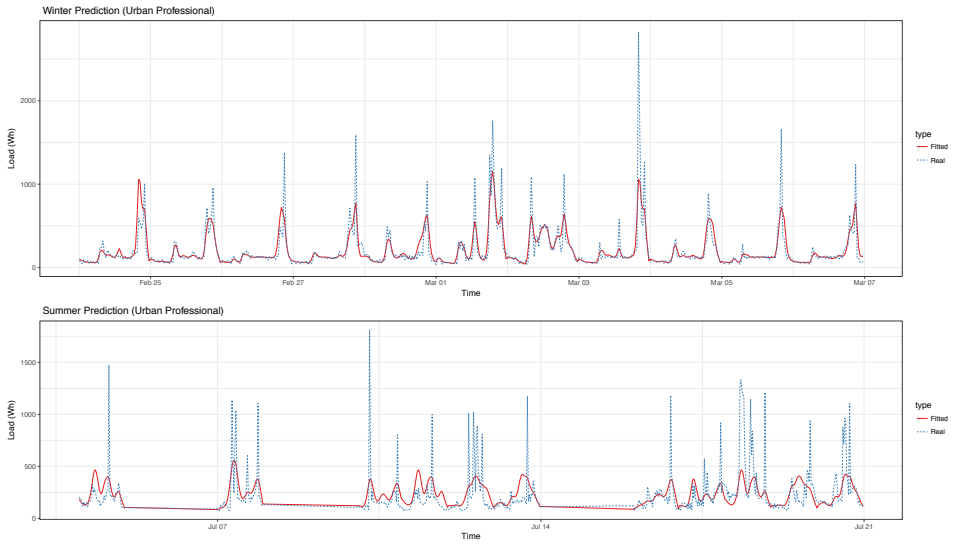


Figure 5.7: GAM fit for a customer that belongs to OA characterised as 'Urban Professional'

From more detailed 3D visualisations, it can be gauged that week in winter for the customers looks way more periodic than that of summer.

When looking at smoothing splines, it can be seen that extrapolated to longer period of time the customer barely has an association with typical energy use but may have more variation across different weeks. Smoothing spline visualisation is helpful here to see that when taking an average, the behaviour has lower variance and has more distinct patterns.

Perhaps, it may be more interesting to assess how well GAMs can fit such

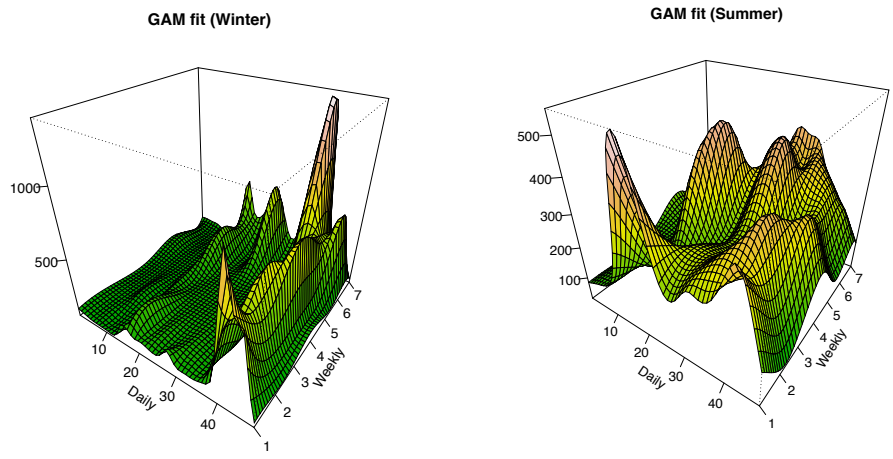


Figure 5.8: GAM fit for a customer that belongs to OA characterised as 'Urban Professional' in 3D.

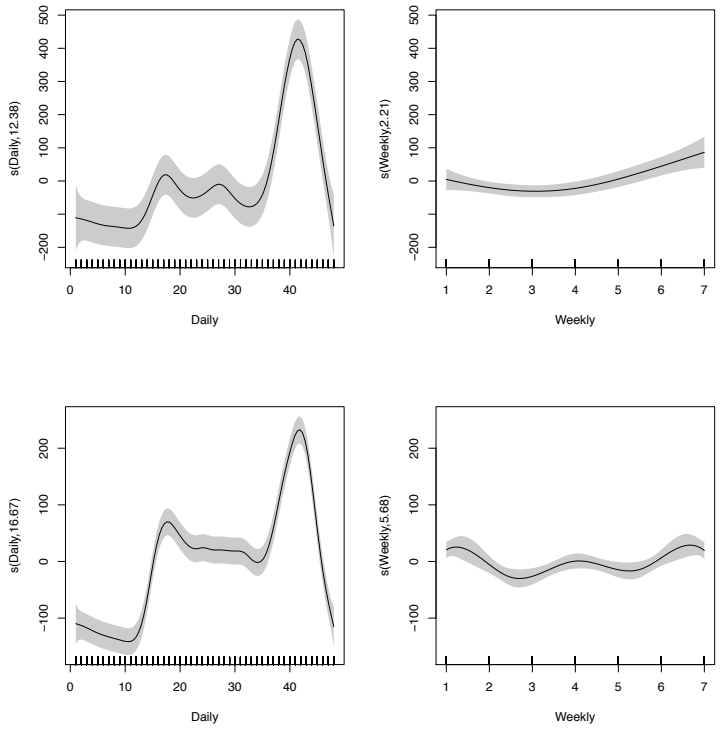


Figure 5.9: The fitted response due to each variable/covariate contribution in winter, 'Urban Professional' .Winter sample (up) and 6 month sample (bottom)

behaviour. Table 5.4 presents the overall results. As can be seen, compared to the previous case, unrestricted model with interactions terms for small sample tend to perform best (R squared of 0.59) in terms of variance explained by the model.

However, the EDF components are slightly different compared to rural customers. Perhaps, due to more irratic behaviours across the reading, it is harder for model to fit a simple spline. EDF is quite high for models with interaction, regardless of wether they are restricted or unrestricted.

As before, large sample only confuses the model. Performance measured by model fit is relatively poor, perhaps due to more diversity of energy use in the long run which can only be explained with an addition of other covariates such as weather or identification of smart meter user characteristics.

	GAM	GAM + Inte. (Unrestricted)	GAM + Inter.	GAM	GAM + Inter. (Unrestricted)	GAM + Inter.
(Intercept)	205.96***	205.96***	205.96***	220.45***	220.45***	220.45***
	(8.87)	(7.61)	(7.26)	(3.01)	(2.89)	(2.89)
EDF: s(Daily)	12.38***			16.67***		
	(15.45)			(20.77)		
EDF: s(Weekly)	2.21***			5.68***		
	(2.64)			(5.94)		
EDF: te(Daily,Weekly)		131.24***			109.43***	
		(168.96)			(140.58)	
EDF: t2(Daily,Weekly)			146.05***			102.52***
			(172.26)			(125.51)
AIC	7131.22	7066.44	7025.79	86743.49	86283.42	86292.43
BIC	7202.05	7635.28	7657.81	86907.73	87035.15	86997.59
Deviance explained	0.41	0.66	0.70	0.16	0.24	0.24
Dispersion	41578.87	30611.35	27838.74	57135.69	52381.96	52513.65
R^2	0.39	0.55	0.59	0.16	0.23	0.23
GCV score	42844.02	40840.36	38584.32	57348.59	53318.32	53392.69
Num. obs.	528	528	528	6288	6288	6288
Num. smooth terms	2	1	1	2	1	1

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.4: Regression output, 'Urban Professional'

Residuals checks (Fig. 4) complement the above results to suggest that once again while small samples are associated with well behaved GAM in terms of satisfying the normality and randomness assumption of the error terms, there are issues for bigger samples.

Regardless of the fact that these samples may be characterised by different

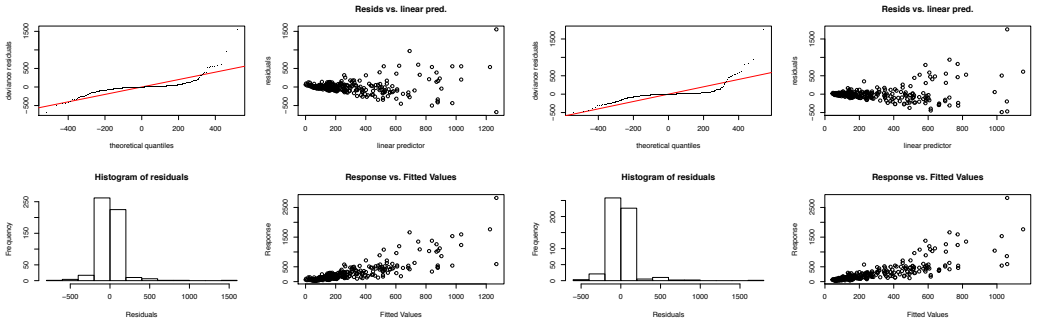


Figure 5.10: Residual Check for Winter Fit,'Urban Professional' . Restricted (left) and unrestricted model (right).

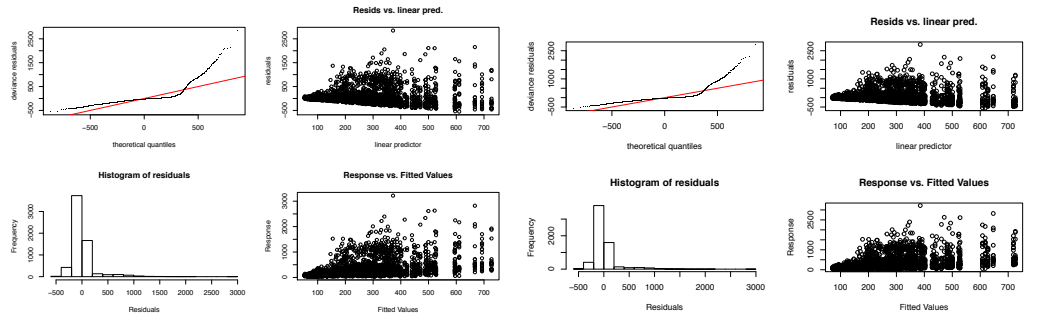


Figure 5.11: Residual Check for 6 months sample, 'Urban Professional'. Restricted (left) and unrestricted model (right).

socio demographics, there appear to be no differences in model fit and explanatory power when it comes to explaining variability of energy use using the parameters based on time.

5.5.2 Gas

This section presents the results for gas consumption for two customers arriving similarly from rural and urban areas. As will be shown, gas can be considered as slightly more variable in the nature of energy use. It is also more periodic compared to that of electricity.

5.5.2.1 Rural Residents

The first example considers the resident of the area that is characterised as rural by OAC. Figure 5.12 report the fitted and real values of energy consumption first for the sample of winter days and then for the summer. Selection of days is chosen such there are no absences by the smart meter user (i.e. no days without consumption).

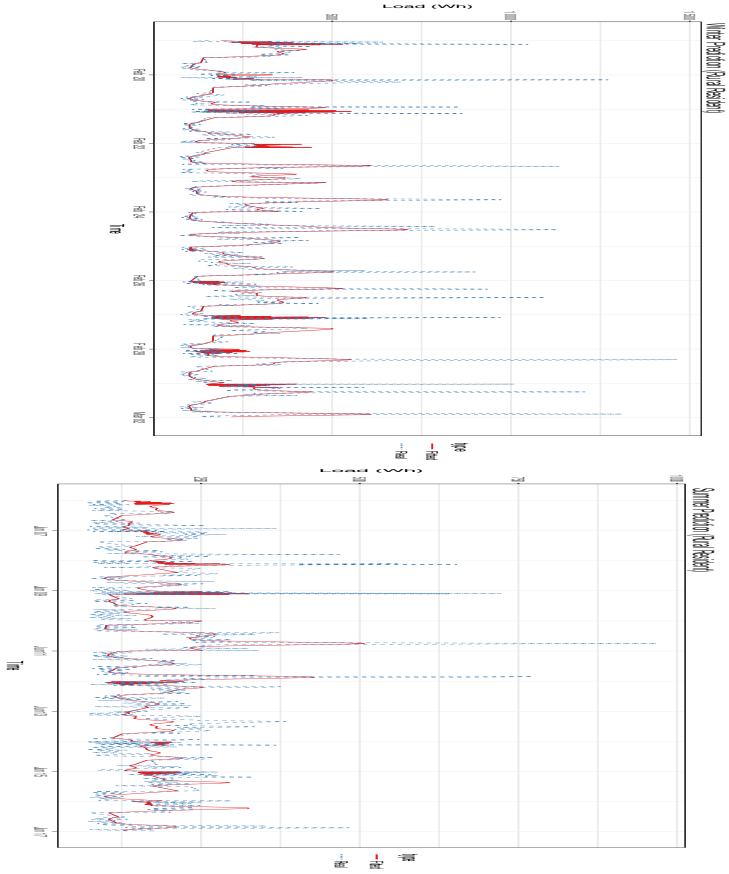


Figure 5.12: GAM fit for a customer that belongs to OA characterised as 'Rural Residents'

As can be observed from above, sampled rural resident consuming gas tends to have semi periodic consumption patterns. Persistent presence at the house/property can be noted from a continuous through the day consumption. One should note, that while magnitudes of consumption in winter and summer differ, the patters themselves may share some similarities.

A more detailed picture of the fit presented by 3D graphic in Fig. 5.13 which illustrates the spline which is fitted using daily and weekly variation. It may be noted that for winter month, variability of consumption is evident for almost every day of the week, while for summer, the largest peaks of consumption correspond to Saturday and Sunday (6th and 7th day).

The average behaviour over time is represented by the smoothing splines below (Fig. 4). Something different is seen here compared to electricity readings that were presented earlier. Daily or weekly consumption has nothing to with an idea

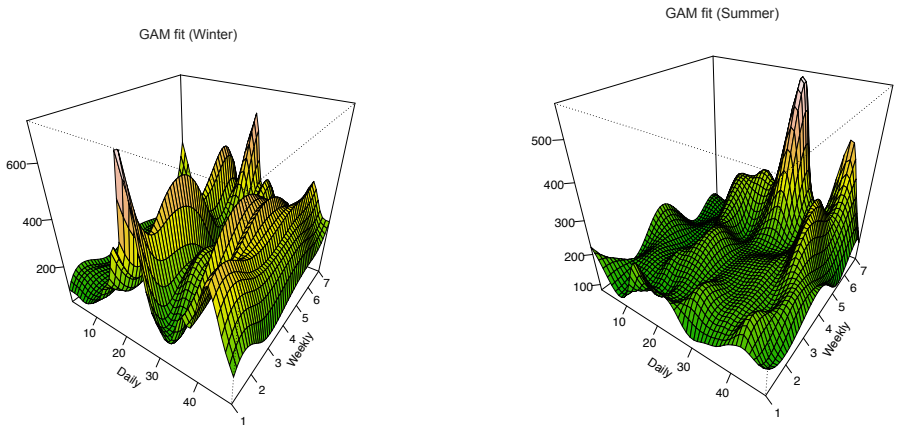


Figure 5.13: GAM fit for a customer that belongs to OA characterised as 'Rural Resident' in 3D.

of a 'typical user'. There is evidence of continuous consumption of gas through the day. For a small winter sample, consumption appears to be constant across the day. Whilst over a six month time span it can be described by a greater peak in the morning that then decreases throughout the day. Mid week for this customer is associated with highest levels of consumption.

A slightly more detailed picture can be seen from regression table below (Table 5.6). The complexity of the splines characterised by smaller EDF suggest that on average gas consumption may require less complex smoothing splines compared to the electricity. There is no weekly significance in explaining the next half hour energy load for smaller sample. Unrestricted model provides a slightly better fit (R squared of 0.45) yet still not able to explain even a half of variation in energy use of this customer. As in the previous sections, the larger sample hold in way more variability that model cannot explain using the time only.

Residuals checks below are presented to complement results above. It can be seen that the model fit is more satisfactory on a smaller sample.

5.5.2.2 Urban Professional

As previously, to complement the visualisation of the consumption trends, 3D visualisation of daily fit are presented in Figure 5.18. High periodicity of double-peaked consumption across the week can be noted. There are almost no differences in the

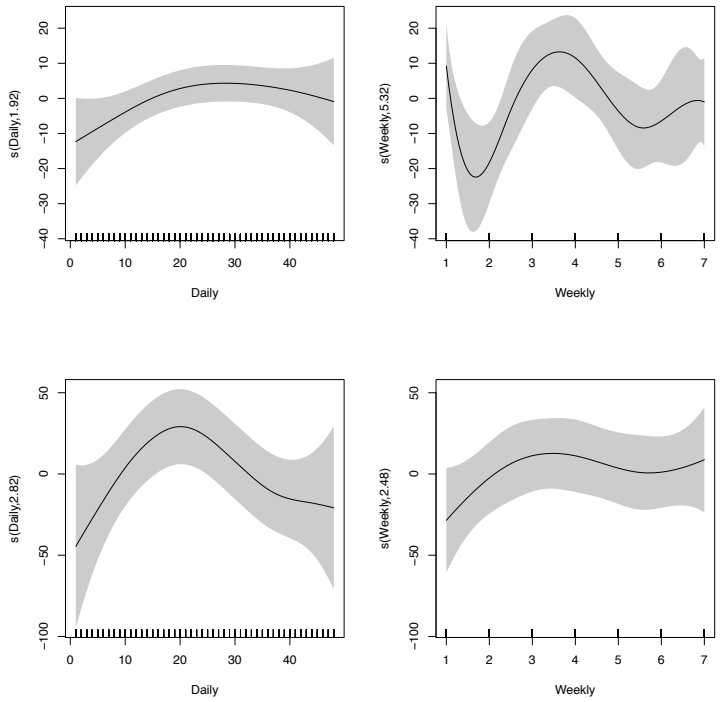


Figure 5.14: The fitted response due to each variable/covariate contribution in winter, 'Rural Resident' .Winter sample (up) and 6 month sample (bottom)

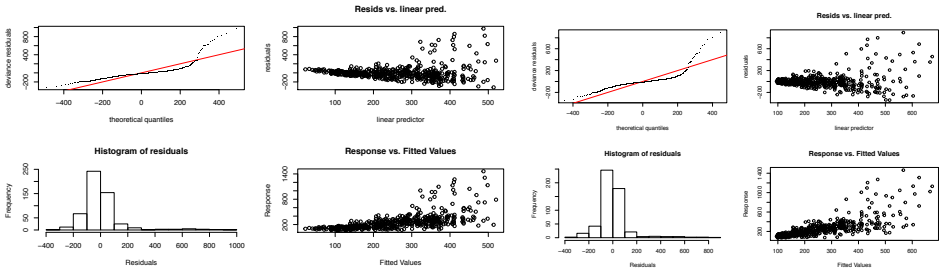


Figure 5.15: Residual Check for Winter Fit, 'Rural Resident'. Restricted (left) and unrestricted model (right).

consumption patterns across the weekdays for this smart meter user.

Sampled urban professional represents a very evident difference to the example from a rural output area. The pattern can be described as highly periodic and also can be described by high predictability when using GAM (R squared of 0.75 for winter). In the winter illustration of Figure 5.17 there is a very consistent occurrence of morning and evening peaks with no consumption over night. There are also almost no effect of the weekend which perhaps may have contributed to such high

	GAM	GAM + Inte. (Unrestricted)	GAM + Inter.	GAM	GAM + Inter. (Unrestricted)	GAM + Inter.
(Intercept)	246.20***	246.20***	246.20***	235.93***	235.93***	235.93***
	(7.28)	(6.25)	(6.88)	(2.26)	(2.26)	(2.26)
EDF: s(Daily)	9.20***			12.93***		
	(11.49)			(16.14)		
EDF: s(Weekly)	2.71			5.44**		
	(3.14)			(5.84)		
EDF: te(Daily,Weekly)		99.27***			55.89***	
		(124.88)			(73.50)	
EDF: t2(Daily,Weekly)			55.94***			27.96***
			(70.92)			(34.71)
AIC	6919.03	6835.09	6901.36	83102.84	83135.54	83102.63
BIC	6978.42	7267.43	7148.69	83240.29	83526.11	83304.74
Deviance explained	0.27	0.55	0.40	0.20	0.21	0.21
R^2	0.25	0.45	0.33	0.20	0.20	0.20
GCV score	28657.26	25462.89	28046.57	32141.88	32311.82	32141.17
Num. obs.	528	528	528	6288	6288	6288
Num. smooth terms	2	1	1	2	1	1

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.5: Regression Output, 'Rural Resident'

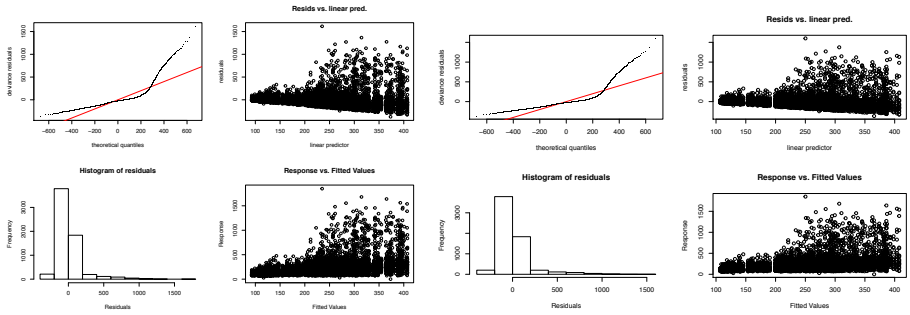


Figure 5.16: Residual Check for 6 months sample, 'Rural Resident'. Restricted (left) and unrestricted model (right).

explanatory power of the fitted model. In the case of summer prediction, one may note a very small consumption loads which are slightly irregular through the week. The absence of consumption may correspond with physical absence of the smart meter user.

Figure 5.19 shows that more variability is present in larger sample. This is relevant for daily variation. However, it can be observed that contribution of individual

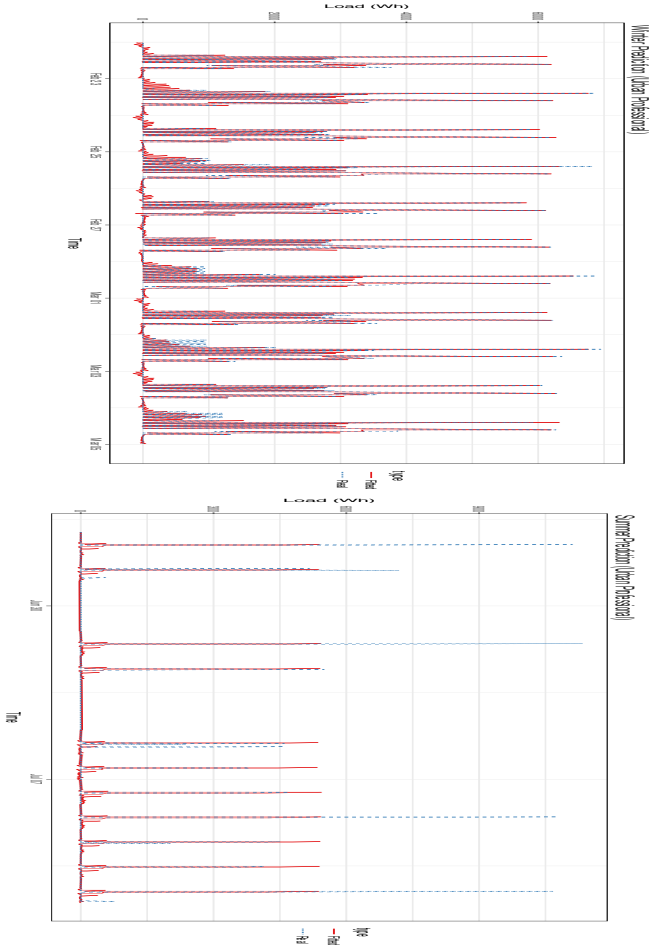


Figure 5.17: GAM fit for a customer that belongs to OA characterised as 'Urban Professionals'

days of the week is almost invisible on a smaller sample yet has some evidence of increasing effects of weekends on a larger sample.

Looking in more details (Table 5.6) one can see that for this particular user variation within the day is significant while total consumption over each day throughout the week doesn't seem to affect the energy load. For unrestricted model the EDF are suggesting high complexity of the smoothing spline. Significance is measured by F tests and lower p-value indicates the we can reject the null hypothesis that daily or weekly variables have no impact on the variation in the energy load. Further measures of fit to look at is the R squared that was mentioned earlier to suggest how well the functional form of presented models explain variability in energy load.

Regardless of whether one uses simple model with no interactions, unrestricted

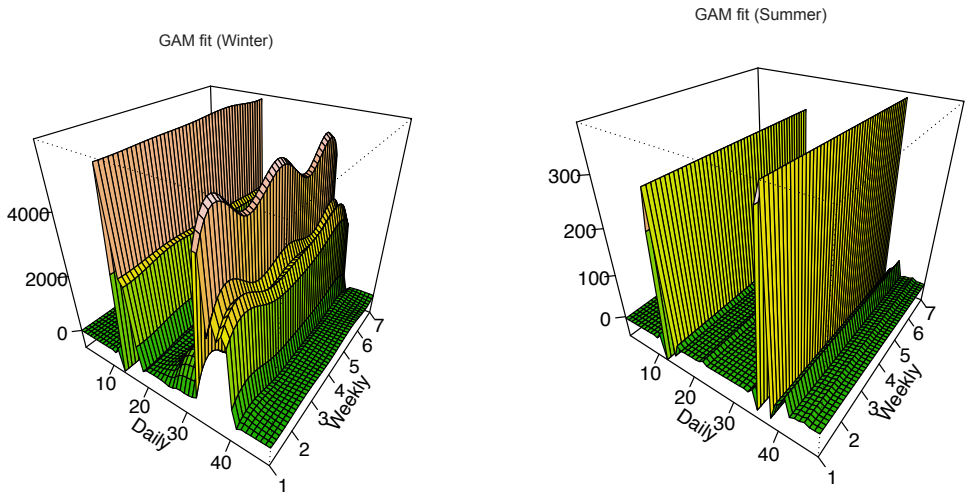


Figure 5.18: GAM fit for a customer that belongs to OA characterised as 'Urban Professionals' in 3D.

(Intercept)	809.22***	809.22***	809.22***	499.86***	499.86***	499.86***
	(9.78)	(7.10)	(9.39)	(19.56)	(19.67)	(19.54)
EDF: s(Daily)	46.97***			46.51***		
	(47.00)			(46.99)		
EDF: s(Weekly)	5.83***			5.78***		
	(5.98)			(5.97)		
EDF: te(Daily,Weekly)		297.19***			93.00***	
		(320.44)			(94.90)	
EDF: t2(Daily,Weekly)			65.85***			55.71***
			(69.31)			(57.75)
AIC	7269.01	7037.55	7237.48	110292.89	110405.38	110282.05
BIC	7502.95	8314.84	7527.13	110659.14	111046.28	110671.41
Deviance explained	0.98	1.00	0.98	0.17	0.17	0.18
Dispersion	50466.87	26618.63	46529.38	2406261.86	2434072.48	2400817.96
R^2	0.98	0.99	0.98	0.17	0.16	0.17
GCV score	56192.53	61158.49	53274.30	2426827.91	2471011.27	2422669.08
Num. obs.	528	528	528	6288	6288	6288
Num. smooth terms	2	1	1	2	1	1

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5.6: Regression Output, 'Urban Professional'

or restricted model, smaller sample pattern can be easily predicted using the dimension of time. Periodic behaviour of the sampled smart meter user can be thus well described by GAM. Larger sample performance however is poor. Meaning that once again long term trends of the consumption may be in need of additional explanatory

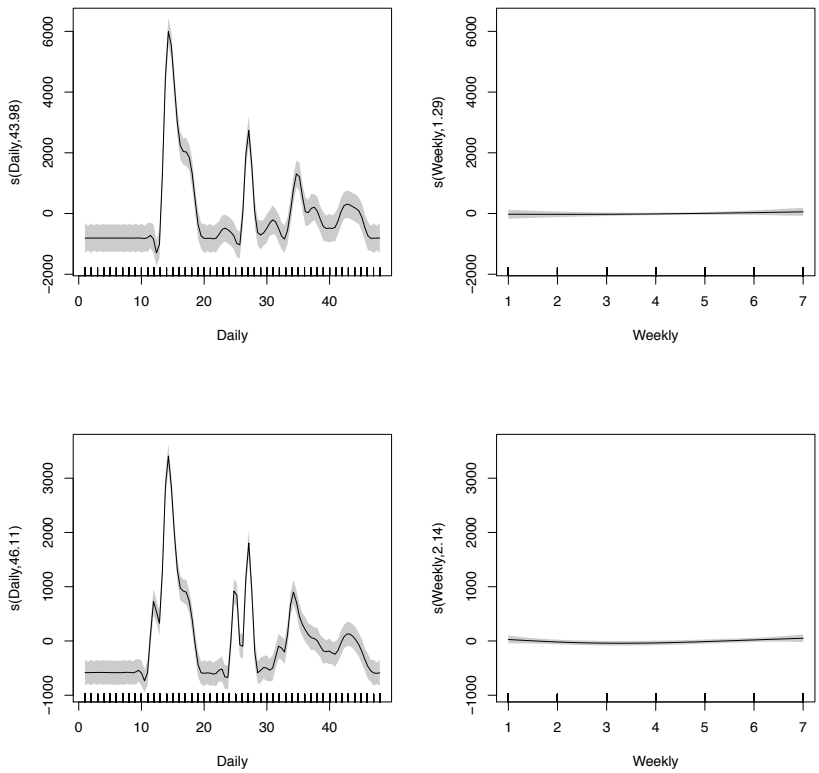


Figure 5.19: The fitted response due to each variable/covariate contribution, 'Urban Porfessional' .Winter sample (up) and 6 month sample (bottom)

factors to be measured and included alongside.

Further residual diagnostics have shown that while for the smaller sample model certainly behaves best among the four experiments presented in the chapter, once bigger sample is taken more variation in energy use seems to occur that model simply collapses. This can be seen from R squared, AIC and BIC.

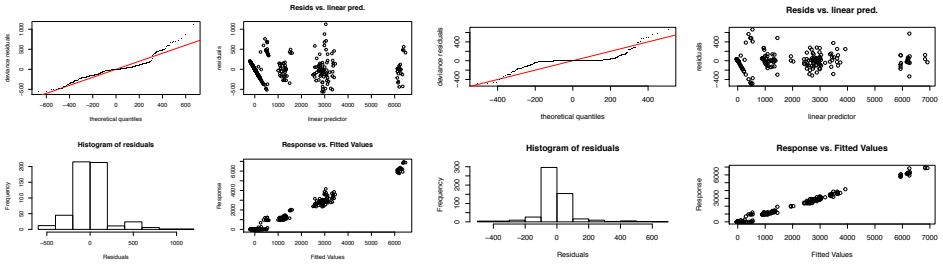


Figure 5.20: Residual Check for Winter Fit,'Urban Professional' . Restricted (left) and unrestricted model (right).

To conclude, the sample rural resident that consumes gas in winter is shown to

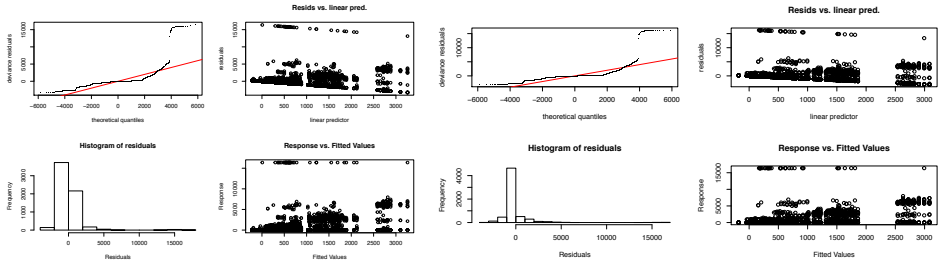


Figure 5.21: Residual Check for 6 months sample, 'Urban Professional'. Restricted (left) and unrestricted model (right).

be highly periodic and predictable given the results of regression analysis.

Please note that these results are relevant only to these particular smart meter users and cannot be extrapolated to all urban/rural residents in the selected area, neither can they suggest that this will be valid for average energy consumer. The results above are further narrowed down to winter sample as this season is associated with more variability. This variability may have been crucial in explaining why GAM model on average were associated with best fitting performance pre-dominantly on smaller winter samples.

5.6 Limitations

It is important to note some of the immediate limitations in the presented analysis are once again associated with the availability of data about the smart meter user characteristics. An an approximation of those, Census OA Classification, has shown some potential to undermine the reasons behind the variability of the energy consumption cycles in the selected sample. However, these results are by no means conclusive and should not be taken as definitive representation of urban and rural differences in energy consumption.

5.6.1 GAMs are complex

Other limitations with the presented experiments are associated with the nature of the GAM model. It is obvious that it is a complex model and given the number of smoothing functions available out there it is also a hard model to be designed such that the best combination of the functional form or combination of functions is chosen. Very simple structures of the model were used in this chapter in order to

preserve some opportunities for interpretability as well as avoid over fitting problem. As in other chapters, the black box solutions are least favourable here. Mainly because if one acts on these findings, the simpler the model combined with its interpretation may serve to be more useful. This can be true for cases where one wants to understand periodicity or regularity of energy use which can be directly presented by smooth basis functions.

5.6.2 Residuals

As was seen from the experiments in the chapter, on average the residual terms in the model output for smaller samples tend to follow a normal distribution shape. However, the uniqueness of energy consumption on each day by different consumers, add noise which in turn adds outliers values that may undermine the assumption about the error term. It is important to say at this stage that outlier values when it comes to energy consumption recordings may actually be quite useful. They are the characteristics of certain behaviour and over time may be systematic. If one has more data on each consumer, say their two-three years of consumption these systematic behaviours may be picked up and modelled more easily.

5.6.3 Imprecision

The results of the GAM fit in this chapter may suffer from imprecision. It is hard to be able to use the results to forecast exact energy consumption in the next hour using these type of model fit due to its smoothing nature. However, this is a rather useful property for generalisability of the results that suggest about periodicity of energy use by particular consumers. So on the other hand, it can still be useful to describe, if not predict behaviour. In the case where many consumers are pulled together, smoothing may be a better strategy as it will be less sensitive to outlier consumption patterns and provide a more general picture of how people consumer energy in groups that can be then referenced to specific geographical areas or even the whole country.

5.7 Discussion and Conclusions

This chapter has presented an attempt to describe energy load by studying past consumption recorded by smart meters. An inference into how different days of the week and different weeks of the months may possibly tell us about future energy use was presented using Generalised Additive Models (GAM) regression analysis. This is the first application of GAM models to residential energy patterns. Previous work only looked at industrial electricity consumption (Laurinec, 2016; Wood, 2006). Possible links to the socio-demographic characteristics of the ares under they study were further accessed to see whether simple distinction between urban and rural areas may have impact on periodicity in consumption and as a consequence, its predictability.

GAM approach is certainly not one of the most straight forward to use and suffers from complexity that undermines overall interpretation of the results if we were to put it in the social context that exists outside the data. As the energy consumption itself can be hardly described using simple linear model, GAM may perhaps be one of the best alternatives one can look at for analysis of smart meter data time series. It was shown that on average, where customers are periodic the model can pick it up efficiently using the mix of smoothing functions that are fitted on different part of the series. It is important to note that such high accuracy is achieved only on a fairly small sample, where various structures of GAM were attempted before choosing the most optimal one. If use on big sample of fitting the same GAM to different smart meter users the results may not necessarily look as neat as in this chapter.

Having more data on consumers, particularly an ability to have greater geographical resolution, would certainly improve the model results. Time series data on weather, activities and appliances use recorded along side smart meter data may complete the understanding of greater variation in the energy use. Precision in customers location, offers further precision in possible temperature changes in the area as well as unique characteristics of buildings types and use that can improve the explanatory power of GAMs.

Chapter 6

Methodology and Results: Customer Label Prediction

6.1 Introduction

This chapter will look at the other and very different type of prediction task known as classification. As with forecasting, this method can be applied to both an individual energy use as well as to grouped customers energy patterns that was defined through clustering algorithm. The chapter will begin with the discussion of the context in which classification problem is attempted. Namely, it is focused on prediction of specific label for the customers that may be characterised by the vulnerability towards paying energy bills. Prediction and identification of these kind of customers is of high importance for both energy companies and the regulators. Unlike two previous chapters which were pre-dominantly methodological, this chapter bears greater substantive application. More details are provided below.

The UK is known to be one of the pioneers in introduction in 1994 a policy of energy suppliers' obligation in energy efficiency. Focus on carbon savings at residential level, this regulation has led to initiatives such as subsided home insulations, free boiler replacements and various modifications to energy company tariffs that can help modify the energy use to be more sustainable. It further expended to considerations of more pressing societal issues such as fuel poverty and its reduction[1].

[1]Such targets are imposed through Energy Company Obligation (ECO) and administered by the government regulator, Office of Gas and Electricity Markets (OFGEM). Apart from the UK, policies

Big energy suppliers are the ones that are targeted mainly by the regulators. These companies are expected to provide house insulations and financial assistance to their customers at no cost, including free installation of smart meters. When deciding on which customers may qualify for the financial support the income to bills spending is used for identification of fuel poverty/energy vulnerability. The complimentary way to access such vulnerability is also looking at how much energy being consumed by the customers compared to their expected use. For instance, if you were supported would you consume more energy to bring yourself closer to adequate levels. Whilst defining how these customers can be picked up by energy supplies may sound straight forward, it remains to be one of the greatest challenges for suppliers to find vulnerable or at risk of becoming vulnerable customers. Current financial support which is offered to customers comes through self selection process where energy customers themselves report that they may benefit from financial support.

Now that energy companies are obtaining more and more data on their customers energy readings from smart meter, the question is whether these customers may be actually identified from smart meter data using various classification and segmentation techniques. In Chapter 4, it was shown that smart meter data when desegregated can be meaningfully segmented into distinct groups of consumption. This chapter aims to test whether smart meter data can be further segmented into groups that have qualitative characteristics or label. In this case, energy consumption vulnerability flag.

6.1.1 Structure of the chapter

The data used in this chapter is on gas consumption is utilised from smart meters installed across northern England and Scotland for the period from 2014 to January 2015, and existing data on household energy vulnerability derived from the fact that customers have been enrolled in various kinds of financial support. Some of the similar methods which were seen in Chapter 4 are re attempted here. As the

to tackle fuel poverty poverty have been introduced in New Zealand (Howden-Chapman et al., 2012; O'Sullivan et al., 2015), Indonesia (Andadari et al., 2014), Japan (Okushima, 2016) and also, in a number of European Union countries such as Italy, France, Belgium and Spain.

case study based on unbalanced sample (majority of the customers in the sample are characterised by absence of vulnerability flag), few techniques to balance the sample were applied to the data in the preparatory step.

This chapter further discusses how the accuracy of analytical models may vary over different types of data and alternative methods. The next section will discuss the data that was used for the experiments. It is followed by methodology overview which outlines the analytical strategy that was used to classify customers to discover different consumption patterns in the data. The last section concludes and provides suggestions for the policy implications of the findings in this chapter as well as some suggestion for future work in customer label prediction[2].

6.2 Label Prediction: Energy Vulnerability

The case study is based on the sample of 1,919 smart meters from a region in northern England and parts of Scotland. Each meter was recording half-hourly consumption in kWh or Wh depending on the source. For the sake of simplicity, a binary vulnerability flag was created to indicate whether one or more support measures associated vulnerability characteristics were applied to the customer.[3]

For the analysis below, all smart meter readings for the month of February that are then used later in our prediction model. Table 6.1 provides an overview of the data. In the given sample 24% of customers were classified as vulnerable and 76% as non-vulnerable. On average, per day, overall consumption is recorded at 98.69 kWh (median 67.44) with standard deviation 39.34 (minimum at 0 and maximum at 181.72).

[2]The work in this chapter is a continuation of the MSc Thesis in which the prediction of energy customer vulnerability was attempted using neural network model. This chapters extends the analysis seen in Ushakova (2015) by suggesting more appropriate modelling techniques as well as it modifies the issues that arise due to unbalanced distribution of the categorical labels in the experimental sample. The data which is used in this chapter was available for further research till March, 2016 and is restricted in use due to non disclosure agreement. This has prevented usage of this data for other sections in the thesis

[3]This includes information on customers who have been enrolled in priority services, belong to a group of credit customers in debt, customers on Fuel direct, customers receiving a grant, vulnerable customers off supply and those who receive a warm home discount

Data	N (smart meters)	N (days)	N (daily readings)	N (total observations)
One month sample	1,919	28	48	2,372,592
Overall dataset	1,919	390	48	33,309,120

Table 6.1: *Data structure.* The structure of the smart meter sample and the one month subsample that is used in the prediction model below.

As a simple visualisation of the data monthly consumption patterns for two groups of customers as categorised by the energy supplier are presented below. Figures below illustrate examples of consumption patterns for retired consumers and families, both in the case of vulnerable and non-vulnerable consumers.

6.2.1 Box plots of consumption patterns for randomly-selected customers in a given demographic class

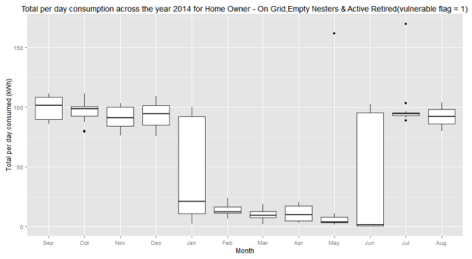


Figure 6.1: Vulnerable *"Retired and Empty Nester"* customer

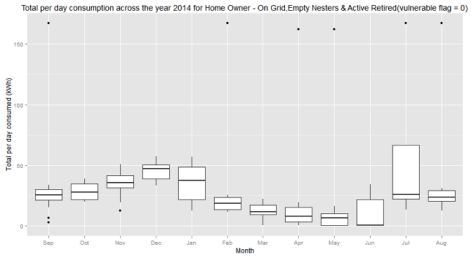


Figure 6.2: Non-vulnerable *"Retired and Empty Nester"* customer

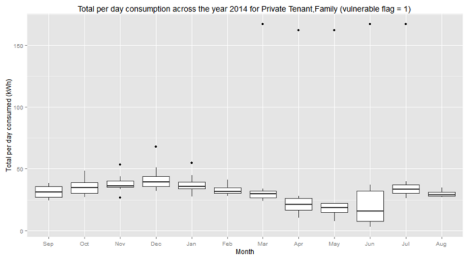


Figure 6.3: Non-vulnerable *"Family"* customer

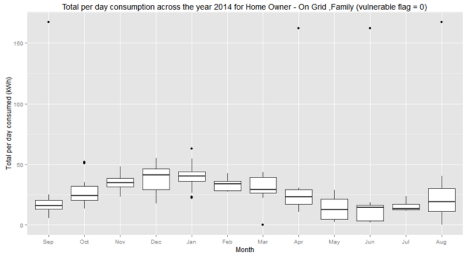


Figure 6.4: Vulnerable *"Family"* customer

For retired individual in Figures 6.1 and 6.2 consumption was generally high but dropped mainly in January. Non-vulnerable customer experienced a smoother profile across the year with more variation in July compared to other months. Overall, retired customers receiving support tend to consume more gas compared to non-vulnerable customers. The family group in Figures 6.3 and 6.4 demonstrates consistent consumption over the year, and in fact increased energy consumption during spring time in comparison to previous figures. This may be attributed to using gas for cooking and more hot water as family size may be also correlated with consumption. Presence of the vulnerability flag does not necessarily imply a sizeable difference in consumption for sampled individuals.

Figures above are based on data of very fine granularity and used here to highlight the complexity of aggregating the data on consumption due to differences in individual consumption profiles. Many additional customer characteristics are omitted, and it cannot be assumed that sampled customers are representative of consumption patterns for a given life-stage group. Tenancy and property characteristics as well as geographical location may play a significant role in the observed differences in consumption. Furthermore, while the data was cleaned by focussing

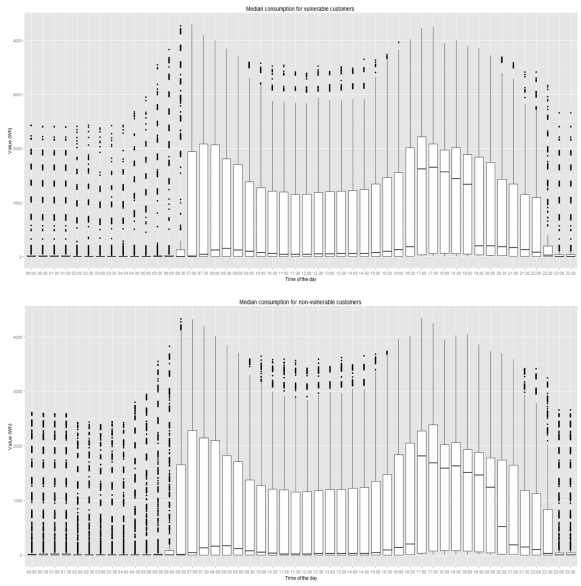


Figure 6.5: Median half-hourly consumption for vulnerable and non-vulnerable consumers for the month of February

on weekday usage there are still a number of outliers, especially for non-vulnerable customers. This may be due to, for example, sudden weather changes.

As a further illustration we look at all individuals for weekdays in one month. Figure 6.5 plots median consumption by vulnerable and non-vulnerable customers during February. It is clear, that the median half-hourly consumption for vulnerable consumers exhibits similar peaks to non-vulnerable consumers, however, there is a slight difference in outliers and magnitude of peaks. Winter, as in previous chapters, was selected for both visualisation and prediction with the underlying assumption that there is greater variation in the consumption patterns throughout the winter. For Scotland and Northern England, according to Met Office (2015) January and February tend to be the coldest months of the year, but February may be more isolated from the effect of winter holidays Cao et al. (2013).

The results in Figure 6.5 indicate high levels of variability within the data, as well as presence of outliers especially for consumption during night hours. While most of the vulnerable sample tend not to use gas during night, there are still a number of individuals exhibiting high consumption levels. It is also very difficult to discern any differences in the consumption patterns between two groups of cus-

tomers.

6.3 Predicting consumer vulnerability

In the previous section, it was shown that visual inspection of two groups may not necessarily suggest differences in the energy use, especially in the case of aggregated patterns. Given the size of the sample and inability to check each individual profile separately, machine learning methods are being applied to the data, as in other chapters, to study whether the algorithm can pick what differentiates two groups from numerical point of view.

The methodology for this study is based on supervised machine learning techniques that are commonly applied in big data analytics and were previously used in the analysis of smart meter data. Some of the familiar tree methods such as Random Forest that was used in clustering section will be presented. Nevertheless, the analysis in this chapter extends existing literature with more targeted classification and prediction — identification of vulnerable customers. Tree methods are further tested against some of the very common classification methods that can be found in literature (i.e neural network, support vector machines and naive Bayes) for further robustness checks.

6.3.1 Balancing the sample

As a pre-stage, given the uneven distribution of consumer vulnerability flag, the methods for sample re-balancing were attempted. These are hybrid method or synthetic minority oversampling technique (SMOTE). SMOTE uses simultaneously both under-sampling and over-sampling of the classes, by reducing the majority class through dropping observations randomly. It has been shown in Kuhn and Johnson (2013) that SMOTE and under sampling are associated with higher receiver operating characteristic (ROC) and improved sensitivity and specificity. SMOTE was used primarily for this case study. One of the main reasons to re-balance the sample is to avoid higher prediction accuracy for one group compared to another. Given that non-vulnerable customers are over-presented in the sample, there may be a risk that model will earn how to predict well the group which is represented by majority

of the patterns and not the one we are primarily interested with: vulnerable energy customers.

6.3.2 Random forest

As random forest was already discussed in details in Chapter 4. This subsection considers a minimal reference to the algorithm structure yet gives slightly more details about the model with the reference to problem specification such that is clear what is the outcome variable and the predictors are.

Random forest classification for vulnerability identification is performed in the familiar by now stages of the random forest described earlier in the thesis. First, the algorithm selects a bootstrap sample to be analysed. The tree is then built through repetitive steps until the optimal combination of variables for predictions with minimal error is found. Each time, the model selects variables at random. In our case, there are 48 variables that correspond to each half-hour smart meter reading every day of the year. The outcome variable is the vulnerability flag for each consumer. The learning algorithm begins on two randomly-selected predictors of vulnerability flag and expands until covering all 48 predictors. The tree is identical to the decision tree mechanism, where the decision is based on how each variable contributes to further splitting of the data until we can reach our final classification split – into vulnerable and non-vulnerable classes. One advantage of using a random forest model is that it allows for the building multiple trees, rather than just one. Through such a process the algorithm mainly looks for trees that would build associations between input and output variables. A higher variation in the data allows the algorithm to easily differentiate what contributes to vulnerable and non-vulnerable classes, and split the data further. Furthermore, there is no need to correct the model for seasonality or time dependencies as random forest logically would separate those in the training stage. A brief overview of the methodology is given below:

As was observed from the data visualisation, for smart meter data it is expected that evening or morning gas consumption levels would have a greater impact on the learning process, while overnight or afternoon consumption should have a relatively smaller influence on the relationship between input and outcome variables.

One of the advantages of using random forest models is low probability of over fitting the data Friedman et al. (2001a) as it is mainly based on decision trees rather than optimisation problems. The optimisation nature of the algorithms are core to the neural network and support vector machine algorithms. As part of the robustness studies, these models were still included and presented briefly below.

6.3.3 Neural networks, support vector machines and naive Bayes

The models discussed here are suggested for the analysis of large datasets Friedman et al. (2001a). The specific choice of the model is often motivated by data variation and whether one may expect a linear or nonlinear relationship between predictors and the outcome.

Neural network methodology is based on defining neurones that connect input variables to the outcome, the multilayer structure of the model allows it to represent complex non-linear mappings. In our specification, the analysis is built on a logistic regression model for the hidden layer that connects smart meter readings to a binary vulnerability flag. Minimisation of the sum of squared errors is done by the gradient descent algorithm.

Gradient descent works by using the first order condition of the function in order to find the local minimum point. By taking small steps from a proxy of gradient for a given function, both local maximum or minimum points can be approached through a number of iterations. For this study the number of iterations was raised depending on the size of the sample due to the fact that each customer has a unique combination of inputs and the model may need a reasonable amount of time to converge. Neural networks have been previously been used for energy consumption point predictionNizami and Al-Garni (1995); Tso and Yau (2007); Haghi and Toole (2013); Lee et al. (2012). Neural network models usually outperform other approaches such as linear regression, decision trees or support vector machines for the point prediction using historical data. However, in our case we observe a rather poor performance, perhaps due to the classification nature of our prediction problem and noisiness of the data. The latter issue complicates finding a unique solution

to the optimization problem.

Support vector machines (SVMs) are based on the minimization of the cost function through a similar gradient descent approach. Instead of having a hidden layer connecting the input and outcome variable, the algorithm is based on initially creating a nonlinear feature space where it then seeks to fit a linear regression that may separate the features into two classes. While the use of SVMs in energy consumption studies is not extensive, several studies show good prediction performance of such models. For example, Mohandes et al. (2004) focus on wind speed prediction from historic daily averages using multi-layer perceptron (MLP) neural networks and support vector machines. Support vector machines outperformed MLP in terms of prediction accuracy. Dong et al. (2005) use SVMs to predict energy consumption of commercial buildings in Singapore.

Finally, a naive Bayes classifier is also considered as in this particular setting it allows is to calculate the probability of a users vulnerability flag by forming a posterior about the outcome. This posterior updates as more smart meter readings are taken. Thus, with more data available greater prediction accuracy. is expected The probability of the outcome variable to be either zero or one is estimated using the maximum likelihood approach. Naive Bayes, as shown in Rish (2001), relies mainly on the assumption that the features are independent of the predicted class, and performs well on highly-interdependent features. The prediction power would gradually decrease if the class zero is over represented in the sample. In our case, after re-balancing the sample, we could not report a highly visible difference in the prediction power using naive Bayes. This is likely attributable to high variation in the half-hour loads. In addition, the heterogeneous levels of interdependencies associated with half-hour consumption may also arise from idiosyncratic usage of natural gas at household level.

In practical work, algorithms often differ in how they utilize predictors that are less statistically important for identifying the relationship between input and outcome Breiman (2001a). Whereas random forest models benefit from such weak inputs, for neural network and support vector machines this additional noise may

Model	Accuracy	Recall	Precision	F-score
Neural Network	60.11%	0.64	0.66	0.65
SVM (radial kernel)	76.2%	0.81	0.99	0.89
Naive Bayes	56.0%	0.81	0.85	0.83
Random Forest	**94.6%**	**0.81**	**0.79**	**0.80**

Table 6.2: Results (ten folds cross validation) for each model that was used to predict vulnerability flag using consumption data

detrimentally affect the solution. Caruana and Niculescu-mizil (2006) show that for highly variable and complex data sets, or those that have information on real-world complex problems, naive Bayes is expected to be outperformed by models like random forest. The results confirm this earlier prediction.

6.4 Results

The accuracy of prediction for each model are presented in this section. The optimal model parameters for each model are selected through ten-fold cross validation. To assess the performance of the models, Table 6.2 reports accuracy, precision, recall, and F-score. The summary of the results suggest that Random Forest outperforms alternative models in terms of overall accuracy. As was also seen in Chapter 4, it only confirms the observation that the random forest may have a greater power in differentiating similar patterns of consumption.

As the choice of alternative models was largely driven by their popularity in academic research, the observed prediction results are in line with the literature in related fields. For example, Lines et al. (2011) focus on appliance consumption predictions and compare naive Bayes, random forest, neural network, and SVM (also using cross-validation to select optimal model parameters). They show that Random Forest slightly outperforms other models on their data.

One of the advantages of using random forest is its interpretability. In this example especially, analysis of variable importance for decision tree split allows for identification of which hours are important for distinction between the two groups. Alternative models like neural network and SVMs are often treated as “black-box” solutions due to the fact that they can be rarely opened up in a similar fashion for

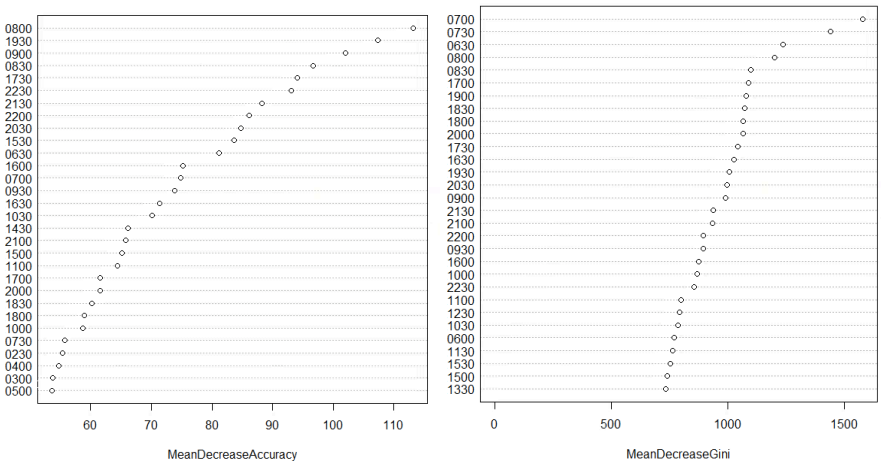


Figure 6.6: Mean Decrease Accuracy and Gini by variable importance

further inspection on why certain model has performed well. . With random forest one can assess which variables are significant for prediction accuracy. Figure 6.6 provides a summary of variable importance tables that indicate the variables in the order of importance for prediction power and their contribution to subsequent tree splits based on Gini impurity criterion.

By importance ranking, as expected, morning and evening peak hours have a strong contribution to prediction accuracy. By Gini, morning hours tend to be more important in their contribution to the split of the decision tree. This is highly intuitive as one may expect that customer vulnerability can be more more evident through the distinct behaviour during the peak hours.

In line with the original argument by Breiman (2001a), the weak inputs in the dataset make achieving high prediction accuracy with neural network or SVM more challenging. Variation that arises from these variables adds more noise and contributes to more confusion in convergence to local minimum point. In our case, the variables' importance in Figure 6.6 shows that almost half of the variables are not critical for prediction accuracy. Random forest appears to have taken weakness/importance into account, thus achieving maximum noise reduction.

6.5 Conclusions and Limitations

The research presented above aimed to answer the question of whether there is a potential to identify vulnerable natural gas customers by using data from smart meters using various machine learning methods. This section further has extended previously shown analysis to more targeted prediction of customer label such as fuel poverty/energy vulnerability.

Vulnerable customers were expected to under-consume in transition to the winter period, yet this was hardly observed in the data. Vulnerable customers have shown more constant and distributed over the day consumption profiles while the non-vulnerable tend to exhibit peaks and have quite uniform patterns of consumption within the sample, implying that they may leave home at certain periods while vulnerable customers may consistently use gas in their homes. Nevertheless, some patterns in both groups were similar which may have been a reason as to why most of the models failed to provide high prediction accuracy. Tree based models once again have shown better performance compared to other methods. it was further shown that in classification of the smart meter readings , the peak hours play an important role for differentiation between two customers groups.

It is important to acknowledge that while this analysis produces some insights, it should be seen as a demonstration, rather than complete solution, for how smart meter data can be used to understand and predict vulnerability. Some further extensions that can improve the presented work would be inclusion of time indicators that suggest how support was provided to customers. This will allow for study of so-called treatment effect of intervention into energy consumption by smart meter user.

Some trivial policy implication can be drawn from this section. Mostly it is a relationship between high heterogeneity of gas consumption households and the meeting of the objectives set by OFGEM , mainly targeting and supporting fuel poor. Energy consumption vulnerability remains an ambiguous concept in both technical and social contexts, which may require further research to build more inclusive and transparent indicators. As suggested in Schmidt and Weigt (2015),

the study of energy demand and consumption requires a highly interdisciplinary approach especially if policymakers are interested in shaping and transforming current energy systems. Thus, both social and political science as well as engineering and data science may be helpful in answering such research questions.

Chapter 7

Scaling Up: Data Reduction and Transformation Techniques for Smart Meter Data

7.1 Introduction

Prior to this chapter, the analysis presented in the thesis was performed on various samples of the large dataset that was specially available for this research. Aggregated and disaggregated samples were considered for tasks of segmentation and forecasting. This chapter aims are to address how one may consider scaling up the data analysis in the previous chapters to the whole sample such that it is suitable for instance, in cases where data may arrive and update at real time. Certainly, standard computing systems will struggle to process such large amount of data if we were applying the computational methods seen earlier to full sample. To give a rough idea, to predict a fuel poverty/energy vulnerability flag on February sample takes around a week to process for a single model. This is an issue that could serve as a barrier for using advanced or more importantly, the most suitable methodology when answering various research questions with smart meter data.

A number of advances in the area of mathematical transformations for time series data may come up handy for tackling the issues of computational limits. Primarily referred to as spectral analysis and signal processing based methods, these

techniques allow for data reduction through the meaningful transformation of the time series data such that the uniqueness of patterns is preserved, however, takes up less space, i.e. the data is compressed. Smoothing, as seen in previous chapter, is one example of such approach, whereby the data is represented as a set of basis functions. However, in smoothing, little attention is given to time and space dimension, meaning that functional form is not driven by periodicity or cyclical behaviour of the data. This chapter extends the survey of transformation methods available to both the Fourier and Wavelet transformations, and examines how they may help in data reduction for clustering and prediction applications on the full dataset. All the techniques are vital for another possible challenge associated with the analysis of smart meter data: privacy. In this context, the ability to transform smart meter data series into a less visually identifiable sequence of activity can help preserve smart meter users' anonymity. Lastly, another application arising for the work in this chapter could be identification of anomaly behaviour or fault issues with smart meters that can be identified by the periods of irregular readings, or the complete absence of readings for certain period of time.

The robustness of the proposed smart meter time series transformations are checked by looking at the recovered pattern of the energy consumption from the transformation. The significance levels of various periodic behaviour over different time spans are also assessed. The strategy is essentially to check how much information can be lost/maintained from the energy consumption patterns by projecting onto certain functional forms. As pointed out by Nason and Von Sachs (1999) 'there is no such thing as one statistical time series analysis as the very many different fields encompassed by time series analysis are in fact so different that the choice of a particular methodology must naturally vary from area to area'. As this thesis is rather an applied work, it wont go in too much details about various modification and extensions that can be added to presented time series transformations. The aim is to keep it simple and generalisable at this stage, while highlighting a definite opportunity for future work in the area of signal processing.

The rest of the chapter is structured as follows. The preliminary sections dis-

cuss the significance of data reduction in the emerging area of big data research and provide a quick overview of the statistical tasks that needs solving. The examples of methods available such as Principal Component Analysis, Fourier transform and Wavelet transform are then presented. Given the description of how these methodologies work, it is suggested that Wavelet transformations might be the most appropriate for smart meter data. The results of wavelet application are presented in the subsequent section and followed by some conclusions and possible further work.

7.2 Importance of Data Reduction

Previously in the thesis, various sampling approaches were used for both clustering and prediction tasks. Specifically, sampling techniques that involve aggregation or narrowing down the temporal resolution of data (i.e. selecting specific month) have rather meaningful reasons as it allowed for researching smart meter user data in more details on the smaller sample, studying very granular and unique variation. However, if the analysis applied in the thesis was to be replicated in an industry setting such detailed approach needs to be compromised. This mainly driven by the requirements for privacy and customers anonymity preservations.

Data reduction is an inevitable process one needs to consider when dealing with data of the size presented in this thesis. However, these approaches can be based on both qualitative and quantitative strategies. Qualitative, as previously introduced, include selection of a month of interest or aggregating readings to some specific unit (i,e, Output Area, Postcode Sector). Quantitative include direct transformations of time series data in other summarised functional forms this can describe data using less features. For instance, if we have a stationary time series, using the average trend would be sufficient to describe the nature of the patterns. Where series are rather non stationary, alternative methods need to account for this, such as de-trending using wavelets.

Another advantage of data reduction is that the initial stage of simplifying the data helps to avoid so called 'curse of dimensionality' problem. More data variability that is included in the model is not always the best choice as more data may

imply more noise and higher likelihood of spurious relationships. It is thus may be wiser to use only systematically important features that describe variation in the data.

7.3 Preliminaries

Spectral analysis is a form of time-series analysis focused on the decomposition of the series into discrete frequencies that can represent trends in the signal. This type of analysis is one of the most popular techniques used in geo-sciences, engineering, astrophysics and generally in domains which deal with temporal data of high resolution. In essence, spectral analysis aims at arriving at some representation of time series that is formed by cosines and sines functions that further represent the periodicity/seasonality characterisation in the data. These functions form the base of the analysis as a way to account for cyclical behaviour or so-called periodic components. The core and certainly the most famous method to do such accounting is the Fourier analysis presented in the subsequent section.

7.4 Principal Component Analysis

One of the most common and perhaps computationally economic methods to reduce the dimensionality of data is via Principal Component Analysis (PCA). Originated from the work of Karl Pearson (Pearson, 1901), PCA is based on the orthogonal transformations of partially or fully correlated features into linear combination of uncorrelated variables (principal components). These principal components are defined and measured by the variance of features under consideration. The principal components are consequently represented by features that have highest variance in the dataset. Often PCA is used to define some of the most important features that can describe relationship between input and outcome variables. One important requirement for PCA to be useful is that input variables have more or less similar scale, as failing to ensure this will result in variables that are presented by larger numerical scale being preferentially selected which is not always representative of variability, especially for data of categorical nature for which ones may want to use Factor Analysis (for more details on methods available please see Comrey and Lee

(2013); Jackson (2005).

Intuitively, one may see that using PCA can be a quite common and easy to grasp approach to reduce smart meter time series data, it is nevertheless mentioned here similarly to k-means was mentioned in the chapter about clustering (Chapter 4). Both PCA and k-means are the methods that researchers often turn to due to their simplicity of application and interpretation. What is important to consider is that the relationship between variables in the data need to be carefully accessed from very profound and even theoretical point of view, before any methods applied.

While can be fairly challenging to apply to auto-correlated time series data such as smart meter data, PCA can be handy where extra variables about smart meter users are available, such as weather or socio demographic characteristics. Where no additional data is available the approaches suggested below may be considered to be more appropriate.

7.5 Fourier Transform

The Fourier Transformation is one of the most common tools used to analyse time series and signal based data. This approach also serves as baseline that has a very strong connection to other signal processing methods such as Wavelets that are applied later in this chapter. However, it will be shown that in the case of smart meter series, Fourier transformation alone may not always be appropriate as smart meter data can be characterised by presence of discontinuous behaviours and occasional spikes. This will demand a transformation that can account for such behaviour and their respective durations.

Fourier transform is one of the most popular techniques that is used to decompose time series data into frequencies that represent the signal and its development began with work of Joseph Fourier (1807) who demonstrated that a 2π periodic function can be represented as:

$$f(x) = \alpha_0 + \sum_{k=1}^{\infty} \left(\alpha_k \cos(kx) + b_k \sin(kx)\right), \tag{7.1}$$

where α_0, α_k and b_k are obtained using:

$$\alpha_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)dx\,, \quad \alpha_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\cos(kx)dx\,, \quad b_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\sin(kx)dx \tag{7.2}$$

The time dimension of the function is now lost once we perform the decomposition and the pattern is instead composed of frequencies, indexed by $k = 1, \ldots$. The transformation is based on the fact that any function of periodic nature can be transformed into a sum of sine and cosine waves (Priestley, 1996). For time series where periodicity is rather systematic, the Fourier transformation may be an appropriate choice. However, for series where periodicity changes at different parts of the function, the Fourier transformation may not be able to represent this additional information.

7.6 Wavelet Transform

After a number of years, the field of signal processing focussed on moving away from a cosine and sine representation of time series. In 1980s developed by Morlet and Grossman (see more in Grossmann and Morlet (1984); Kronland-Martinet et al. (1987)) the so-called wavelets acted to expand the Fourier method by adding dimensions of time into the count of distinct features. Unlike the Fourier method, this allows localisation in both time, and frequency (or scale). This feature is particularly useful if one is interested with data compression and time series noise removal (Graps, 1995).

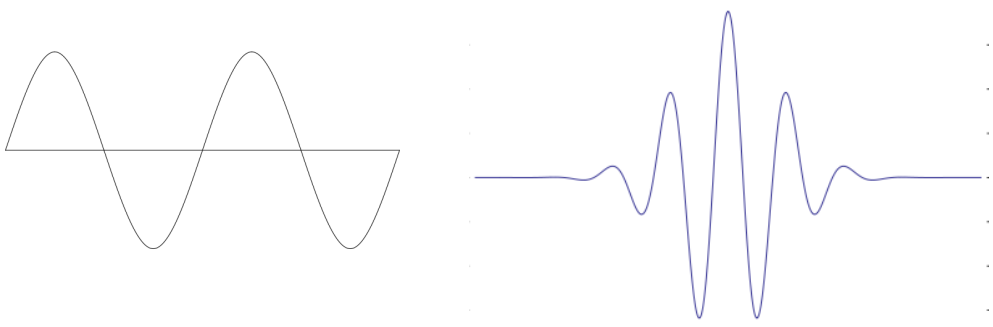


Figure 7.1: Illustration of Fourier (left) and wavelet basis functions (right)

A family of wavelet basis functions is principally described by a mother wavelet $\psi(x)$ and its scaled and shifted children (Graps, 1995):

$$\psi_{s,l}(x) = 2^{-\frac{s}{2}}\psi(2^{-s}x - l)\ . \tag{7.3}$$

In the notation above, s describes the wavelet function width, and l the wavelet position.

For this section the Discrete Wavelet Transform (DWT) will be used. Despite its use here for data reduction, the DWT can also be used as alternative similarity measure between time series (Fritz et al., 2012), particularly for indirect clustering approach that was discussed earlier in Chapter 3. The main objective of the approach that is used here is to use the wavelet coefficients to approximate the original series. Various scales of transformation can be used. If we use a so-called dyadic sampling scheme as we move towards greater scale a fewer number of coefficients are needed to represent the series (for more details please see Handcock and Gile (2010)).

7.7 Results

To begin with, the wavelet transformation is applied to smart meter data, with a random annual pattern of consumption being selected[1] . A month subsample was then selected for simplicity of visualisation. Please note that this is a subsample of annual pattern that was selected once again due to computational issues. Winter gas consumption was picked due on average being more variable and suffer or benefit from seasonal effects as was seen from the previous chapters.

The first illustration presents the raw data (Figure 7.2). It is then followed by the reconstructed series which used only distinct features of the wavelet (Figure 7.3). Lastly, the variety of wavelet transformation coefficients are presented in Figure 7.4. In a nutshell, this figures represent how the series can be transformed further such that only few features can be used to represent the pattern uniqueness.

[1] For this section, the packages 'wavelets' and 'WaveletComp' were used in R, for more details on the package and general application of methodology in R please see Nason (2010); Rösch and Schmidbauer (2014)

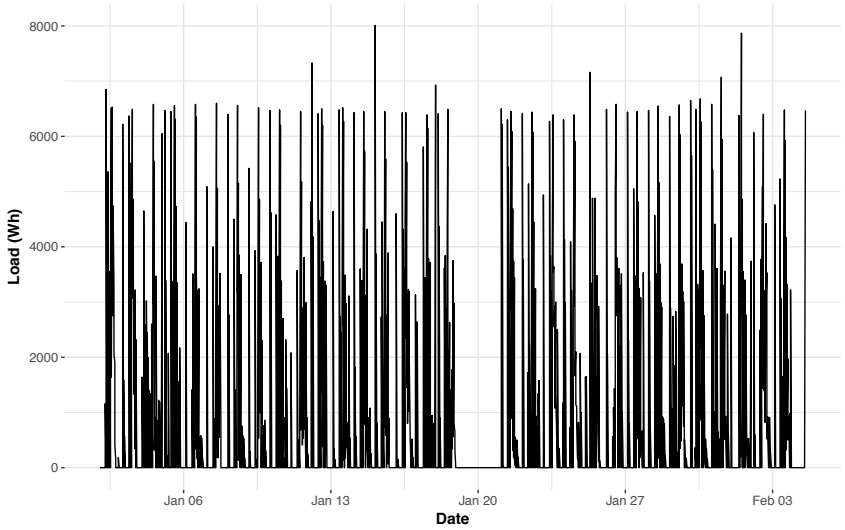


Figure 7.2: The sampled pattern that will be used for transformation. *Please note that there is identification of either absence or faulty in smart meter taking records around June 20th.*

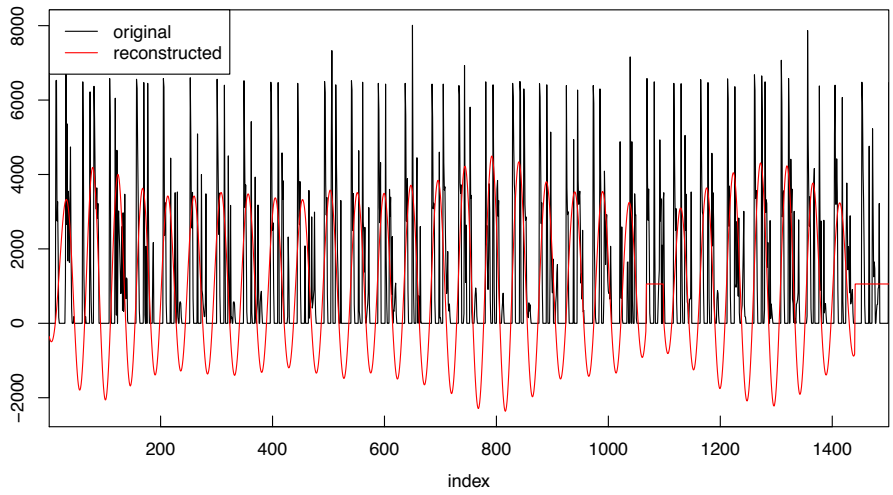


Figure 7.3: Reconstructed trend using wavelet waves

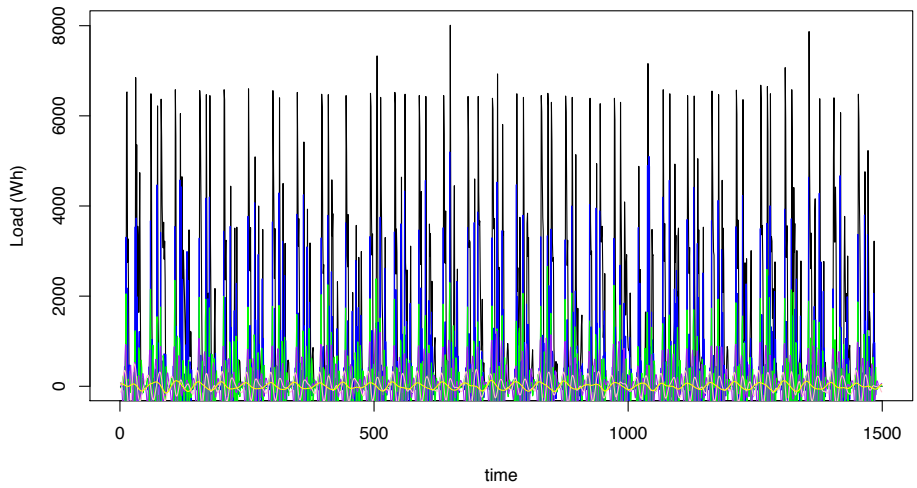


Figure 7.4: Wavelet coefficients. *Each colour represents different scale coefficients*

Images of the wavelet coefficients (Figures 7.5 and 7.6) are often referred to as scalograms, and illustrate how the signal is broken down over different scales, and how this varies over time. It provides a useful aid to understanding what has happened across the time within the environment of a single smart meter user. In these images the wavelet analysis on the given sample and the annual series are presented. The images illustrate how the variance in the time series is distributed across different time spans, otherwise known as scales. Where it is warmer (yellow and red colour palette), the contribution of these wavelet functions can be considered to be more significant, meaning that within such time spans there was a significant contribution from this scale of waveform. Note: the purple region indicates regions of scale/time space in which the wavelet cannot be fully evaluated (the size of the wavelet extends beyond the range of the data). In the case of month only sample, this significance is present with the time equivalent to a day, meaning that each day may be different from one another. They y-axes on the images represent the period conducted of half hourly readings, where 48 of such readings represents a single day.

A more interesting picture may be obtained from the wavelet analysis of annual data, where now the month tend to be an important explanatory time scale of the

variation in the data. The period represented about 21 days. Just about each three weeks. These periods may be indicative of climate and day light changes that may have impacted the energy use within the household.

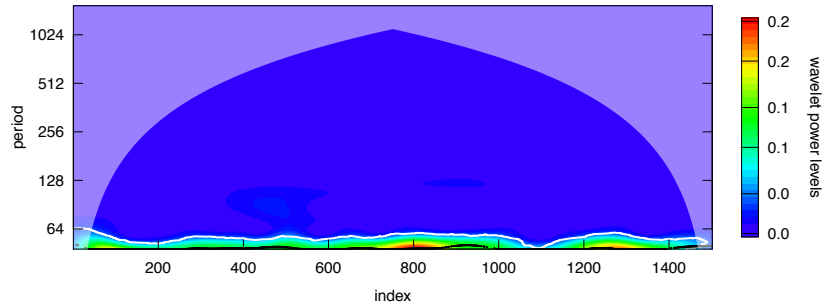


Figure 7.5: Signal significance illustration. *The y axis indicates the which period is associated with half hours periods. Most of the significance is associated with period under 64 half hour intervals which indicates just about a day*

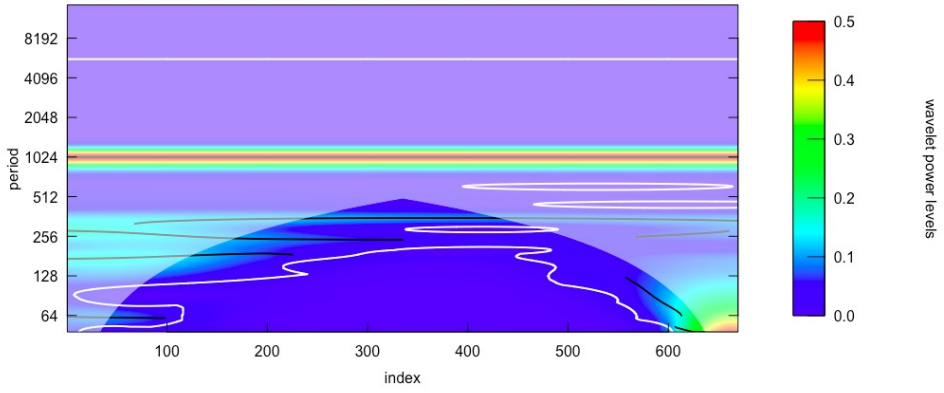


Figure 7.6: Signal significance illustration. *The y axis indicates the period which is associated with half hours periods. Most of the significance is associated with period under 1024 half hour intervals which indicates just about a month*

At a very fine temporal scale, an example of the wavelet transformation for

average daily energy consumption is presented below (Figure 7.7. Many of the wavelet scales in this setting have an average value of zero, meaning that only a few wavelet scales may explain, or describe most of the data.

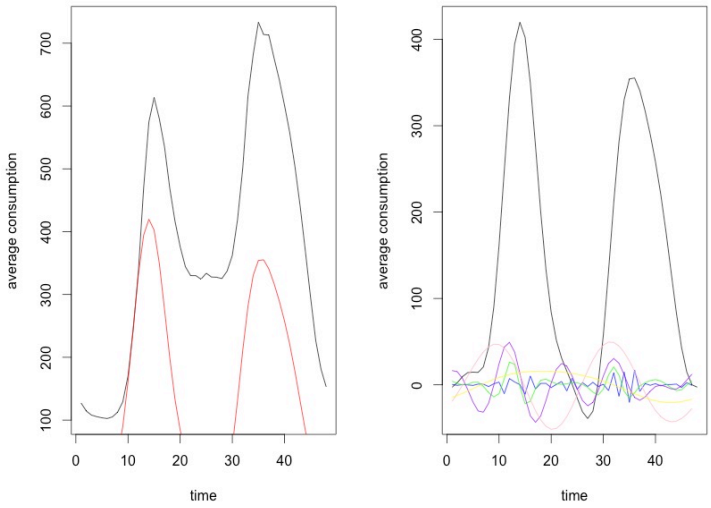


Figure 7.7: Wavelets decomposition of the average daily energy consumption temporal profile. *Figure on the left represents the real temporal profile in black and the profile recovered from wavelets transformation in red. Figure on the right represented the levels of transformation starting with the first level decomposition in black and fifth level decomposition in blue. As can be seen wavelets tend to pick the peak hours as the representative pattern of the given time series.*

7.8 Conclusions

This chapter provided a very basic attempt to transform time series data into a two-dimensional time-scale plane. The task was to see how the data can be reduced to more manageable form. For example, the wavelet analysis at a daily scale (Figure 7.7) suggests the most significant features/trends in the data can be described by only a few wavelet basis functions. Storing only these relevant coefficients, and setting the rest to zero, enables us to compress the dataset whilst retaining key information. Furthermore, presented transformations also may allow for perseverance of uniqueness of certain smart meter profiles. Given the ethical considerations of how smart meter data can be shared and utilised where available at highly granular individual level, techniques like wavelet transformations may be used to transform

the data from the raw and privacy sensitive form to more generalised information which can tell one enough about the pattern structure but may be harder to trace back to individual consumers.

Wavelet signal processing is an area that has been relatively untouched by researchers that look at smart meter data. This is can be both surprising and unsurprising, as while these methods are highly relevant for such data, they can prove more useful when large data is available, most datasets analysed in social science are remaining to be on a relatively small scale.

This chapter can also be seen as an extension to the preceding work on regression analysis that was used to describe periodicity of energy consumption. Wavelet analysis can be treated as an alternative tool to study the effects of various seasonality which can be captured in data, meaning that there is no need to impose seasonality structure by researcher as was done in the regression. Wavelet analysis helps to identify what is seasonal and what is not. As such, it allows for more in depth analysis of periodicity for each individual user or regions a whole.

Some limitations of the present analysis are associated mainly with interpretability of the results. It is often had to attribute a precise significance level to any observed structure in the wavelet coefficients. The images and visualisations presented in the chapter can at best describe what is happening after transformations have been applied. However, this can be a problem, as subjective judgements of what the results represent may drive very different implications. Thus, interpretation of wavelets representations need to be taken with care and maybe less preferred to slightly more interpretable methods such as Fourier transformations of even simple PCA.

Chapter 8

Discussion, Conclusions and Future Work

'We are drowning in information and starving for knowledge.'

- Naisbitt (2015),

This chapter presents some conclusions and final thoughts on the overall thesis. It further provides more thorough details on the potential contribution the thesis makes to research fields that are concerned with energy, smart meter data and methodology for big datasets of time series structure available to social scientists at large. The limitations of the performed work and future directions that can be taken by the research community will be presented and discussed alongside. Additionally, issues that are associated directly with the data will be discussed in order to draw attention to additional sources that can supplement large collections of smart meter data readings.

The rest of the chapter is structured as follows. Data used in the thesis is overviewed first to remind the reader of the samples, their temporal and geographical resolutions that were taken for the analysis. The difficulty of reliably performing data linkage that motivated the thesis to be focused solely on smart meter data are also covered in this section. Some solutions to these problems that can be considered by future research are presented afterwards. This is followed by the summary of research questions and findings. Contributions to the disciplines of energy and methodology in social science research in general are presented in the subsequent

sections. Further work and discussion of how the limitations in this work can be potentially addressed by future research will round up the chapter.

8.1 Data Driven Limitations

As thesis is largely driven by the nature of data, challenges associated with the data structure and sampling may be considered as the hardest to overcome and to a large extent remain unsolved. The data used in this work consisted of a National Sample of smart meter data for about 400,000 users. These data were referenced at postcode sector level and had half hour temporal granularity. The other samples that were used for analysis were the so-called Bristol sample, this was available at greater geographical resolution, Census Output Area. Finally, the Fuel Poverty Sample was available at postcode level and supplemented with data on financial vulnerability of energy customers. Chapter 3 has provided a detailed assessment of National and Bristol sample to motivate discussions of how smart meter data can be effectively visualised, what are the possible issues with using aggregated descriptive measures. It was shown that to describe smart meter data using both temporal and geographical resolution is challenging yet the combination of mapping techniques and careful selection of specific time intervals that are of interest may be helpful. Chapter 3 has presented further problems associated with various geographic resolution. These have motivated the obstacles to data linkage given the consideration of ethical issues. Indeed, the ethical use of the data may be considered to greatly restrict the depth of findings that can be derived from smart meter data.

This piece of research has shown that Big Data does not necessarily mean that the richness of insights, arriving from such data is proportional to its size. Where available, for the most part, such large data sets are actually associated with more noise and complexity which may not necessarily aid one's understanding of the issue under investigation. Big data that is reduced to small data and data of more traditional size still can be considered to serve as more insightful and profound way to look at smart meter data. So overall, one may say that big data analysis is rather a comprehensive analysis of big data chunks. Analysis on the whole datasets is

certainly challenging given the computational capacities available to research community and industry practitioners.

As was suggested throughout the thesis, most of the results remain to be rather inconclusive and bear illustrative nature. Due to limitations associated with what could be known about smart meter users, it was challenging to provide confidently the inferential conclusions about why consumption varies across the users. It is important to note, that there are number of issues and dimensions that could have been added to energy consumption to provide a clearer picture, however, these remained largely ignored at this stage. For instance, weather and climate conditions that are associated with the areas under study. Other factors which have not been analysed are energy price and the energy tariff the customer is on. Due to unavailability of the data, these effects were impossible to include in this study, yet it is important to acknowledge the significant contribution these factors could make to the variability of consumption.

8.1.1 Possible Solutions

In the Appendix of the chapter on data (Chapter 3) a number of additional datasets that can possibly be connected to smart meter data were suggested. These datasets are only those that are available openly or through administrative networks. What wasn't considered thus far, is how additional data can be possibly collected along with smart meter data to generate more beneficial outputs for analysts in both industry and academia. While there is a significant amount of research available to date that looks at smart meter records, socio demographic classifications and appliances use (Haben et al. (2016); Albert and Rajagopal (2013); Haghi and Toole (2013); Beckel et al. (2014) it would be interesting to connect the readings to household activities in more targeted way. An attempt to do that is currently underway in Oxford and lead by Philipp Grunewald under the umbrella of the project "METER" (for more details please see Grunewald et al. (2017)). The project seeks to understand how time use and energy use activities correspond to each other. By giving the participants of his study a small device where they can log their activities over the day, he then connects these data to the readings from the participant smart meter data. In

a nutshell, this approach helps to reveal what is happening in the household beyond the readings and also, what is smart meter user routine may look like and which household characteristics may influence overall activities over the day. This study, whilst currently only taking into consideration a single day per participants could be expanded for longer time periods. However, one of the challenges with such studies is self selection by participants, which often indicates greater energy use awareness and affluence. What is still missing in the current research agenda is the targeting of minority groups, less affluent energy customers and customers that may know little or nothing about how the cost of energy is reflected through their activities. An attempt to that can be seen from the fuel poverty case study presented in the thesis. Nevertheless, due to data access restrictions, there is certainly a scope to improve presented work. In the subsequent section, some ideas of how qualitative research can aid the completeness of data will be also suggested.

8.2 Research Questions and Findings

One of the main research question that was posed in the thesis is how much insights can be generated from smart meter data when available on its own, without any further data attached. Sub research areas were thus are presented by narrowing down this question to the following sub questions:

- How smart meter data can be visualised effectively such that spatial and temporal variability of energy use can be accessed?
- Given the size of the dataset, are there optimal strategies to select samples for further analysis?
- Can smart meter data be classified meaningfully? Are there national clusters of energy use that can be used to characterise the national population (Great Britain)?
- Can energy consumption be analysed and predicted using regression tools and if yes, how accurately? Are there any differences in predictability based on where smart meter user may live?

- Is it possible for fuel poor energy customers to be identified from smart meter data? If yes, to which extent this is feasible and what are the limitations?

- How helpful is geo-demographic classification such as Census OAC to explain the variability in energy use?

- How can clustering and prediction be performed more efficiently (i.e. using less computational power)?

To address the first question of classification feasibility, Chapter 4 has presented some interesting and relatively stable results using Gaussian Mixture Models. Various temporal and spatial aggregation were taken for analysis to study how suppressing individual user dynamics may change the results of segmentation. Reasons why the affects of aggregation were selected for further analysis are mainly driven by the researchers' temptation to aggregate data such that sample is reduced without considering the fact that such strategy may lead to poor results or the results that lack understanding of uniqueness of energy consumption use.

Clustering results were further explored under narrowed temporal and space scenarios to see how different time or different region may be representative of the results obtain on the whole dataset. This was rather illustrative, however, still demonstrates how different time and space resolution may be associated with further diversity of energy consumption variation. Such analysis informs the research community about how hard it is to generalise about energy use given its incredible variety and uniqueness.

The Chapter 5 had looked at various ways how energy consumption time series can be described and studies using regression analysis. Generalised Additive Models (GAMs) were introduced to approach this task. This thesis is the first attempt to use GAMs for residential energy customers load description. For this trial, a few random electricity and gas smart meter users were selected for the analysis from areas that were characterised by Census 2011 Geo demographic classification as urban or rural. It was shown that GAM models that consider interactions of weekly and daily seasonality may perform well in describing the energy use. However, one

of the main observations is that GAM do work well predominantly on small samples of readings that cover about 10-11 days. Once the samples if extended to few months, the performance of the models alters to be very poor. This is not surprising result that was observed across the thesis when using other methods as well. The uniqueness of each user consumption patterns and behaviours may not be fully explained by time only and more data need to be added to achieve better model fit.

Other types of prediction task such as customer label prediction were also studied for feasibility given the variation in the energy consumption data. While Chapter 5 offered methods to study the periodicity in energy consumption using statistical properties of the data, Chapter 6 has taken a rather subjective and non statistical outcome for prediction. Specifically, this related to label prediction for a fuel poverty indicator. This allowed the evaluation of how feasible it is to use smart meter readings to classify customers such that some of the socio-economic characteristics may be predicted from energy readings alone.

To augment the smart meter data, Census OAC geo demographic classification was used in Chapters 3 and 5. Mostly, for the Bristol sample that offers greater geographical resolution of Output Area. Whilst still inconclusive, due to an inability to validate the observed results, some associations between energy variability and the socio demographic characteristics of areas where smart meter user resides were evident throughout the thesis. Future work that may have a hold of smart meter data for all of the residential properties in the areas certainly can advance the results presented at this stage.

Most of the work nevertheless was performed on smaller samples. The methods such as Gaussian Mixture Model and Generalised Additive Model require significant computational power which currently can be challenging to obtain on an average dual core computer. It was also shown that overall models themselves tend to operate well on smaller samples as additional complexity arriving from larger sample add rather noise that confuses the derivation of the appropriate model fit. The previous chapter has suggested some ways in which data may be transformed such that uniqueness of the energy consumption patterns can be preserved but its

can be summarised with fewer data points. Wavelet transformations borrowed from signal processing literature were attempted to asses how well the wave functions can be used in the context of smart meter data. The wavelet analysis of smart meter data can serve not just as data reduction tool but also descriptive tool that can help in defining significant time spans of consumption variability. Such analysis can be performed at individual and aggregated level and allow for more in depth understanding of unique seasonal patterns embedded in the energy consumption of each smart meter user.

This thesis has assessed numerous ways for describing the patterns of energy consumption using clustering techniques, regression analysis and data reduction methods such as wavelet transformations. Whilst clustering has been performed on raw data to maintain all the unique features of energy use, future work should attempt the clustering of model parameters that can be derived from GAMs and Wavelets. This may help in segmenting larger pieces of data where preserving customers anonymity is a priority.

8.3 Contribution and Applications

There are a number of rather indirect knowledge advancements from this work, for instance, in tasks such as assessment of smart meter data, data visualisation and sample organisation. However, the central contribution of this work is rather applied and methodological. Given the data limitations discussed above, the research agenda was turned into an investigation of how smart meter data may be explored to provide fruitful insights about customer behaviour and energy use using only its temporal structure.

The thesis relied heavily the methods that are available in statistical and computer science literature, thereby no new algorithms were designed as a result of this research. However, it was discussed previously by Diamantoulakis et al. (2015) that significant research opportunities for Big Data research in the energy field lie in development of new machine learning methods that can aid development of dynamic energy management systems and contribute to more efficient management of smart

grid. In this thesis, the aim was to show what possible issues may be associated with the application of available methods and highlight the space for more technical work that can be taken by computer scientists. Before new methods are designed, it maybe important to revise what the scholarly work already have in place as methods such as Mixture (Scrucca et al. (2016)) and Generalised Additive Models (Wood, 2006; Hastie and Tibshirani, 1990) have shown to be highly appropriate for smart meter data.

A broader contribution of this work that hopefully goes beyond smart meter data analysis, was an aim to create a marriage between social science research and applied computational methods, which hopefully can be seen through the combination of technical and interpretative parts of the thesis.

Direct application for this these can be seen more easily on the premises of energy suppliers and government departments that have got a hold of smart meter data [1]. Clustering techniques seen in the thesis can be used for effective targeting of advertisements campaigns as well as for demand side management strategies such as peak hour consumption shifting. Forecasting and regression analysis of energy use can point at more predictable customers as well as rather hectic and unusual behaviours.

In summary, the thesis attempts to show how much can be explained by time on its own when it comes to smart meter user data. The conclusion and perhaps a stepping stone for further research of smart meter data is that partially, variability of energy use that is driven by time and seasonality, and is capable of segmenting and predicting energy use without any additional data, however, performance of does really vary from case to case.

8.4 Suggestion for Future Work

Smart meter data are an essential ingredient for any innovation that is interested with development and management of newborn smart cities and inhabited in them Internet of Things (IOT). Most of these recent innovations are centred around effi-

[1] The methodology and results from Chapter 4 of this thesis were used to inform clustering techniques at the data supplier premises

cient energy use and how such can be addressed with providing more home control using Artificial Intelligence technologies (Augusto and Nugent, 2006). Example of these include remote control of washing machines, dishwashers, ovens and more generally, electric heating control and efficient lighting.

Future social and economic developments and welfare certainly depends on how well governments and energy providers are managing their energy supply and use (Armaroli and Balzani, 2007). This work provides a recommendation as it informs, at least in part, which additional data may make aid inference about energy consumption variation using smart meter data, and making the resulting insights more robust.

Where extra data is available, possibilities can be way more extended. For instance, some interesting work has been performed using probabilistic dynamic spatio-temporal graphical models, mostly in disease research yet can be easily extrapolated to smart meter data that have additional data on customer characteristics. Example of this can be found in Ahmed and Xing (2009).

Data integration on different levels would certainly allow for a more complete picture of any analysis based on smart meter data. Nevertheless, one needs to be thoughtful about how data on consumers is collected to provide long term and robust insights on consumer behaviour.

Future work can certainly incorporate additional variables which may explain behaviour, for instance integrating these into time series (i.e. weather plus energy data) to see whether there may be an improvement in GAM predictions. Furthermore, methodology in this thesis may be applied way beyond smart meter data and can be considered by researchers looking at other energy recordings that have similar time resolution. These could be research projects that are interested in understanding renewables energy use and how much renewable energy may be needed to satisfy a single smart meter user energy load.

8.4.1 Matter of perspective

It is important to mention that the benefits of understanding smart meter data can be different depending on whether one adopts a supplier perspective or a customer

perspective. The regulatory pressures will drive different interest about which insights are important for energy supplier. For instance, energy cost saving will be prioritised by consumer while demand side management may be prioritised by the energy supplier.

8.4.2 Mixed methods approaches: benefits of qualitative research and data

This research, primarily based on quantitative data can only demonstrate a limited understanding of energy consumption behaviour when presented only by smart meter data. It is further limited to the natures of statistical approached used to analyse it. A thorough understanding of environments in which consumption takes place is certainly vital. For instance, two families which have different daily activities and habits though living in absolutely identical properties can be associated with greater consumption diversity. Chapter 3 touched upon this issue rather briefly, where there are a few datasets already open to research community may aid the smart meter data analysis if available at sufficient enough geographical resolution. However, what is missing is a thorough understanding of what drives certain energy use. One can endlessly link numerous datasets that hold information about property attributed and household characteristics of smart meter users but the 'elephant in the room' will not disappear till the researchers start conducting interviews with the users, studying their habits and how impact of particular lifestyles or activity patterns can be seen from smart meter data. Energy use and in particular, how inefficient it can be in the modern society, probably is described best using Jevon's paradox (Alcott et al., 2012). Whilst most often the technology around has been improved and designed to be more energy efficient. Having energy efficient bulb may often lead to more bulbs in the house. Someone who used to drive a car everyday of their life less likely to switch to the bicycle no mater how much sustainable they want this world be. Such a theory suggests more work needs to be done with consumers individually. A review of initiatives that are focused on consumers feedback and engagement was performed by Gangale et al. (2013). Furthermore, a number of studies have evidence that where customers have shown greater awareness and knowledge about

the benefits of smart meter and the information it provides, there were immediate opportunities to save on energy bills as well as contribute to the overall attempts of energy load on the grid at peak times(Faruqui et al., 2010a; Buchanan et al., 2014).

8.4.3 Infrastructure for Data Analysis

A further challenge associated with smart meter data is how one can analyse such big datasets in a short period of time without spending months on data cleaning and then another few months on application of various data analytics techniques. Despite considerations on how to design an effective data storage solutions that can preserve anonymity of individual records, there is strong need for the development of computer systems that can allow processing of such massive datasets using parallel computing. There are some solutions that are put in place and mainly targeted at business communities such as Microsoft Azure and Amazon Web Services. These data analytics infrastructures offer an access to remote computing that have an incredibly large amount of memory and speed such that big data sets as the one in this thesis can be possible analysed without a need to be split in chunks. These systems also benefit from a user friendly environment. However, the main issue is the cost of these services and also the issue of anonymity preservation. A week on Amazon Web Services using an average memory and speed system can come up to as much as £600. This is certainly may deem unaffordable for an academic research and until universities start investing into the design of similar systems to be put in place for academic use, the lack of appropriate data infrastructure will be slowing down advances in big data analytics. Large university such as University College London, have an attempt to do so by providing the researchers with access to parallel computing , known as 'Legion'. Although this can be highly useful for data available openly, given all the necessary security requirements associated with smart meter data, it may deem impossible at this stage to use such systems for this or similar type of research.

8.4.4 Conceptualisation of energy use

An interesting direction for further research is in the conceptualisation of energy use. Energy consumption can be thought of in a similar fashions as the various good and services consumption studied by economists. Reflecting on the Literature Review (Chapter 2) and work of Bernard et al. (1988); Lutzenhiser et al. (1997); Wilhite and Wilk (1987); Hackett and Lutzenhiser (1991) one may consider testing the following model of energy consumption

Final consumption = Fixed consumption + Variable consumption

This may be particularly useful in addressing fuel poverty. The hypothesis being that customers with a lower margin of variation around the Fixed consumption may be tested and investigated further. Likewise, customers whose Variable consumption tends to be stable may be investigated for sudden spikes in energy load. This could present another way of classifying energy behaviour which is more theoretical rather then reliable solely on statistical methods. Furthermore, as was seen in rather fundamental studies on understanding the factors that contribute to variation in energy consumption such as those by Raaij and Verhallen (1982); Steemers and Yun (2009); Lutzenhiser (1993), a number of variables may be included in explanatory analysis of energy consumption variation. However, challenges remain and are in desperate need of qualitative research, namely: quantifying variables such as lifestyles, energy knowledge, attitudes and social norms, values and personality (Lutzenhiser, 1993; Lutzenhiser et al., 1997).

Furthermore, there are many temporal processes which are hard to capture, for instance, habit formations, learning, and internalisation (Raaij and Verhallen, 1982). Change in household size, family transformations such as child birth, divorce or loss of a family member will also inevitably have an effect on temporal dynamics. This can be modelled separately and present a new area of energy consumption research. Figure 8.1 provides a summary of factors that can be included in such analysis simultaneously.

Another conceptualisation tool could be the use of optimisation. Energy consumption can be transformed into minimisation and maximisation problems which

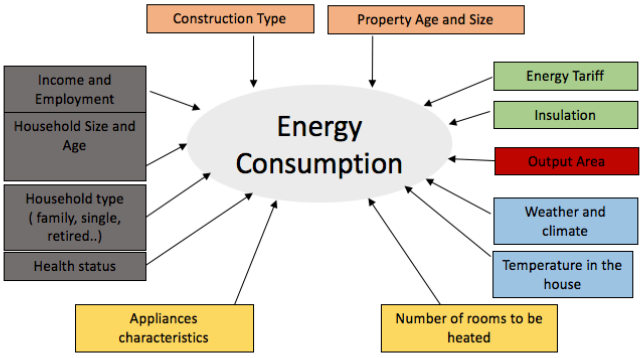


Figure 8.1: What lies behind energy consumption pattern: factors that can be tested as contributors to energy consumption variation in the conceptualised model

can be solved under budget constraints as well as constraints which are build on the necessary amount of light and heat that are sufficient for well being of smart meter user. Some examples of attempting to do so can be seen from Samadi et al. (2010).

8.4.5 Fuel poverty identification

Only a part of the thesis was dedicated to a very specific application of smart meter data. That is in the area of fuel poverty identification. It is very much debatable and certainly a very challenging area to provide a definite answers on how one can define what fuel poverty is, not to mention how it can be reduced and eliminated. Smart meter data have been rarely used in the past to predict and classify potentially fuel poor customers. Future work may consider collecting very precise data on fuel poverty. In close connection with smart meter data such data may allow for identification of customers that need support across the UK much faster that it is currently possible using energy costs models that were presented in Chapter 3.

8.5 Future Applications

There is a strong need for data integration on different levels i.e. individual , property level, neighbourhood level. Researchers need to be smarter about the data

collection on consumers to make the insights from smart meter data valuable and have forward looking advantages instead of just short term gains. So far, smart meter data was associated with significant excitement about the potential it can give for research and industry knowledge advancements. Yet, the way data is collected in the first place needs to be well thought out depending on the type of applications.

Privacy concerns which need to be prioritised in the modern age of data deluge should consequently lead to more specific data protection legislations that are targeted at smart meter users. There is so much that can be observed about a single household from the smart meter data. Yet until well protected, these data should remain to be available at perhaps similar resolution that was seen in the thesis (i.e. high granular temporally but of very low geographical resolution). Energy companies need to put systems in place which will require an extra layers of data transformations applied to raw data before it can arrive at hands of larger data analytics groups.

With respect to some industrial application of the analysis in the thesis, non quantitative customer labelling and classification remain challenging. When using classification methods such as tree methods that were presented to classify customers as potentially vulnerable with respect to energy costs, one may need to allow for cases where extra information is available to define whether a customer is financially vulnerable. This could be information arriving from direct contact with the customer, or by means of surveys/interviews. As many automated technologies that are available nowadays such as those used by Experian or retail companies providing recommendations on purchases such as Amazon or Netflix, it is obvious that the algorithm can not always be reliable. Especially with smart meter data, given the fact that energy is an essential service for human wellbeing, machine learning and AI cannot function well without human contribution, and may be harmful when it comes to the issues of individual privacy.

8.5.1 No free lunch

The *No free lunch theorem* (Wolpert and Macready, 1997) suggests that there is no single model that magically can work well for all kinds of tasks and problems, and

whilst some algorithms may work well on one sample of data it has no implication that the same approaches can behave as well on other samples. In this thesis each of the methods was carefully chosen given the nature of the data that was available. Additionally, experimenting with various sub samples of data allowed one to overview the suitability of methods for different kinds of data samples. According to the *no free lunch theorem* one of the main consequences or lessons to keep in mind when working with big data is that when different sample of data are used (or even the same data complemented with a few additional variables), one may need to re-consider which models and methods would fit best.

8.6 Closing Statement

To conclude this thesis, instead of suggesting that the solutions to any problems presented with smart meter data were revealed in this work, it may be said instead that till someone shows that there is a better and more efficient way to analyse these data, this work will serve as a useful stepping stone.

Largely, this thesis has considered a previously uncharted sample of data and unlocked the potential of smart meter data for customers classification, and regression analysis of energy use. These applications will hopefully serve useful in various domains, be it policy making or energy company customer support strategies. It is a first attempt to provide a comprehensive methodological and interpretative review of various methods that can be used to generate valuable insights from such data.

So much and so little was revealed from smart meter data on its own. However, given how much information the data holds, it is exciting to see what future research that will have more detailed information about customers, their activities, lifestyles and everyday practices, may reveal and tell about trends and behaviour of energy customers, not just in the UK but across the whole world.

Appendices

Appendices

Chapter 2: Appendix

Source	Name of the Data	Time period	Geog. Reference	Granul.	Description
DCLG	English Housing Survey	2008-2015	NA	Annual statistic	Updated each year and provides data on energy efficiency, insulation and tenure trends/does not cover entire UK, sample of around 6-7,000 houses is drawn randomly each year for investigation
UKDS and DECC	Energy Performance Certificates	2005-2012	Region	Annual statistic	Data is provided by DECC and was collected under National Energy Efficiency Data-Framework (NEED). Sample is sufficiently large and covers over 4 million households from England and Wales. Data besides, energy efficiency bands have additional variables on age, type of property, floor area, annual gas and electricity consumption as well as fuel poverty indicators.
CDRC	House Ages and Prices	1899-2015	LSOA and MSOA	ONS (quarterly) and VOA (annual)	The data was collected originally by ONS and VOA. The dwelling age counts at LSOA level alongside recorded median house prices at MSOA level.
ONS	Census 2011	2011	Census Output Area	Decennial statistic	Offers a fairly detailed description of all households and properties in the UK. Useful variables could include household size, employment characteristics, dwelling age, country of origin and others. However, the data has no consideration for recent ($<$ 10 year) temporal variations and may contain missing data.
UKDS	Understanding Society	2009-2015	LSOA and medium level Local Authority Districts	Annual statistic	Multi-dimensional household survey that re-interviews the same individuals every 24 months with the sample being over 40,000 individuals. Among other information, the detailed data on household composition, family background, employment as well as housing and payments related to subsistence are provided.
DECC	English Housing Survey: Fuel Poverty Dataset	2012	Census Output Area	Annual statistic	A detailed data set on financial circumstances of the sampled households with reference to the amount of energy that is used for different appliances (i.e. used for space and water heating); also includes the data on the eligibility for fuel poverty support schemes such as the Warm Front grant. As a limitation, it does not cover the entire UK, with the sample being limited to 12,000 households).

Table 1: *The openly available data sets that may aid the understanding of variation in energy consumption*

A selection of the data sets that are currently openly available and may contribute to the analysis of energy consumption variation in the UK are presented in the Table 1. At large, these data are limited to annual statistics, while the Census is only available as decennial. In terms of geographical references, there is a visible heterogeneity as some data sets are more granular than others. This would imply that if one is interested in linking the data sets it is necessary to first aggregate the data to different geography.

Chapter 4: Appendix

	Aggregated			Disaggregated		
Time Interval	**Mean**	**Median**	**St.Dev**	**Mean**	**Median**	**St.Dev**
1	259.80	234.92	140.08	181.96	0.00	1178.16
2	220.58	197.30	124.38	164.12	0.00	1145.76
3	199.61	177.38	118.64	159.73	0.00	1140.11
4	188.91	167.45	116.58	156.93	0.00	1143.15
5	183.96	162.23	116.29	160.52	0.00	1143.68
6	184.54	163.47	117.05	158.71	0.00	1145.71
7	187.95	166.85	119.18	166.91	0.00	1182.66
8	203.57	180.67	126.72	180.73	0.00	1184.06
9	229.44	206.03	136.46	214.08	0.00	1238.64
10	306.70	273.04	182.71	314.94	0.00	1400.71
11	421.46	381.55	226.57	467.85	0.00	1626.99
12	686.83	629.86	343.77	694.64	0.00	1913.96
13	976.60	904.71	439.61	953.85	20.00	2175.56
14	1306.30	1226.27	531.64	1146.52	70.00	2286.96
15	1510.45	1419.06	592.26	1216.66	90.00	2271.78
16	1558.24	1481.12	555.39	1474.98	90.00	2150.01
17	1421.02	1358.17	473.94	1337.45	120.00	2020.68
18	1276.81	1227.62	399.74	880.61	90.00	1857.05
19	1093.03	1057.07	328.46	766.19	22.00	1759.55
20	951.90	925.39	277.99	656.50	0.12	1696.31
21	854.64	831.74	252.07	583.86	0.00	1625.30
22	781.08	762.16	233.69	549.54	0.00	1594.56
23	741.01	723.04	224.81	552.25	0.00	1598.88
24	741.73	722.33	227.28	551.11	0.00	1608.77
25	722.07	704.34	218.00	563.18	0.00	1634.49
26	750.96	728.97	230.48	553.34	0.00	1622.94
27	728.22	710.34	217.77	559.58	0.00	1644.10
28	220.58	715.41	212.91	164.12	0.00	1639.50
29	737.52	177.38	210.52	602.01	0.00	1681.72
30	787.08	774.86	116.58	643.80	0.00	1143.15
31	850.69	839.90	232.32	764.25	10.00	1828.98
32	1046.35	1029.30	289.77	930.65	67.00	2006.49
33	1234.72	1214.22	330.11	1171.51	100.00	2204.06
34	1472.50	1441.87	394.22	1285.21	150.00	2252.11
35	1603.22	1563.55	415.40	1369.46	200.00	2299.62
36	1698.90	1654.65	436.31	1314.84	191.00	2195.42
37	1644.06	1603.94	409.87	1316.56	177.00	2204.19
38	1644.44	1603.13	407.36	1241.10	138.50	2147.75
39	1565.43	1526.71	391.76	1176.25	112.00	2079.76
40	1492.84	1455.79	383.25	1083.67	90.00	2007.02
41	1390.42	1349.09	369.52	994.66	89.00	1952.78
42	1278.60	1236.79	353.70	1197.23	67.00	1893.60
43	1144.64	1100.02	334.37	1018.55	90.00	1829.42
44	977.39	929.89	304.52	630.75	23.00	1562.56
45	797.32	753.50	267.70	483.32	0.00	1403.52
46	605.45	565.37	228.64	350.38	0.00	1536.12
47	431.50	398.29	187.90	278.56	0.00	1482.01
48	320.94	293.57	157.03	245.98	0.00	1502.46

Figure 2: Descriptive statistics for the samples we used for the experiments

Chapter 5:Appendix

As an extension to the analysis in Chapter 5, a brief analysis for 'Ageing' socio demographic group is presented below.

	Gas		
	Mean (half hour)	St Dev (half hour)	N
Ageing Rural Resident (annual)	901.92 wh	1686.54Wh	14833
Ageing Rural Resident (Jan-Mar)	1685.70Wh	2241.47Wh	3374
Ageing Rural Resident (May-Aug)	239.67Wh	581.65 Wh	3605
Ageing Urban Resident (annual)	508.89Wh	1194.86Wh	15169
Ageing Urban Resident (Jan-Mar)	1052.77Wh	1717.00Wh	3374
Ageing Urban Resident (May-Aug)	92.01Wh	284.13Wh	3605

Table 2: Descriptive Statistics for Gas Samples, "Ageing" group

Ageing

This section considers a very similar comparison as was seen in Chapter 5 yet for a smart meter users that reside within the areas that are characterised as having large proportion of ageing population. Urban and rural ageing populations areas were selected for comparison. As will be shown in fact the urban and rural ageing population can be described as distinct in quite similar fashion as the experiments in the chapter.

Rural Ageing

Sampled smart meter user from 'Rural Ageing' area can be characterised by quite persistent consumption throughout the day with partial periodicity in terms of the consumption peaks, yet with way more variability in between as the presented user certainly consumes quite continuously throughout the day (Figure 3). The model fit measured by R squared is about 0.81 for winter and about 0.61 for summer. The model performance can be considered as relatively good in explaining overall variability of this smart meter user's consumption. This is contrary to rural customer that was studied earlier, meaning that there may be more periodicity expected for smart meter user that reside in the urban area with pre-dominantly ageing population.

Summer consumption can be characterised as double peaked yet with almost

no or little consumption in between the peaks. This may be explained by absence of heating which can be more evident during the winter seasons. The period of possible physical absence can also be noticed from the summer pattern.

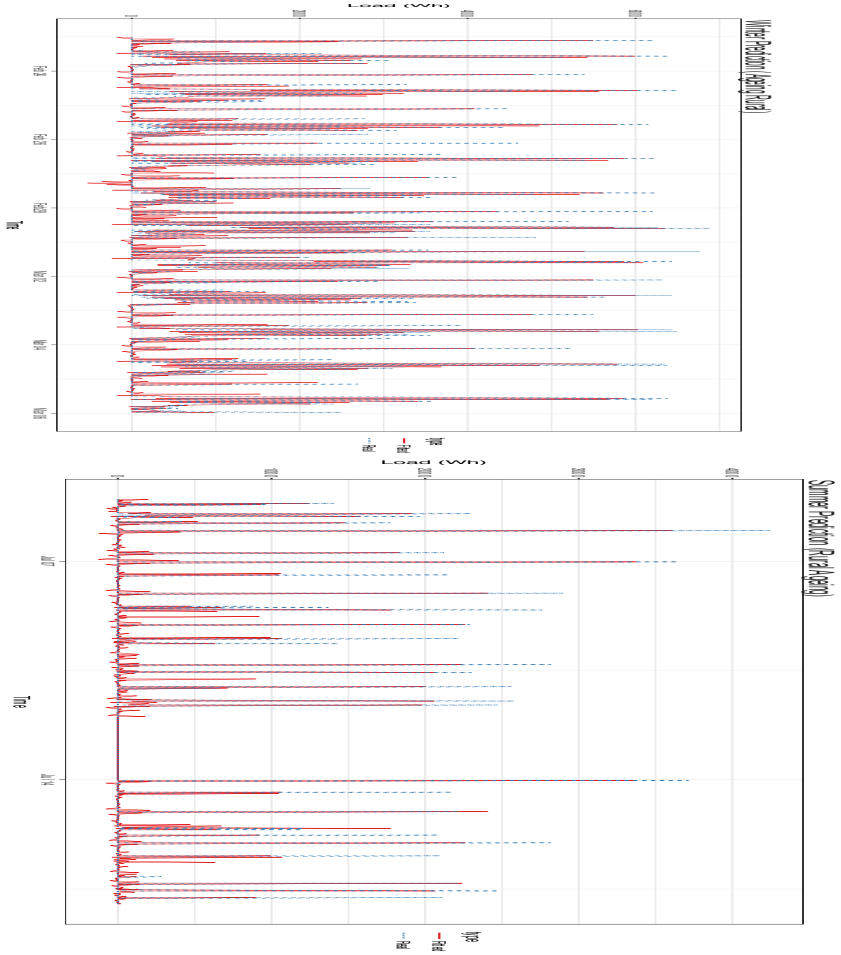


Figure 3: GAM fit for a customer that belongs to OA characterised as 'Rural Ageing'

The breakdown of the model fit by hours of the day and days of the week are presented in Figure 4. For winter, it may be noted that consumption during the first half of the day tend to be greater in magnitude compared to that of evening. In summary, consumption look persistent in its trend for each day of the week but can be described by decreasing consumption across the day.

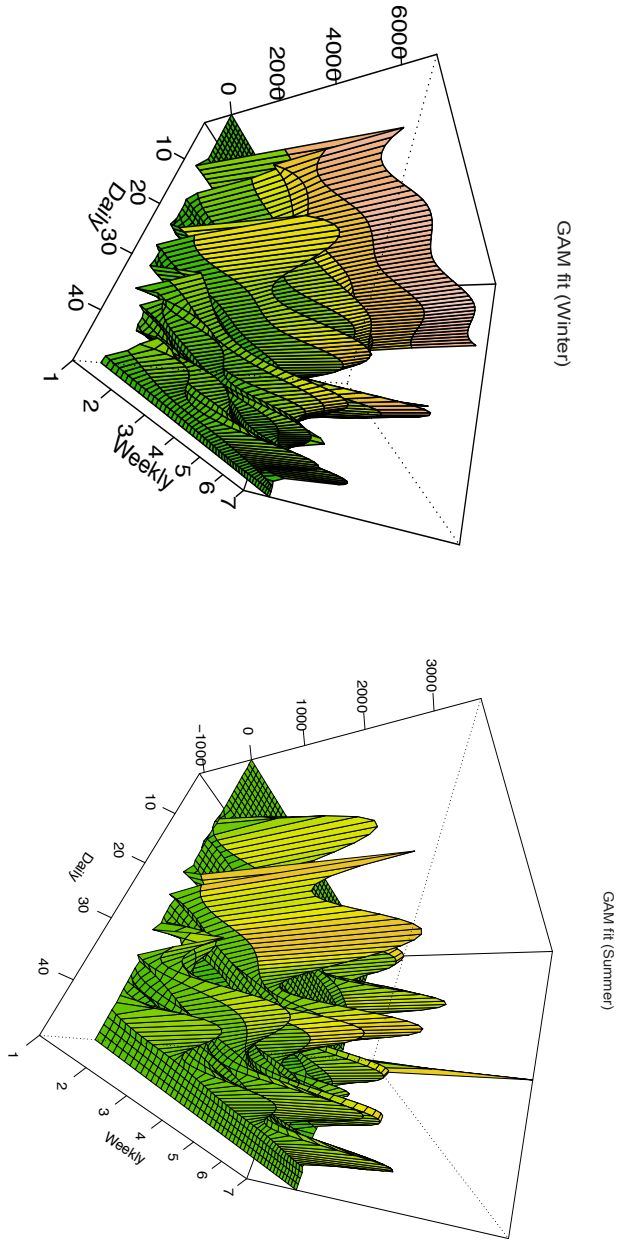


Figure 4: GAM fit for a customer that belongs to OA characterised as 'Rural Ageing' in 3D.

In the case of the summer, consumption levels appear rather hectic. There seem to be more consumption first days of the week (Monday to Wednesday). Overall, there are more than just two peaks during the days. This makes it slightly challenging to conclude which kind of activities may be present in this type of property.

Urban Ageing

Sampled smart meter user from the urban area characterised by large number of ageing population may remind the reader of the 'Urban Professional' case that was studied previously. Figure 5illustrates winter and summer trend and predictions. Extremely high periodicity of the behaviour can be noted for both winter and summer. The model fit for winter measured by R squared is 0.92. Summer however can be described quite poorly (only about 0.28). The summer poor fit may be described by inconsistency of the peaks magnitude across the sampled time.

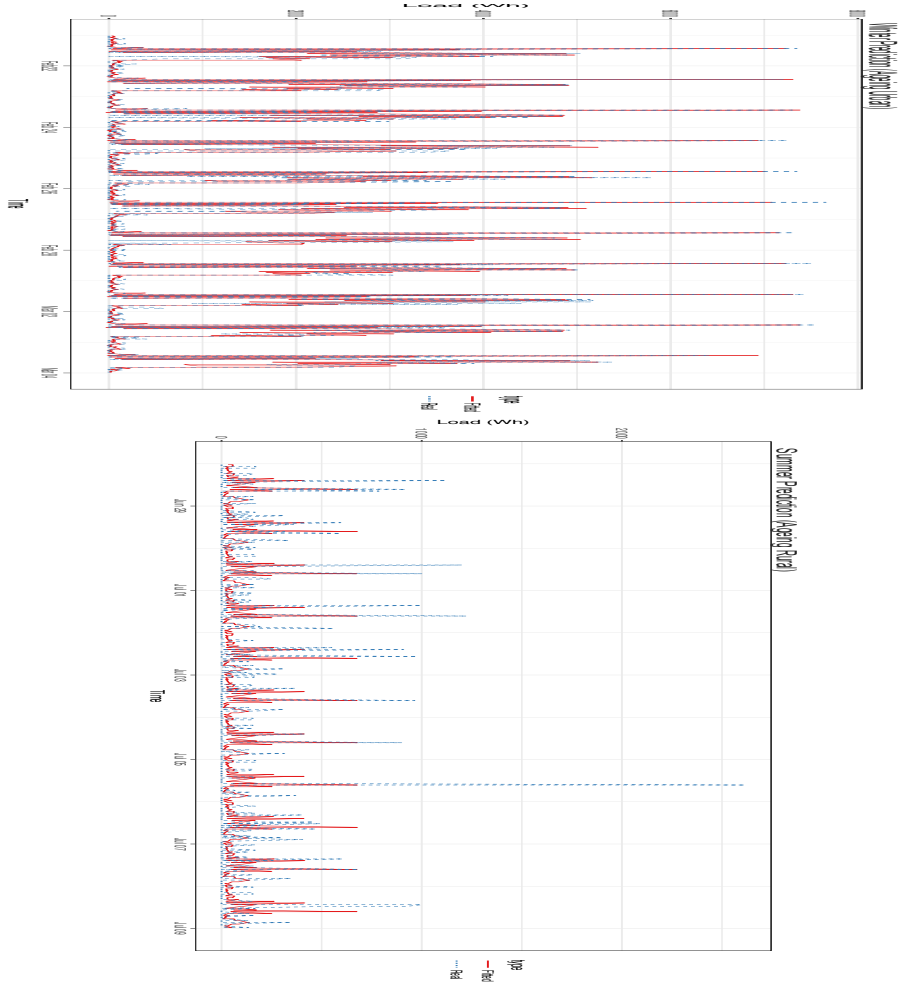


Figure 5: GAM fit for a customer that belongs to OA characterised as 'Urban Ageing'

The 3D visualisations that illustrate in more details the consumption differences across the time during the day and across the days of the week are given in Figure 6. From winter fit, one may note quite sharp consumption levels with clearly

defined morning and evening peaks that are persistent regardless of the away of the week. Very similar picture is observed for the summer with slightly shorter consumption timeframe for the morning and evening peaks but nevertheless clearly defined and present across all the days of the week.

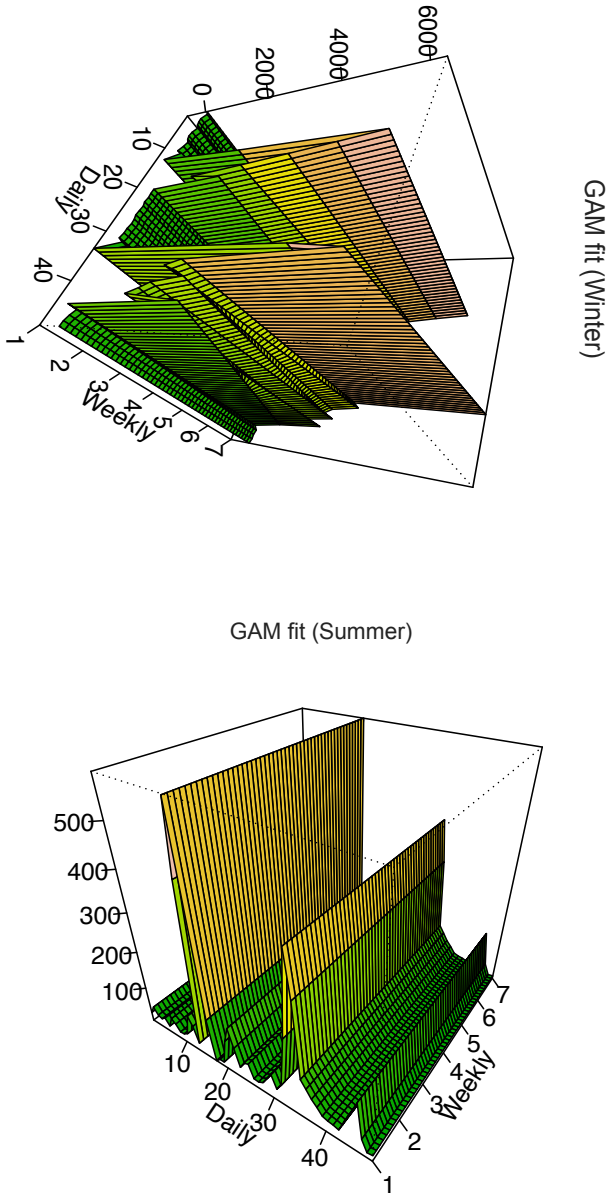


Figure 6: GAM fit for a customer that belongs to OA characterised as 'Urban Ageing' in 3D.

In summary, it can be noted that there are certainly well defined differences

in terms of periodicity and predictability of energy consumption patterns between urban and rural group. Narrowing down analysis not just for urban and rural but further to ageing population that reside in these distinct areas have demonstrated that on average ageing smart meter users tend to be more persistent in the consumption behaviour that does not alternate across the days of the week. These customers are also more likely to be present at the property.

Bibliography

A. Ahmed and E. P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

H. N. Akouemo and R. J. Povinelli. Data improving in time series using arx and ann models. *IEEE Transactions on Power Systems*, 32(5):3352–3359, 2017.

A. Albert and R. Rajagopal. Smart meter driven segmentation: What your consumption says about you. *IEEE Transactions on Power Systems*, 28(4):4019–4030, 2013.

B. Alcott, M. Giampietro, K. Mayumi, and J. Polimeni. *The Jevons paradox and the myth of resource efficiency improvements*. Routledge, 2012.

D. A. Alhadeff. *Microeconomics and human behavior: Toward a new synthesis of economics and psychology*. Univ of California Press, 1982.

R. K. Andadari, P. Mulder, and P. Rietveld. Energy poverty reduction by fuel switching. impact evaluation of the {LPG} conversion program in indonesia. *Energy Policy*, 66:436 – 449, 2014. ISSN 0301-4215. doi: http://dx.doi.org/10.1016/j.enpol.2013.11.021.

B. Anderson, S. Lin, A. Newing, A. Bahaj, and P. James. Electricity consumption and household characteristics: Implications for census-taking in a smart metered future. *Computers, Environment and Urban Systems*, 63:58–67, 2017.

C. Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

R. Anderson and S. Fuloria. Who controls the off switch? In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 96–101. IEEE, 2010.

N. Armaroli and V. Balzani. The future of energy supply: challenges and opportunities. *Angewandte Chemie International Edition*, 46(1-2):52–66, 2007.

J. C. Augusto and C. D. Nugent. *Designing smart homes: the role of artificial intelligence*, volume 4008. Springer, 2006.

T. J. Barnes and M. W. Wilson. Big data, social physics, and spatial analysis: The early years. *Big Data & Society*, 1(1):2053951714535365, 2014.

C. Beckel, L. Sadamori, T. Staake, and S. Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, 2014.

M. J. Bernard, J. McBride, and D. J. Desmond. Events – the third variable in daily household energy consumption. *Working Paper*, 1988.

B. Boardman. Energy efficiency and fuel poverty. In *PRASEG Annual Conference*, volume 5, 1998.

B. Boardman. *Fixing Fuel Poverty: Challenges and Solutions*. Earthscan, 2010.

J.-M. Bohli, C. Sorge, and O. Ugus. A privacy model for smart metering. In *Communications Workshops (ICC), 2010 IEEE International Conference on*, pages 1–5. IEEE, 2010.

R. N. Bolton. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing science*, 17(1): 45–65, 1998.

P. Bourdieu. *Distinction: A social critique of the judgement of taste*. Harvard University Press, 1984.

S. Bouzarovski, S. Petrova, and S. Tirado-Herrero. From fuel poverty to energy vulnerability: The importance of services, needs and practices. Technical report, SPRU-Science and Technology Policy Research, University of Sussex, 2014.

G. Box and D. Cox. An analysis of transformations revisited, rebutted. Technical report, WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER, 1981.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

D. Boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

L. Breiman. Fitting additive models to regression data. *Computational statistics and data analysis*, 15:13–46, 1993.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001a. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001b.

K. Buchanan, R. Russo, and B. Anderson. Feeding back about eco-feedback: How do consumers use and respond to energy monitors? *Energy Policy*, 73:138 – 146, 2014. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2014.05.008.

K. P. Burnham and D. R. Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

H.-A. Cao, C. Beckel, and T. Staake. Are domestic load profiles stable over time? an attempt to identify target households for demand side management campaigns. In *Industrial Electronics Society, IECON 2013 - 39th Annual Conference of the IEEE*, pages 4733–4738. IEEE, 2013.

R. Caruana and A. Niculescu-mizil. An empirical comparison of supervised learning algorithms. In *In Proc. 23 rd Intl. Conf. Machine learning (ICML 2006*, pages 161–168, 2006.

A. Cavoukian, A. Fisher, S. Killen, and D. A. Hoffman. Remote home health care technologies: how to ensure privacy? build it in: Privacy by design. *Identity in the Information Society*, 3(2):363–378, 2010.

C. Chatfield, J. Zidek, and J. Lindsey. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2010.

C. P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

J. Cherfas. Skeptics and visionaries examine energy saving. *Science*, 251:154–156, 1991.

G. Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68–80, 2012.

A. L. Comrey and H. B. Lee. *A first course in factor analysis*. Psychology Press, 2013.

N. Cressie. Spatial prediction and ordinary kriging. *Mathematical geology*, 20(4): 405–421, 1988.

S. L. Cutter, B. J. Boruff, and W. L. Shirley. Social vulnerability to environmental hazards. *Social science quarterly*, 84(2):242–261, 2003.

DCLG. English housing survey headline report 2014-15. 2015.

DECC. Housing energy fact file 2012: energy use in homes. 2012.

DECC. Annual report on fuel poverty statistics 2013. 2013.

P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis. Big data analytics for dynamic energy management in smart grids. *Big Data Research*, 2(3):94 – 101, 2015. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2015.03.003. Big Data, Analytics, and High-Performance Computing.

D. A. Dickey and W. A. Fuller. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: Journal of the Econometric Society*, pages 1057–1072, 1981.

M. Dodge and R. Kitchin. Codes of life: Identification codes and the machine-readable world. *Environment and Planning D: Society and Space*, 23(6):851–881, 2005.

B. Dong, C. Cao, and S. E. Lee. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37(5): 545–553, 2005.

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

A. Druckman, M. Chitnis, S. Sorrell, and T. Jackson. Missing carbon reductions? exploring rebound and backfire effects in uk households. *Energy Policy*, 39(6): 3572–3581, 2011.

W. Fan and A. Bifet. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, 14(2):1–5, 2013.

A. Faruqui, S. Sergici, and A. Sharif. The impact of informational feedback on energy consumptiona survey of the experimental evidence. *Energy*, 35(4):1598 – 1608, 2010a. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2009.07.042. Demand Response Resources: the US and International Experience.

A. Faruqui, S. Sergici, and A. Sharif. The impact of informational feedback on

energy consumptiona survey of the experimental evidence. *Energy*, 35(4):1598–1608, 2010b.

D.-W.-I. C. Flath, D.-W.-I. D. Nicolay, T. Conte, P. D. C. van Dinther, and L. Filipova-Neumann. Cluster analysis of smart metering data. *Business & Information Systems Engineering*, 4(1):31–39, 2012.

V. Foster, J.-P. Tre, and Q. Wodon. Energy prices, energy efficiency, and fuel poverty. *Unpublished paper. Latin America and Caribbean Regional Studies Program, Washington, DC: The World Bank*, 2000.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001a.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001b.

J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

H. Fritz, L. A. Garcıa-Escudero, and A. Mayo-Iscar. tclust: An r package for a trimming approach to cluster analysis. *Journal of Statistical Software*, 47(12): 1–26, 2012.

F. Gangale, A. Mengolini, and I. Onyeji. Consumer engagement: An insight from smart grid projects in europe. *Energy Policy*, 60:621 – 628, 2013. ISSN 0301-4215. doi: https://doi.org/10.1016/j.enpol.2013.05.031.

R. C. Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.

C. G. Gibson. *Singular points of smooth mappings*, volume 25. Pitman publishing, 1979.

P. B. Goes. Editor's comments: big data and is research. *Mis Quarterly*, 38(3): iii–viii, 2014.

S. Goodwin. Visualisation for household energy analysis: techniques for exploring multiple variables across scale and geography. 2015.

R. Granell, C. Axon, and D. Wallom. Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. 30:1–8, 11 2015.

A. Graps. An introduction to wavelets. *IEEE computational science and engineering*, 2(2):50–61, 1995.

A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis*, 15 (4):723–736, 1984.

P. Grunewald, M. Diakonova, D. Zilli, J. Bernard, and A. Matousek. What we do matters–a time-use app to capture energy relevant activities. 2017.

O. Guerra-Santin and L. Itard. Occupants' behaviour: determinants and effects on residential heating consumption. *Building Research & Information*, 38(3):318–338, 2010.

S. Haben, M. Rowe, D. V. Greetham, P. Grindrod, W. Holderbaum, B. Potter, and C. Singleton. Mathematical solutions for electricity networks in a low carbon future. 2013.

S. Haben, C. Singleton, and P. Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE transactions on smart grid*, 7(1):136–144, 2016.

B. Hackett and L. Lutzenhiser. Social structures and economic conduct: interpreting variations in household energy consumption. In *Sociological forum*, volume 6, pages 449–470. Springer, 1991.

A. Haghi and O. Toole. The use of smart meter data to forecast electricity demand,. *CS229 Course paper*, 2013.

D. J. Hand. big dataand data sharing. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(3):629–631, 2016.

M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5, 2010.

T. Hargreaves, M. Nye, and J. Burgess. Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors. *Energy policy*, 38(10):6111–6119, 2010.

T. J. Hastie and R. J. Tibshirani. Generalized additive models, volume 43 of monographs on statistics and applied probability, 1990.

J. Haworth. *Spatio-temporal forecasting of network data*. PhD thesis, UCL (University College London), 2014.

R. Herrnstein. A behavioral alternative to utility maximization. *Applied behavioral economics*, 1, 1988.

J. Hills. Getting the measure of fuel poverty: Final report of the fuel poverty review. *Department of Energy and Climate Change (DECC)*, CASE report 72, 2012a.

J. Hills. Getting the measure of fuel poverty: Final report of the fuel poverty review. 2012b.

P. Howden-Chapman, H. Viggers, R. Chapman, K. O?Sullivan, L. T. Barnard, and B. Lloyd. Tackling cold housing and fuel poverty in new zealand: A review of policies, research, and health impacts. *Energy Policy*, 49:134 – 142, 2012. ISSN 0301-4215. doi: http://dx.doi.org/10.1016/j.enpol.2011.09.044. Special Section: Fuel Poverty Comes of Age: Commemorating 21 Years of Research and Policy.

C. Hsiao, D. C. Mountain, and K. H. Illman. A bayesian integration of end-use metering and conditional-demand analysis. *Journal of Business & Economic Statistics*, 13(3):315–326, 1995.

G. M. Huebner, I. Hamilton, Z. Chalabi, D. Shipworth, and T. Oreszczyn. *Applied Energy*, 159:589 – 600, 2015. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2015.09.028.

J. E. Jackson. *A user's guide to principal components*, volume 587. John Wiley & Sons, 2005.

A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

A. K. Jain and R. C. Dubes. Algorithms for clustering data. 1988.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

A. Jarrah Nezhad, T. K. Wijaya, M. Vasirani, and K. Aberer. Smartd: smart meter data analytics dashboard. In *Proceedings of the 5th international conference on Future energy systems*, pages 213–214. ACM, 2014.

S. Johnston, A. Lee, and H. McGregor. Engineering as captive discourse. *Techné: Research in Philosophy and Technology*, 1(3/4):128–136, 1996.

C. M. Judd, G. H. McClelland, and C. S. Ryan. *Data analysis: A model comparison approach*. Routledge, 2011.

J. Kandt. The social and spatial context of urban health inequalities: towards an interpretive geodemographic framework. 2015.

A. Kavousian, R. Rajagopal, and M. Fischer. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, 55:184–194, 2013.

R. Kitchin. Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, 3(3):262–267, 2013.

R. Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

M. Koivisto, P. Heine, I. Mellin, and M. Lehtonen. Clustering of connection points and load modeling in distribution systems. *IEEE Transactions on Power Systems*, 28(2):1255–1265, 2013.

R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of sound patterns through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(02):273–302, 1987.

M. Kuhn and K. Johnson. *Applied predictive modeling*. Springer, 2013.

J. Kwac, J. Flora, and R. Rajagopal. Household energy consumption segmentation using hourly data. *Smart Grid, IEEE Transactions on*, 5(1):420–430, 2014.

D. Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70), 2001.

P. Laurinec. Doing magic and analysing seasonal time series with gam (generalised additive model) in r, howpublished = https://petolau.github.io/Analyzing-double-seasonal-time-series-with-GAM-in-R/ , 2016. Accessed: 2018-05-10.

P. Laurinec and M. Lucká. Comparison of representations of time series for clustering smart meter data. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1, 2016.

J. Lee, Y.-c. Kim, and G.-L. Park. An analysis of smart meter readings using artificial neural networks. *Convergence and Hybrid Information Technology*, pages 182–188, 2012.

B. Legendre and O. Ricci. Measuring fuel poverty in france: Which households are the most fuel vulnerable? *Energy Economics*, 49:620–628, 2015.

J. I. Lerner and D. K. Mulligan. Taking the long view on the fourth amendment: Stored records and the sanctity of the home. *Stan. Tech. L. Rev.*, page 3, 2008.

X. Li, C. P. Bowers, and T. Schnier. Classification of energy consumption in buildings with outlier detection. *IEEE Transactions on Industrial Electronics*, 57(11): 3639–3644, 2010.

T. W. Liao. Clustering of time series data-a survey. *Pattern recognition*, 38(11): 1857–1874, 2005.

J. Lines, A. Bagnall, P. Caiger-Smith, and S. Anderson. *Classification of Household Devices by Electricity Usage Profiles*, pages 403–412. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-23878-9. doi: 10.1007/978-3-642-23878-9_48.

M. A. Lisovich, D. K. Mulligan, and S. B. Wicker. Inferring personal information from demand-response systems. *IEEE Security & Privacy*, 8(1), 2010.

P. Longley, J. Cheshire, and A. Singleton. *Consumer Data Research*. UCL Press, 2018.

J. LUO, T. HONG, and M. YUE. Real-time anomaly detection for very short-term load forecasting. *Journal of Modern Power Systems and Clean Energy*, 6(2): 235–243, Mar 2018. ISSN 2196-5420. doi: 10.1007/s40565-017-0351-7.

L. Lutzenhiser. A cultural model of household energy consumption. *Energy*, 17(1): 47–60, 1992.

L. Lutzenhiser. Social and behavioral aspects of energy use. *Annual review of Energy and the Environment*, 18(1):247–289, 1993.

L. Lutzenhiser et al. Social structure, culture, and technology: Modeling the driving forces of household energy consumption. *Environmentally significant consumption: Research directions*, 129, 1997.

V. Mayer-Schönberger and K. Cukier. Big data–a revolution that will transform how we live, think and work, 2013.

P. McDaniel and S. McLaughlin. Security and privacy challenges in the smart grid. *IEEE Security & Privacy*, 7(3), 2009.

B. McDonald, P. Pudney, and J. Rong. Pattern recognition and segmentation of smart meter data. *ANZIAM Journal*, 54:105–150, 2014.

E. McKenna, I. Richardson, and M. Thomson. Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy*, 41:807–814, 2012.

F. McLoughlin, A. Duffy, and M. Conlon. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An irish case study. *Energy and Buildings*, 48:240–248, 2012a.

F. McLoughlin, A. Duffy, and M. Conlon. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An irish case study. *Energy and Buildings*, 48:240–248, 2012b.

H. Meier and K. Rehdanz. Determinants of residential space heating expenditures in great britain. *Energy Economics*, 32(5):949–959, 2010.

Met Office. Met Office: Climate Summaries. `http://www.metoffice.gov.uk/climate/uk/summaries`, 2015. [Accessed: 2015-09-12].

L. Middlemiss and R. Gillard. How can you live like that?: energy vulnerability and the dynamic experience of fuel poverty in the uk. 2014.

M. Mohandes, T. Halawani, S. Rehman, and A. A. Hussain. Support vector machines for wind speed prediction. *Renewable Energy*, 29(6):939–947, 2004.

C. S. Möller-Levet, F. Klawonn, K.-H. Cho, and O. Wolkenhauer. Fuzzy clustering of short time-series and unevenly distributed sampling points. In *International Symposium on Intelligent Data Analysis*, pages 330–340. Springer, 2003.

R. Moore. Definitions of fuel poverty: Implications for policy. *Energy Policy*, 49: 19–26, 2012.

P. Moran. A test for serial correlation of residuals. *Biometrica*, 37:178–181, 1950.

D. Muchlinski, D. Siroky, J. He, and M. Kocher. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87–103, 2015.

J. Naisbitt. Megatrends (1982). *Ten New Directions Transforming Our Lives*, 2015.

G. Nason. *Wavelet methods in statistics with R*. Springer Science & Business Media, 2010.

G. P. Nason and R. Von Sachs. Wavelets in time-series analysis. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 357(1760):2511–2526, 1999.

A. Newing, B. Anderson, A. Bahaj, and P. James. The role of digital trace data in supporting the collection of population statistics–the case for smart metered electricity consumption data. *Population, Space and Place*, 22(8):849–863, 2016.

T. A. Nguyen and M. Aiello. Energy intelligent buildings based on user activity: A survey. *Energy and buildings*, 56:244–257, 2013.

S. J. Nizami and A. Z. Al-Garni. Forecasting electric energy consumption using neural networks. *Energy Policy*, 23(12):1097 – 1104, 1995. ISSN 0301-4215. doi: http://dx.doi.org/10.1016/0301-4215(95)00116-6.

T. Oates. Identifying distinctive subsequences in multivariate time series by clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 322–326. ACM, 1999.

OFGEM. Retail energy markets in 2015. 2015.

S. Okushima. Measuring energy poverty in japan, 2004?2013. *Energy Policy*, 98: 557 – 564, 2016. ISSN 0301-4215. doi: http://dx.doi.org/10.1016/j.enpol.2016.09.005.

K. C. O'Sullivan, P. L. Howden-Chapman, and G. M. Fougere. Fuel poverty, policy, and equity in new zealand: the promise of prepayment metering. *Energy Research & Social Science*, 7:99–107, 2015.

K. Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno. Handling bad or missing smart meter data through advanced data imputation. In *Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society*, pages 1–5. IEEE, 2016.

S. Piantadosi, D. P. Byar, and S. B. Green. The ecological fallacy. *American journal of epidemiology*, 127(5):893–904, 1988.

M. Priestley. Wavelets and time-dependent spectral analysis. *Journal of Time Series Analysis*, 17(1):85–103, 1996.

E. L. Quinn. Smart metering and privacy: Existing laws and competing policies. 2009.

W. v. Raaij and T. Verhallen. Patterns of residential energy behavior. *Journal of Economic Psychology*, 4, 1982.

A. C. Rencher and G. B. Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.

I. Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

D. Roberts, E. Vera-Toscano, and E. Phimister. Fuel poverty in the uk: Is there a difference between rural and urban areas? *Energy policy*, 87:216–223, 2015.

S. Roberts. Energy, equity and the future of the fuel poor. *Energy Policy*, 36(12): 4471–4474, 2008.

A. Rösch and H. Schmidbauer. Waveletcomp 1.1: A guided tour through the r package. 2014.

J. Rosenow, R. Platt, and B. Flanagan. Fuel poverty and energy efficiency obligations–a critical assessment of the supplier obligation in the uk. *Energy Policy*, 62:1194–1203, 2013.

J. L. M. Saboia. Autoregressive integrated moving average (arima) models for birth forecasting. *Journal of the American Statistical Association*, 72(358):264–270, 1977.

P. Samadi, A.-H. Mohsenian-Rad, R. Schober, V. W. Wong, and J. Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 415–420. IEEE, 2010.

I. B. Sánchez, I. D. Espinós, L. M. Sarrión, A. Q. López, and I. N. Burgos. Clients segmentation according to their domestic energy consumption by the use of self-organizing maps. In *Energy Market, 2009. EMM. 6th International Conference on the European*, pages 1–6. IEEE, 2009.

M. Santamouris, K. Pavlou, A. Synnefa, K. Niachou, and D. Kolokotsa. Recent progress on passive cooling techniques: Advanced technological developments to improve survivability levels in low-income households. *Energy and Buildings*, 39(7):859–866, 2007.

S. Schmidt and H. Weigt. Interdisciplinary energy research and energy consumption: What, why, and how? *Energy Research & Social Science*, 10:206–219, 2015.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289, 2016.

F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

T. Sefton. Targeting fuel poverty in england: is the government getting warm? *Fiscal Studies*, 23(3):369–399, 2002.

S. M. Shellman. Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis*, 12(1):97–104, 2004.

R. Silipo and P. Winters. Big data , smart energy , and predictive analytics time series prediction of smart energy data,. Technical report, KNIME, 2013.

J. Skea. The renaissance of energy innovation. *Energy & Environmental Science*, 7 (1):21–24, 2014.

B. K. Sovacool. Conceptualizing urban household energy use: climbing the "energy services ladder". *Energy Policy*, 39(3):1659–1668, March 2011.

K. Steemers and G. Y. Yun. Household energy consumption: a study of the role of occupants. *Building Research & Information*, 37(5-6):625–637, 2009.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307, 2008.

P. Subrahmanyam, D. Wagner, D. Mulligan, E. Jones, U. Shankar, and J. Lerner. Network security architecture for demand response/sensor networks. *Consultant Report to California Energy Commission*, 2005.

L. G. Swan and V. I. Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and sustainable energy reviews*, 13(8):1819–1835, 2009.

S. B. Taieb, R. Huser, R. J. Hyndman, and M. G. Genton. Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. *IEEE Transactions on Smart Grid*, 7(5):2448–2455, 2016.

L. Taylor, R. Schroeder, and E. Meyer. Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, 1(2):2053951714536877, 2014.

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

G. K. Tso and K. K. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761 – 1768, 2007. ISSN 0360-5442. doi: http://dx.doi.org/10.1016/j.energy.2006.11.010.

G. Tullock. A comment on daniel klein's" a plea to economists who favor liberty". *Eastern Economic Journal*, 27(2):203–207, 2001.

A. Ushakova. Can we identify vulnerable energy customers in the uk using smart meter data? *MSc Dissertation, UCL*, 2015.

C. Véliz and P. Grunewald. Protecting data privacy is key to a smart energy future. *Nature Energy*, 3(9):702, 2018.

R. Walker, P. McKenzie, C. Liddell, and C. Morris. Area-based targeting of fuel poverty in northern ireland: An evidenced-based approach. *Applied Geography*, 34:639–649, 2012.

Y. Wang, Q. Chen, T. Hong, and C. Kang. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, 2018.

M. Weiss, A. Helfenstein, F. Mattern, and T. Staake. Leveraging smart meter data to recognize home appliances. In *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*, pages 190–197. IEEE, 2012.

H. Wilhite and R. Wilk. A method for self-recording household energy-use behavior. *Energy and buildings*, 10(1):73–79, 1987.

R. A. Winett and P. Ester. Behavioral science and energy conservation: Conceptualizations, strategies, outcomes, energy policy applications. *Journal of Economic Psychology*, 3(3-4):203–229, 1983.

I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

S. N. Wood. mgcv: Gams and generalized ridge regression for r. *R news*, 1(2): 20–25, 2001.

S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99 (467):673–686, 2004.

S. N. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2006.

J. M. Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2015.

X. Wu, X. Zhu, G.-Q. Wu, and W. Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016.

R. Yao and K. Steemers. A method of formulating energy load profile for domestic buildings in the uk. *Energy and Buildings*, 37(6):663–671, 2005.

Z. Yu, F. Haghighat, B. C. Fung, and H. Yoshino. A decision tree method for building energy demand modeling. *Energy and Buildings*, 42(10):1637–1646, 2010.

T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang. A new index and classification approach for load pattern analysis of large electricity customers. *IEEE Transactions on Power Systems*, 27(1):153 – 160, 2012. doi: 10.1109/TPWRS.2011.2167524.