MONASH University

Time series modelling and forecasting of disaggregated electricity data

Cameron Roach

B.Sc., The University of Melbourne

A thesis submitted for the degree of Doctor of Philosophy at

Monash University in 2020

Department of Econometrics and Business Statistics

Contents

Copyright notice

Abstract

The electricity industry is collecting large volumes of data from various sources. From regional grid demand to individual sensor readings in buildings, there is a wide range of disaggregated data sources requiring new techniques for forecasting and inference. A better understanding of how electricity is being used by consumers has the potential to increase energy efficiency and improve grid planning and management. This thesis presents several novel approaches to understanding these varied data sources. Two published papers and another two working papers are included.

Our first contribution is to present a methodology for creating coherent probabilistic forecasts in hierarchical settings. We find that our approach improves forecast performance compared to an appropriate benchmark model when assessed using the pinball loss scoring function. The effectiveness of the methodology is demonstrated using electricity consumption data from eight bottom-level zones and two aggregated zones in New England.

The second contribution of this thesis is to present an approach to understanding the effects of commercial office building attributes on electricity demand. We use smart meter data and mixed effects models to estimate each building characteristic's effect size and statistical significance throughout the day. This allows for the creation of demand impact profiles which clearly show how each attribute increases or decreases electricity demand.

Our third contribution focuses on using mixed models to improve point forecast accuracy for commercial office building electricity demand. Several benchmark models are assessed against various mixed effects models. We find that our proposed subject-specific curves model produces the best one-day ahead forecasts based on a variety of forecast accuracy

metrics. Furthermore, we demonstrate how such a model can be used to conduct scenario analyses by varying which attributes are present in a building.

The fourth and final contribution of this thesis focuses on visualising building management system sensor readings. Commercial buildings are often fitted with thousands of sensors that collect various readings for equipment operation and indoor environment quality. Interpreting this data is challenging due to the scale of data collection. We propose extracting time series and metadata features which are then transformed by several different dimensionality reduction techniques. We demonstrate how this approach can be used to detect anomalies in building operation.

In addition to the contributions listed above, all code for published papers has been made available publicly to help encourage future research in this area.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes two original papers published in peer reviewed journals and two working papers. The central theme is modelling and forecasting disaggregated time series data within the energy sector. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the Department of Econometrics and Business Statistics under the supervision of Rob Hyndman and Souhaib Ben Taieb.

In the case of Chapters 2, 3 and 5 I was the sole author of each paper. In Chapter 4 I am the lead author. The inclusion of co-authors in Chapter 4 reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research. My contribution to the work in each chapter is summarised in Table 1.

Thesis chapter	Publication title	Status (published, in press, accepted or returned for revision)	Nature and % of student contribution	Co-author name(s), nature and % of co-author's contribution	Co-author(s), Monash student Y/N
2	Reconciled boosted models for GEFCom2017 hierarchical probabilistic load forecasting	Published	100%	N/A	N
3	Estimating electricity impact profiles for building characteristics using smart meter data and mixed models	Published	100%	N/A	N
4	Subject-specific curves for time series forecasting of smart meter demand	Submitted	80%	Rob Hyndman (10%), Souhaib Ben Taieb (10%)	N
5	Exploring unusual sensor behaviour in buildings using BMS data and unsupervised learning techniques	Submitted	100%	N/A	N

Table 1: *Summary of contributions by author for each publication and working paper.*

I have not renumbered sections of submitted or published papers. Papers have been included as they were published.

Student name: Cameron Roach

Student signature:

Date: July 13, 2020

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor name: Rob J Hyndman

Main Supervisor signature:

Date: July 13, 2020

Acknowledgements

Throughout the creation of this thesis I often thought of the quote: *"We are what we repeatedly do."* by Will Durant. There have been many challenges over the last four years, physically and emotionally, and I will forever be grateful to those that showed me support during trying times. During the more difficult days I would remember these words and try to maintain some semblance of good research habits and self-discipline. There is no doubt room for improvement there, as there always will be, but for now I am happy with what I repeatedly do.

I'd like to single out a few people that have been particularly essential to the completion of this project. First, I must thank Craig Roussac of Buildings Alive who has shown support and faith in me throughout this project. Most importantly, he was receptive to the original project idea and for this I remain grateful. My supervisors Rob Hyndman and Souhaib Ben Taieb have both been supportive and given thoughtful feedback, however it was their patience during times of illness and recovery that I am most grateful for. It is so often said that the PhD process can be an isolating journey compounded by distant supervisors, but both were as far removed from this cold archetype as could be. I finish with the utmost respect for each, both professionally and personally. I would like to thank my parents. Despite their sometimes injudicious enquiries into when I might be finished, they continued to support and encourage me while I undertook this challenge. They nursed me back to health whenever it did wane, and without them I wouldn't have made it past my first year.

Finally, on more formal note, it would be remiss not to mention this research project was supported by funding from Buildings Alive. I would like to thank Buildings Alive for making data available and their guidance in understanding commercial building equipment and behaviour. This research was also supported by use of the Nectar Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS). I would again like to thank my supervisors Rob Hyndman and Souhaib Ben Taieb for their invaluable support and advice throughout the research process. I would like to thank my PhD panel which included Gael Martin, Di Cook, Farshid Vahid and Didier Nibbering for their thoughtful feedback and advice. Finally, I would like to thank the reviewers of submitted papers for their thorough and constructive feedback.

Preface

The article "Reconciled boosted models for GEFCom2017 hierarchical probabilistic load forecasting" presented in Chapter 2 has been published in the *International Journal of Forecasting*. Results from this article were also presented at the *International Symposium on Forecasting* held in Cairns, Australia, in June 2017.

The article "Estimating electricity impact profiles for building characteristics using smart meter data and mixed models" presented in Chapter 3 has been published in the journal *Energy and Buildings*.

The article "Subject-specific curves for time series forecasting of smart meter demand" presented in Chapter 4 was submitted to the *Journal of Forecasting*.

The article "Exploring unusual sensor behaviour in buildings using BMS data and unsupervised learning techniques" in Chapter 5 has been submitted to the journal *Energy and Buildings* and is currently under review. Preliminary results were presented at the *2018 Summer Study on Energy Efficiency in Buildings* held in Pacific Grove, USA, in August 2018 by the American Council for an Energy-Efficient Economy.

Chapter 1

Introduction

1.1 Overview

Due to the finite nature of electricity generation, the planning and usage of electricity networks is critical to a well functioning society (Esteves et al., 2015). The energy sector collects large volumes of data that can be used to help accomplish this (Pérez-Chacón et al., 2018; Yu et al., 2015). Network operators, distributors and private companies collect data that includes grid demand, individual building demand via smart meters and building management system (BMS) sensor readings. These data sources and the time series techniques used to analyse them are continually being explored and refined.

Grid demand is often recorded by zones, such as states or large regions, and represents the total demand in each. It is often used to forecast electricity demand which is in turn used for planning and grid management. Smart meters record electricity demand for individual households and businesses. As smart meters gather data for individual homes and buildings they allow for more fine-grain understanding of usage characteristics and trends. They are being installed in large numbers across many major economies. In the US there are an estimated 87 million installed smart meters as of 2018 (U.S. Energy Information Administration, 2018); China's state grid has finished their roll-out of electric smart meters (Research in China, 2019); and in Australia there are an estimated 2.8 million smart meters in Victoria alone (The State of Victoria Department of Environment,

Land, Water and Planning, 2016) while the national number is set to increase with the enforcement of a rule change by the Australian Energy Market Commission ensuring all new and replacement meters should be smart or advanced (Chan and Boddington, 2019; Australian Energy Market Commission, 2015). As smart meters typically record electricity demand at 15 or 30 minute intervals, they produce a large number of time series that can be used for inference and forecasting. BMS's are typically installed in large commercial buildings and are implemented to control heating, ventilation and air-conditioning (HVAC) systems and other equipment. These BMS systems are comprised of many sensors spread throughout each building collecting data on indoor environment quality, equipment status and electricity usage of individual items.

Some of the intended goals of data collection at such a large scale is to improve short and long-term forecasts for better grid planning; disaggregating consumption to appliances or equipment; and detecting anomalies in building performance and load management (Wang et al., 2018). In this thesis we explore each of these data sources and focus on two key goals:

1. To improve the performance of existing forecasting procedures.
2. To allow decision makers to better understand energy usage within buildings.

New methodologies are introduced in each chapter, each of which addresses one of these goals.

Expectations around the requirements for adequate energy demand forecasts have evolved in recent years. Where once point forecasts produced using univariate techniques were adequate (Taylor, Menezes, and McSharry, 2006); now probabilistic forecasts are expected (Hong et al., 2016; Hong, Xie, and Black, 2019) which allow us to assess both the anticipated demand and the uncertainty surrounding it. In fact, several recent reviews on electricity demand forecasting methodologies did not focus on probabilistic or quantile forecasts (Suganthi and Samuel, 2012; Yildiz, Bilbao, and Sproul, 2017; Singh et al., 2012), which highlights the rapid development that has occurred in recent years. Another recent review by Hong and Fan (2016) discussed the shift in focus towards probabilistic forecasting. Quantifying the uncertainty of forecasts through quantiles or density functions is of

paramount importance when producing peak demand forecasts where the peaks between years can vary dramatically, and is now common practice for some grid operators when producing peak demand forecasts (Australian Energy Market Operator, 2019).

A recent development in load forecasting makes use of the hierarchical structure of forecasts to improve accuracy. In cases where forecasts are to be produced for all levels of a hierarchy, we want to ensure that the forecasts at each level are consistent. By consistent, we mean all child nodes of the hierarchy sum to their parent nodes. Consistency not only produces sensible forecasts, but also results in better accuracy. While producing consistent hierarchical point forecasts is a relatively simple process, producing consistent *probabilistic* forecasts is a more challenging and less explored area. Addressing this research topic was the focus of the Global Energy Forecasting Competition 2017 (Hong, Xie, and Black, 2019) which produced the material contained in Chapter 2.

The above considerations apply equally to smart meter demand. Several studies have begun to focus on producing probabilistic forecasts at the building level (Ben Taieb, Taylor, and Hyndman, 2020; Ben Taieb et al., 2016; Arora and Taylor, 2016; Hong and Fan, 2016). While much of the existing smart meter analytics research has focused on improving forecasting performance, there are further uses for the data when combined with supplementary data sets. This is sometimes referred to as data fusion. Wang et al. (2018) discuss multivariate data fusion and state that in relation to smart meter data, "Very few papers consider weather data, survey data from consumers, or some other data. Integrating more external data… may reveal more information." Furthermore, they highlight that appropriate visualisation approaches to express the importance of various components is an overlooked area in need of exploration. A true boon for energy analytics is the ability to disaggregate demand by appliance or building characteristic and proceeding with a data fusion approach is a possible path forward. In Chapter 3 we attempt this by modelling smart meter data in combination with building characteristic data using mixed effects models. In Chapter 4 we also explore using the inherent similarity between commercial buildings to improve point forecast accuracy, again using mixed effects models. This has the added benefit of allowing for scenario analyses by varying the values of predictors such as the type of chiller system being used or distribution system.

While load forecasting has been the main focus of smart meter data (Wang et al., 2018), it has also been used to understand how buildings are working. Examples include clustering time series using mean load profiles, where each observation time of the day is a feature, before applying a suitable clustering algorithm such as k-means (Pérez-Chacón et al., 2018; Flath et al., 2012; Räsänen and Kolehmainen, 2009); or by engineering features based on the observed time series values before applying clustering (Haben, Singleton, and Grindrod, 2016). These unsupervised learning approaches allow us to group users into different categories which can then be used for customer segmentation. However, they fail to clarify how energy is being used by consumers. Clustering does not reveal how equipment or building characteristics affect electricity demand. Furthermore, approaches that rely on regular time series readings are not applicable in those cases where irregular time series intervals are present. Notably, we observe this in Chapter 5 when working with BMS sensor data. Here, irregular time series are present due to sensors being polled sequentially and at different rates. So while the aforementioned clustering approaches are useful, they can be further improved by allowing for irregular time series and the addition of complementary data sets. Understanding underlying behaviour of buildings using smart meter and BMS sensor data are the focus of Chapters 3 and 5, respectively.

Data collected by BMS systems can be used to improve energy efficiency. BMS systems are typically comprised of thousands of sensors throughout a commercial building each of which collects time series data on qualities such as equipment status, air temperature, damper position and electricity demand. Early efforts to leverage these point sensors to save energy relied on hard coded thresholds or rules that would raise alarms if they were violated. Lately, efforts have been made to explore and summarise the data in more effective ways to better enable facility managers to understand how their building is working. Research is still maturing in this field due to the difficulty of obtaining data from multiple buildings.

1.2 Challenges

Reliable forecasting is of paramount importance for grid management. Planning for adequate transmission capabilities to ensure supply and demand balance is a critical

component of a stable grid. Point forecasts alone fail to address this need due to the volatility of electricity demand. Instead, probabilistic forecasts that give a distribution of values are preferred. Management decisions can be based on the known probability of different demand levels. Some grid operators now incorporate probability into their demand forecasts. For example, the Australian Energy Market Operator produces demand forecasts at different probability of exceedance (PoE) levels, where a PoE of α is simply the $1-\alpha$ quantile of a density forecast (Australian Energy Market Operator, 2019).

Typically, a grid will be arranged in various zones each of which may require a demand forecast. Yet there may also be a need for estimates of total demand to ensure generation across interconnected zones can meet demand. In cases where demand at multiple levels of a hierarchy are required we can improve the overall forecast accuracy by using hierarchical forecasting techniques (Hyndman et al., 2011; Hyndman, Lee, and Wang, 2016; Wickramasuriya, Athanasopoulos, Hyndman, et al., 2015). Ensuring consistent forecasts is trivial for point forecasts as we can simply ensure the sum of bottom level zones is equal to any parent zone. Combining probabilistic forecasts in a hierarchical setting is more complicated and is a little researched field. We can not simply add the densities of bottom level zones to obtain a consistent top-level zone. In Chapter 2 we present an approach that produces consistent probabilistic forecasts across hierarchical zones while also showing that we can improve forecast accuracy during the hierarchical reconciliation step.

Unfortunately, there is a lack of work that explores drawing inferences from smart meter data. This is possibly due to a lack of suitable data sets. While some smart meter data sets are available publicly, these rarely contain attribute data. By highlighting the possibilities of combining smart meter data sets with building attribute data sets we hope to encourage the collection of complementary data sets that can be used with smart meter data. Chapters 3 and 4 show two applications of combining different data sets in both inference and forecasting settings. Chapter 3 fits mixed effects models to observe how different building characteristics influence demand at different times of the day and year. Demand impact profiles are produced which offer a clear visualisation of each attribute's importance throughout the day. Chapter 4 shows that using mixed effects models can improve forecast accuracy. Adding fixed effects for building characteristics allows for scenario

analysis, whereby we can create time series of expected electricity demand for different building configurations.

Many buildings have configuration issues. Incorrectly configured BMS's can result in a large wastage of energy. Stuck dampers for instance can result in heating and cooling occurring in the same room. Overly narrow dead bands can cause short-cycling and hunting. As there are often hundreds or thousands of sensors in a single building, it can be difficult to detect and diagnose anomalies in the system's operation. To help make this task easier for facility managers, we propose a dimensionality reduction approach for anomaly detection. We discuss this in detail in Chapter 5.

1.2.1 Open data

One difficulty of conducting research in this area is the scarcity of open-source smart meter and BMS data. Due to privacy and security issues, distributors and private companies are often reluctant to release data publicly (Wang et al., 2018). In order to conduct this research, I worked with Buildings Alive, a building performance company focusing on energy efficiency through rapid response. The smart meter data, building characteristics and BMS readings used in Chapters 3, 4 and 5 were made available by them. These data were not released publicly due to privacy concerns.

Electricity demand data at a more aggregated level is easier to come by. Electricity demand for eight zones in the New England electricity network was made available by ISO New England for the Global Energy Forecasting Competition 2017. These data were used for the research contained in Chapter 2 and were released publicly in an R package (see Section 1.3.3).

1.3 Contributions

This thesis was completed by publication. Four papers are included. Two have been published in peer-reviewed journals and one is currently under review. A fourth working paper will soon be submitted to a peer-reviewed journal.

1.3.1 Publications and Conferences

The material in Chapter 2 has been published in the *International Journal of Forecasting*. Results from this chapter were also presented in the *International Symposium on Forecasting* held in Cairns, Australia, in June 2017.

The contribution in Chapter 3 of this thesis has been published in the journal *Energy and Buildings*.

Chapter 4 was submitted for publication to the *Journal of Forecasting*.

Material in Chapter 5 was presented at the *2018 Summer Study on Energy Efficiency in Buildings* held in Pacific Grove, USA, in August 2018 by the American Council for an Energy-Efficient Economy. Preliminary work was also presented at an invited talk for the *Melbourne Data Science Meetup* held in Melbourne, Australia in November 2019. This chapter has been submitted for publication to the journal *Energy and Buildings*.

1.3.2 Other research activities

Throughout my thesis I have collaborated with the building energy efficiency company Buildings Alive. Additional research not presented here was conducted and included studies on classifying BMS sensor types using neural networks. Long short-term memory networks were used to classify what piece of equipment (e.g. air-handling unit, pump, etc.) or measurement type (e.g. on/off status, supply air temperature, return air temperature) each sensor belonged to using metadata that gave the name of each sensor. Unfortunately, when tested against a random forest trained off bigrams and trigrams of the name metadata we found that the simple benchmark was just as effective and so further research was postponed.

1.3.3 Software and open data

The code used to produce each of the papers is stored in dedicated GitHub repositories. Each repository contains functions and documentation that should make replicating these analyses possible. The R programming language (R Core Team, 2019) was the main language used for most papers, though Python's Scikit-learn library (Pedregosa et al.,

2011) was utilised for dimension reduction in Chapter 5. Code for the production of each accepted paper has been made available as dedicated GitHub repositories in the following locations:

- Chapter 2: https://github.com/camroach87/1701-gefcom
- Chapter 3: https://github.com/camroach87/1801-mmme.

As the raw data files used in Chapter 2 were only available in Excel spreadsheets, a tidied version of the data were released as an R package at https://github.com/camroach87/gefcom2017data. The R processing scripts, raw data and a tidy data set are contained in the repository.

Code for chapters that are still to be accepted by a journal will be made available upon acceptance at:

- Chapter 4: https://github.com/camroach87/1901-sscts
- Chapter 5: https://github.com/camroach87/1802-ufd.

Chapter 2

Reconciled boosted models for GEFCom2017 hierarchical probabilistic load forecasting

The following paper was published in the *International Journal of Forecasting* Volume 35, Issue 4, October–December 2019, Pages 1439-1450.

All code to reproduce the paper was made available at https://github.com/camroach87/1701-gefcom.

International Journal of Forecasting 35 (2019) 1439–1450

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Reconciled boosted models for GEFCom2017 hierarchical probabilistic load forecasting

Cameron Roach

Department of Econometrics & Business Statistics, Monash University, Australia

A B S T R A C T

When forecasting time series in a hierarchical configuration, it is necessary to ensure that the forecasts reconcile at all levels. The 2017 Global Energy Forecasting Competition (GEFCom2017) focused on addressing this topic. Quantile forecasts for eight zones and two aggregated zones in New England were required for every hour of a future month. This paper presents a new methodology for forecasting quantiles in a hierarchy which outperforms a commonly-used benchmark model. A simulation-based approach was used to generate demand forecasts. Adjustments were made to each of the demand simulations to ensure that all zonal forecasts reconciled appropriately, and a weighted reconciliation approach was implemented to ensure that the bottom-level zonal forecasts summed correctly to the aggregated zonal forecasts. We show that reconciling in this manner improves the forecast accuracy. A discussion of the results and modelling performances is presented, and brief reviews of hierarchical time series forecasting and gradient boosting are also included.

1. Introduction

Hierarchical time series forecasting occurs in situations where a dependent variable of interest can be disaggregated across the nodes of a hierarchy. Examples include forecasting the sales of a product both within towns and by state; forecasting economic indicators within states and for an entire country; or, in this case, forecasting the demand in both bottom-level and aggregated zones of an electricity network. When forecasting hierarchical time series, the base forecasts typically do not reconcile as one would expect; that is, the forecasts of the electricity demand in the bottom-level zones may not sum up to the forecasts of the aggregated zones. Hence, it is often necessary to carry out a reconciliation step to adjust these base forecasts.

This paper proposes a methodology for the hierarchical forecasting of electricity demand across eight zones in New England. In addition to the eight bottom-level zones, electricity demands for two aggregated zones are also forecast. This methodology was used in GEFCom2017 in the defined data track. Electricity and weather data were supplied by ISO New England. As this was the defined data track, only the electricity demand, dew point temperature, dry bulb temperature and calendar data were allowed as model inputs. We were presented with an ex-ante forecasting problem requiring forecasts of the 10th, 20th, … and 90th quantiles of the demand distribution for every hour of a future month for all zones.

Quantile forecasts for demand are produced by simulating weather scenarios for every zone in the forecast month. A demand model is then used to predict the demand for every zone and hour over the forecast horizon. Residuals are also simulated and added, which produces simulations of the actual demand rather than just the conditional mean. The zonal forecasts are then adjusted to ensure that they reconcile appropriately within each simulation, and quantiles are calculated for each hour using the reconciled demand simulations.

The boosted demand model is fitted using the XGBoost algorithm (Chen & Guestrin, 2016). Regularization with L1 and L2 penalties is applied in order to avoid over-fitting.

Recent work on hierarchical reconciliation has focused on adjusting the base forecasts to obtain reconciled forecasts with an improved accuracy (Hyndman, Ahmed, Athanasopoulos, & Shang, 2011; Wickramasuriya, Athanasopoulos, & Hyndman, 2015; Hyndman, Lee, and Wang, 2016). These methods focus only on adjusting the forecasts of the conditional mean. Despite the energy industry's shift towards probabilistic forecasting (Hong, Pinson, Fan, Zareipour, Troccoli, & Hyndman, 2016), the literature on

E-mail address: cameron.roach@monash.edu.

https://doi.org/10.1016/j.ijforecast.2018.09.009
0169-2070/

reconciling probabilistic forecasts in a hierarchical setting remains limited. To the best of the author's knowledge, the only relevant paper is that by Ben Taieb, Taylor, and Hyndman (2017), which proposes a methodology for producing coherent hierarchical probabilistic forecasts of smart meter demand. One contribution of the present paper is to enrich the literature on quantile forecasting for hierarchical electricity demand.

The paper has the following structure: Section 2 provides concise reviews of the relevant literature on hierarchical energy forecasting and gradient boosting. The competition's data and methodology are described in Sections 3 and 4. Section 5 discusses the modelling results, and concluding remarks are provided in Section 6.

2. Background theory

2.1. Hierarchical forecasting

When several time series exist in a hierarchy, it is necessary to ensure that the forecasts at each level of the hierarchy reconcile in a sensible manner. When time series exist in a hierarchy, they can be expressed in terms of a summing matrix $\mathbf{S}$ (Hyndman et al., 2011). This summing matrix allows all nodes to be expressed in terms of the bottom-level nodes. For an observation that occurs at time t,

$$\mathbf{y}_t = \mathbf{S}\mathbf{y}_{bt}, \tag{1}$$

where $\mathbf{y}_t$ gives the observed values for all aggregated and bottom-level nodes and $\mathbf{y}_{bt}$ gives the observed values for only the bottom-level nodes.

Hyndman et al. (2011) showed that the base forecasts can be reconciled using only the summing matrix $\mathbf{S}$. At forecast horizon h,

$$\tilde{\mathbf{y}}_h = \mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\hat{\mathbf{y}}_h, \tag{2}$$

where $\tilde{\mathbf{y}}_h$ are the reconciled forecasts and $\hat{\mathbf{y}}_h$ are the base forecasts. This is referred to as ordinary least squares (OLS) reconciliation, and was shown to outperform both bottom-up and top-down reconciliation approaches for both simulated and real-world data.

Subsequent studies by Hyndman et al. (2016) and Wickramasuriya et al. (2015) showed that reconciliation could be improved by incorporating a matrix of weights. A generalized least squares (GLS) approach is given by

$$\tilde{\mathbf{y}}_h = \mathbf{S}(\mathbf{S}'\boldsymbol{\Sigma}_h^{\dagger}\mathbf{S})^{-1}\mathbf{S}'\boldsymbol{\Sigma}_{\mathbf{h}}^{\dagger}\hat{\mathbf{y}}_h, \tag{3}$$

where $\boldsymbol{\Sigma}_h$ is the covariance matrix of the residuals for forecast horizon h and $\boldsymbol{\Sigma}_h^{\dagger}$ is the Moore–Penrose generalized inverse. It is often difficult to calculate $\boldsymbol{\Sigma}_h$, so an alternative weighted least squares (WLS) method can be used instead. If we let $\mathbf{W}$ be a diagonal matrix with elements that serve as the weights, the WLS approach is

$$\tilde{\mathbf{y}}_h = \mathbf{S}(\mathbf{S}'\mathbf{W}\mathbf{S})^{-1}\mathbf{S}'\mathbf{W}\hat{\mathbf{y}}_h. \tag{4}$$

Hyndman et al. (2016) suggested that the diagonal elements of $\mathbf{W}$ could be equal to the inverse of the h-step forecast variances. When using an ARIMA time series model, the h-step-ahead forecast variances can be approximated as proportional to the one-step-ahead forecast variances. Furthermore, since each fitted value is effectively a one-step-ahead forecast, residuals can be used to calculate these variances, making Eq. (4) a practical means of reconciliation. However, this approach is not necessarily feasible when dealing with other model types that do not produce one-step-ahead forecasts when fitted to historical values. Given this, Section 4.4.1 proposes two alternative weight matrices that can be constructed easily for any model.

2.2. Gradient boosting

Gradient boosting has been used with good results in many machine learning challenges (see for example Ben Taieb & Hyndman, 2014, and Koren, 2009). Gradient boosting was first proposed by Schapire (1990), and rigorous statistical overviews of boosting were carried out by Friedman, Hastie, and Tibshirani (2000) and Friedman (2001). Chen and Guestrin (2016) proposed the extreme gradient boosting (XGBoost) algorithm, which allowed for easy scaling while using less computational resources.

Essentially, boosting works by training an ensemble of weak learners that can then provide better predictions than a single model would be able to. Suppose that we are given a data set with n observations and p predictors, $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Then, the predictions are given by

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} \nu f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F},$$

where K is the number of weak learners used, ν is a shrinkage parameter that controls the learning rate, and $\mathcal{F}$ is the model space of the weak learners. Each $f_k(\mathbf{x}_i)$ is fitted in a stage-wise manner to the residuals r_i of the previous fit. Initially, the residuals are set equal to the observed response, $r_i = y_i$ for all i. Then, a weak learner f_k is fitted to the data set $\{(\mathbf{x}_i, r_i)\}_{i=1}^{n}$ for each step k, and the residuals are updated according to $r_i = r_i - \nu f_k(\mathbf{x}_i)$.

The weak learner is fitted by minimising the objective function

$$\mathcal{L}(\phi) = \sum_i l\left(\hat{r}_i, r_i\right) + \sum_k \Omega(f_k),$$

where l is a loss function and Ω is a penalty function that helps avoid over-fitting. The terms for L1 and L2 regularization are included within Ω, so the penalty function can carry out the lasso, ridge and elastic net types of penalisation effectively.

3. Data

We now present a brief overview of the data and the forecasting problem. A detailed discussion of the GEFCom2017 data is provided by Hong, Xie, and Black (2019). Hourly electricity data for eight zones spanning New England were made available by ISO New England, and hourly weather data comprising both the dry bulb and dew point temperatures were also provided. This analysis uses data from January 2005 to April 2017. When training a model for forecasting for a particular month, I used data from January 2005 up to two months prior to the start of the

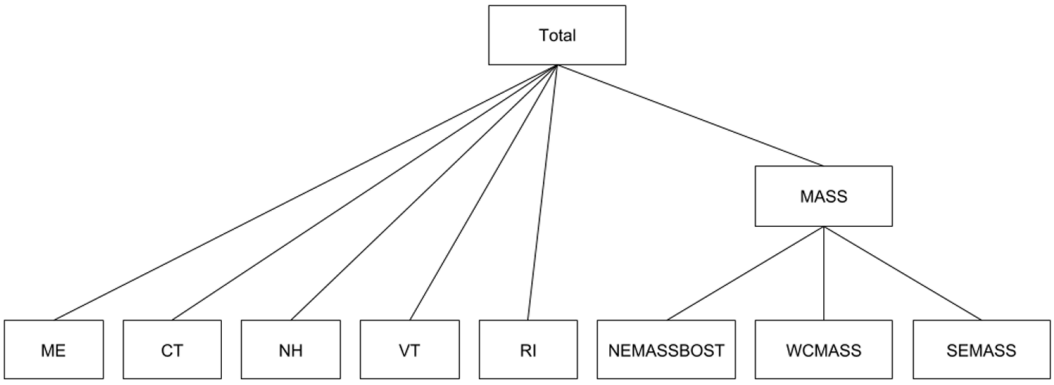

Fig. 1. Load forecasting hierarchy for GEFCom2017. There are two aggregated zones and eight bottom-level zones.

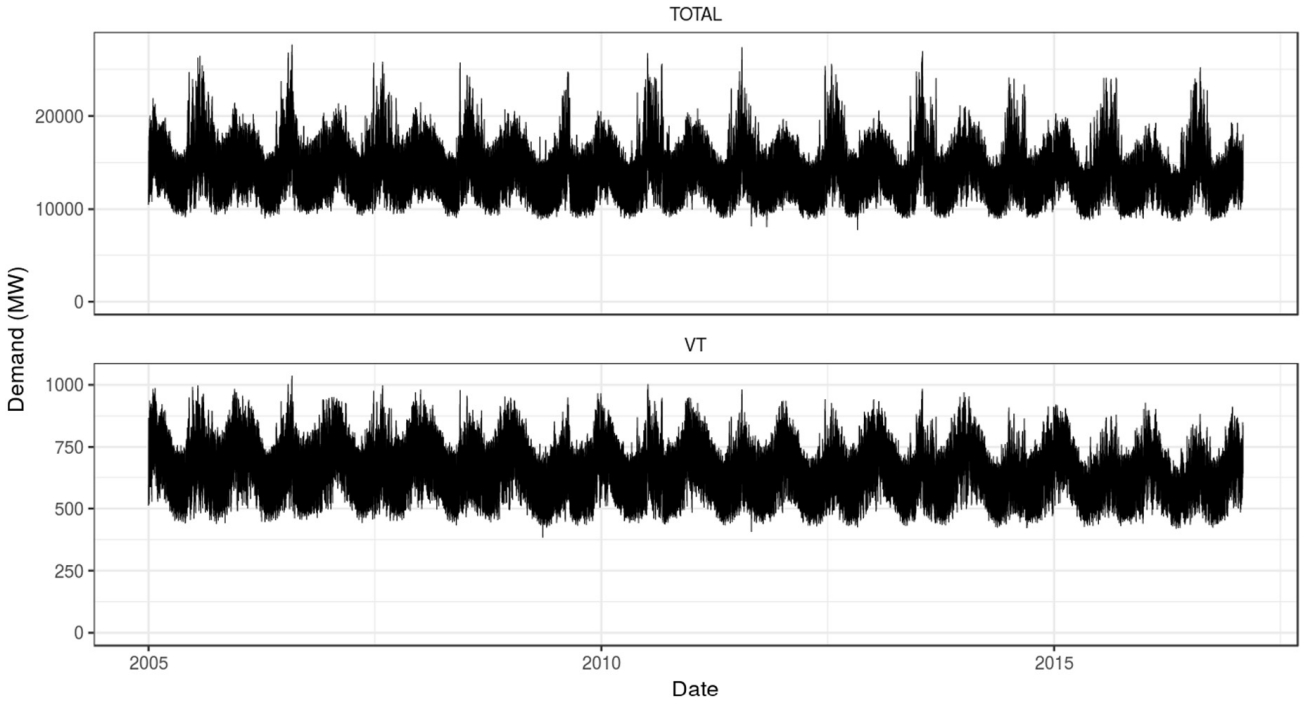

Fig. 2. Electricity demand for both the total of all zones and the bottom-level zone Vermont. Strong seasonality and volatility are observed for both the total and the bottom-level zone.

forecast period. For example, when forecasting April 2017, only data from January 2005 to January 2017 are used for training.[1] Public holiday data were also allowed in the competition. Massachusetts (MASS) consists of three bottom-level zones: Southeast Massachusetts (SEMASS), Western/Central Massachusetts (WCMASS) and Northeast Massachusetts (NEMASSBOST). The remaining bottom-level zones are Maine (ME), Connecticut (CT), New Hampshire (NH), Rhode Island (RI) and Vermont (VT). The sum of all eight bottom-level zones is designated "TOTAL". Fig. 1 shows the structure of the hierarchy.

[1] This two-month gap is generally consistent with how data arrived during the competition. To be clear, I do not expect that a two-month gap between the end of the training period and the start of the forecast horizon will improve forecasts; it is used only to ensure reasonable consistency with the competition proceedings.

3.1. Electricity demand

Fig. 2 shows the time series data for one top-level zone (Total) and one bottom-level zone (Vermont). Daylight saving time (DST) hours have been omitted, as they either contain a reading of 0 MW or are the sum of two periods.

3.2. Weather variables

The defined data track of GEFCom2017 allows only the dry bulb and dew point temperatures to be used as model predictors. Scatter plots of the demand and these two temperature variables for Maine are shown in Fig. 3. Note that similar relationships are present in all other zones. Fig. 4 shows that the two variables are strongly correlated except at higher temperatures. It seems reasonable to expect improvements in predictive power from including both temperature variables in the model.

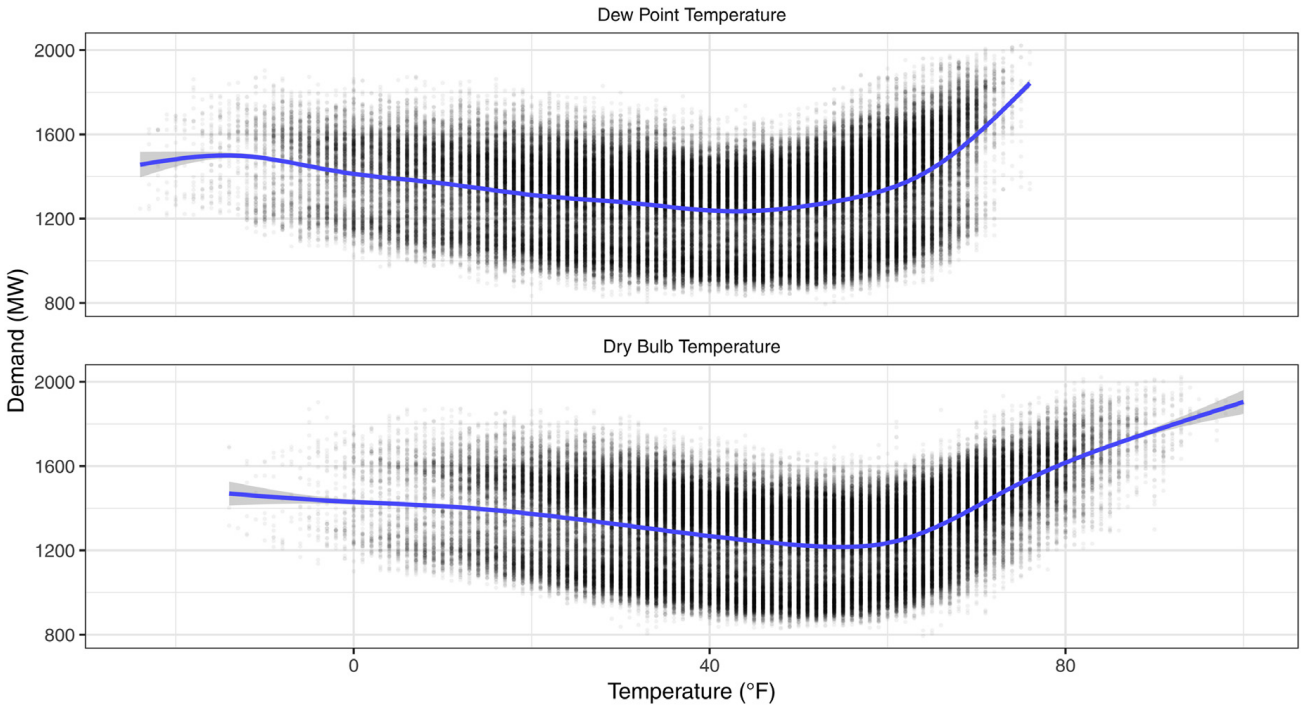

Fig. 3. Scatter plots of the demand and temperature variables in Maine.

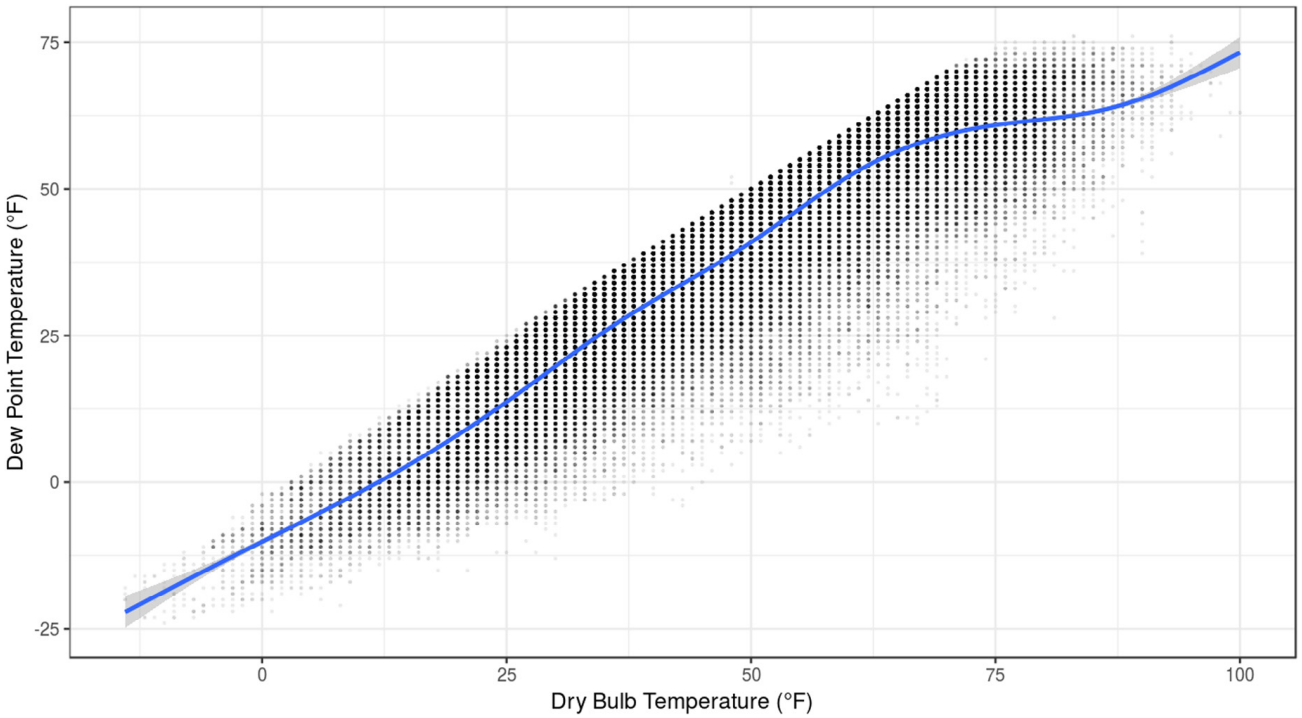

Fig. 4. Correlation between the dry bulb temperature and the dew point temperature in Maine. A non-linear relationship is evident in this scatter plot.

Each bottom-level zone has data from one weather station for each of these temperature variables. Naturally, aggregated zones have several stations available. All weather stations that belonged to a given aggregated zone were averaged to obtain the temperature variables. Using all weather variables separately was tested against this approach, but averaged temperature values were found to perform similarly when validating on a test data set.

3.3. Hierarchy structure

The hierarchy consists of eight bottom-level nodes and two aggregated nodes. It is an unbalanced hierarchy, with

three of the bottom-level nodes combining to form Massachusetts and the remaining bottom-level nodes and Massachusetts aggregating to form the total. A visualisation of this structure is provided in Fig. 1.

Fig. 1 can be represented in matrix notation using the summing matrix **S** from Eq. (1). Expressing the GEFCom2017 hierarchy in the form of Eq. (1) gives

$$\begin{bmatrix} y_{TOTAL,t} \\ y_{ME,t} \\ y_{NH,t} \\ y_{VT,t} \\ y_{CT,t} \\ y_{RI,t} \\ y_{MASS,t} \\ y_{SEMASS,t} \\ y_{WCMASS,t} \\ y_{NEMASSBOST,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} y_{ME,t} \\ y_{NH,t} \\ y_{VT,t} \\ y_{CT,t} \\ y_{RI,t} \\ y_{SEMASS,t} \\ y_{WCMASS,t} \\ y_{NEMASSBOST,t} \end{bmatrix},$$

where $y_{k,t}$ is the demand for zone k at time t.

4. Methodology

The following sections provide a detailed description of the forecasting methodology.[2] For a given month, I fitted a separate model for each zone using a gradient boosting algorithm. I then assessed the performances of L1 and L2 regularization using cross-validation. After selecting the regularization parameters that performed best, I forecast the demand for each zone over the forecast horizon using weather and residual simulations. This created demand simulations for each zone. Each demand simulation was reconciled so as to ensure that the sum of the child nodes was equal to their parent nodes. The final step involved calculating quantiles of the demand simulations for each hour of the forecast horizon.

4.1. Training and test data sets

Each month between June 2016 and April 2017 (the final month of the competition) was used as a test data set. While only four test sets (January 2017 to April 2017) were assessed in the competition, this paper expands on this in order to compare the baseline (Vanilla) and boosted models across each month of an entire year.

The models were trained using data from January 2005 to two months prior to the start of the forecast period. In general, this gap is consistent with how data arrived on the ISO New England website[3] during the competition, as there was usually a two-month processing time for new data. As was discussed in Section 3.1, daylight saving time (DST) hours were omitted. The training data set was used when carrying out parameter tuning via five-fold cross-validation. Residuals were calculated for the training set and were later used during the residual simulation step (see Section 4.3.2).

4.2. Model specification

4.2.1. Boosted model

I used a linearly boosted model from the XGBoost library (Chen, He, Benesty, Khotilovich, & Tang, 2017). The models were fitted in R (R Core Team, 2017) using the caret package (Kuhn, 2017) to carry out cross-validation and parameter tuning. A linear booster was chosen over a tree booster as the two gave similar results but the linear booster typically ran faster. Five-fold cross-validation was used when tuning, as this offered an acceptable compromise between the computational burden and variation in the folds.

An approach similar to that of Ziel and Liu (2016) was used when choosing predictors. For zone k, the following model was used:

$$y_{kt} = c_k(t) + f_k(\mathbf{w}_{kt}) + \epsilon_{kt}, \tag{5}$$

where at time t,

- y_{kt} is the demand;
- $c_k(t)$ is a linear function that models the effects of calendar variables;
- $f_k(\mathbf{w}_{kt})$ is a linear function that models the effects of weather variables;
- $\mathbf{w}_{kt}$ is a vector containing all weather and lagged weather variables; and
- ϵ_{kt} is the model error.

Eq. (5) is of the form discussed in Section 2.2. The calendar variables in $c_k(t)$ include

- public holidays;
- hour of day;
- day of week;
- day of year; and
- a trend term which is a natural number ordering the observations.

The weather variables in $\mathbf{w}_{kt}$ include

- current dry bulb and dew point temperatures;
- 72 hourly lags for dry bulb temperature; and
- 72 hourly lags for dew point temperature.

The choice of 72 hourly lags was made somewhat arbitrarily. The main goal was to include temperature data from the previous three days in order to capture any thermal inertia effects in buildings. This is an important factor in energy demand (Ben Taieb & Hyndman, 2014). More lags could well be added, but this would increase the computation time and I wished to avoid that. There are a total of 156 predictors. Note that the predictors were not scaled prior to fitting the models.

[2] A tutorial with R code is available from https://camroach87.github.io/post/2018-09-28-gefcom2017-tut-1/.

[3] https://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/zone-info.

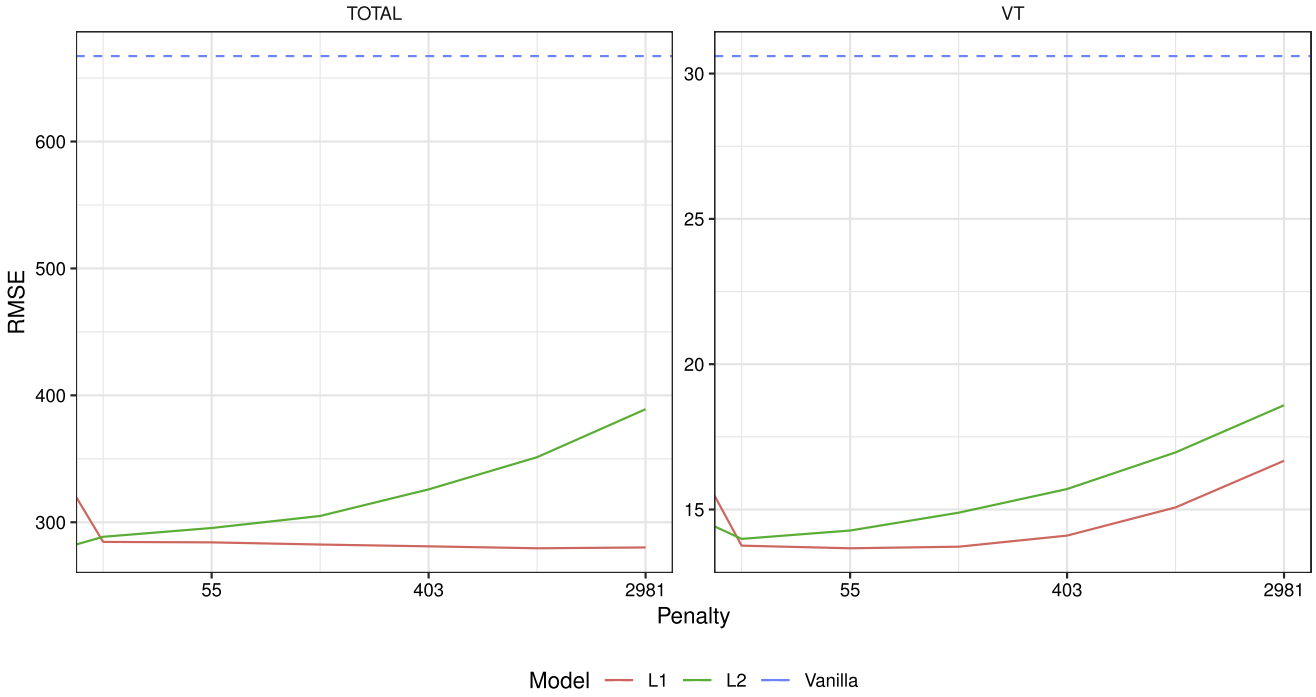

Fig. 5. Five-fold cross-validation RMSE scores. The x-axis gives the magnitude of the penalty size for both the L1 and L2 regularizations. Results are shown for one aggregated zone and one bottom-level zone, but similar results are observed for all other zones.

4.2.2. Vanilla model

For zone k, Tao's Vanilla model (Hong, 2010) is

$$\begin{aligned} y_{kt} = {} & \alpha_{0k} + \sum_{m=1}^{11} \alpha_{1km} M_{mt} + \sum_{d=1}^{6} \alpha_{2kd} D_{dt} \\ & + \sum_{h=1}^{23} \alpha_{3kh} H_{ht} + \sum_{d=1}^{6} \sum_{h=1}^{23} \alpha_{4kdh} D_{dt} H_{ht} \\ & + \alpha_{5k} \text{Trend}_k + f_k(T_{kt}) + \epsilon_{kt}, \end{aligned}$$

where at time t,

- T_{kt} is the dry bulb temperature;
- $f_k(T_{kt})$ models temperature effects;
- M_{mt} is a dummy variable for month $m \in \{1, 2, \ldots, 11\}$;
- D_{dt} is a dummy variable for day of the week $d \in \{1, 2, \ldots, 6\}$;
- H_{ht} is a dummy variable for hour $h \in \{1, 2, \ldots, 23\}$; and
- Trend_k is a natural number that orders the observations.

The temperature effects are modelled by

$$\begin{aligned} f_k(T_{kt}) = {} & \beta_{1k} T_{kt} + \beta_{2k} T_{kt}^2 + \beta_{3k} T_{kt}^3 \\ & + \sum_{m=1}^{11} (\beta_{4km} T_{kt} + \beta_{5km} T_{kt}^2 + \beta_{6km} T_{kt}^3) M_{mt} \\ & + \sum_{h=1}^{23} (\beta_{7kh} T_{kt} + \beta_{8kh} T_{kt}^2 + \beta_{9kh} T_{kt}^3) H_{ht}. \end{aligned}$$

Weather simulations for the models are constructed by shuffling historical weather data backward and forward by a maximum of four days. Each historical year and shuffled time series within serves as a simulation. As I was attempting to simulate the actual demand values, I also simulated residuals using variable-length block bootstrapping. Residuals were not simulated in the Vanilla model, which was consistent with the benchmark method of GEFCom2017.

4.2.3. Regularization

Due to the high dimensionality of our model, there was a risk of over-fitting to the training data. To manage this risk, I fitted several models with L1 and L2 regularization and different penalty values, then performed five-fold cross validation on the training data in order to select the best model. These regularized models were also tested against a baseline model. The baseline model chosen was Tao's Vanilla model (Hong, 2010), which has been used previously as a benchmark model (Hong, Pinson, & Fan, 2014) and was also used in GEFCom2017.

The change in root mean square errors (RMSEs) during five-fold cross-validation is shown in Fig. 5. Both the L1 and L2 regularized models outperformed the Vanilla model. With a sufficiently large penalty, the L1 model gave the best RMSE results.

4.3. Simulating in a hierarchy

The challenge requires competitors to forecast nine quantiles (10th, 20th, … and 90th) for every hour in a future month. This is an ex-ante forecasting problem, as we do not have any data for predictors in this situation. To forecast a demand distribution, I first simulated weather scenarios. Residuals were simulated by sampling from days with similar calendar characteristics.

As the New England zones form a hierarchy, it is necessary to preserve the correlations between them. For example, the weather in one zone will be correlated closely

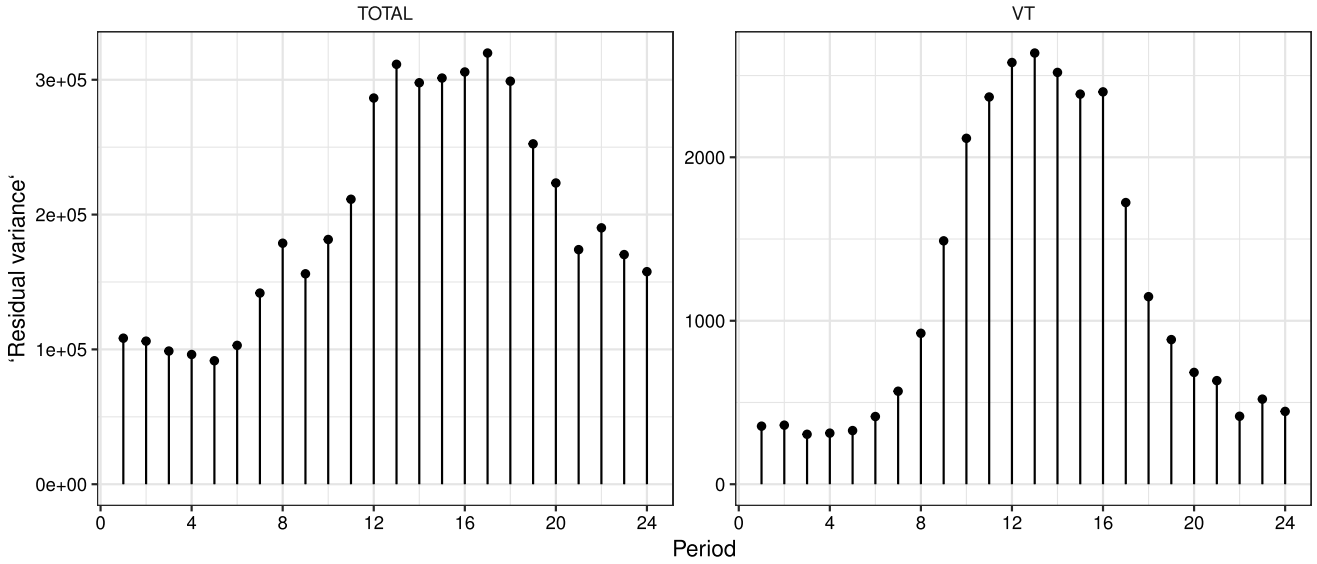

Fig. 6. Residual variance for each hourly period of the day. The variances have been calculated using residuals from all 12 training data sets. Results for one aggregated zone and one bottom-level zone are shown, though similar residual variance behaviours are observed in all other zones. The residual variance is highest during the middle of the day and lowest close to midnight.

with that an adjacent zone. Hence, the simulations need to reflect this. Correlations between zones are also present for residuals (Fig. 7), so care was taken when simulating residuals as well.

4.3.1. Weather simulations

Weather simulations were produced using the shifted-date method (Xie & Hong, 2018). Historical weather time series were shifted backward and forward by a maximum of four days each way, resulting in nine weather scenarios for each year. Eleven years of historical weather data were used, giving a total of 99 weather scenarios. This approach has the advantage of ensuring that realistic weather simulations are produced, as well as preserving weather correlation between zones.

A double seasonal block bootstrap approach similar to that of Hyndman and Fan (2010) was tested against this shifting approach but was found to perform worse. This is most likely to be due to the unrealistic discontinuities that are introduced at block boundaries during the bootstrapping process. This was not an issue for their paper's goal of predicting the maximum demand, but is a problem here.

4.3.2. Residual resampling

When predicting the demand for simulated weather data, the fitted model returns only a conditional mean. The error term in Eq. (5) also needs to be accounted for. To do this, I sampled from the historical residuals and added this sample to the predicted demand. This combination of the conditional mean and residuals produced a realistic demand simulation. The historical residuals were calculated by predicting the demand on the training data set and taking the difference between the predicted and actual demand.

When simulating the residuals, I sampled a sequence of historical residuals in order to preserve the correlation between adjacent observations in the time series. A variable-length block bootstrapping approach similar to that of Hyndman and Fan (2010) was used. A block of variable length was sampled from historical years at close to the same point of the year. The day of year that the block started from was allowed to vary by as much as seven days from the day of the year for which I required residuals, and the length of the block was distributed uniformly between 14 and 21 days. These numbers were somewhat arbitrary and can be varied, but produced reasonably realistic autocorrelation functions (ACFs) when compared to the actual (see Fig. 8 for an example). The correlations in the simulated residuals tend to be lower than those in the actuals due to discontinuities that are introduced at the borders of the blocks. To try to reduce the magnitude of the discontinuities, these block boundaries occurred at midnight when the variance of the residuals was usually lowest (Fig. 6).

It was also important to make sure that whatever dates were chosen when resampling were consistent between zones. Sampling different historical dates for each zone would lead to a breakdown in the inter-zone residual correlation, resulting in less realistic simulations. The residual correlations between zones are shown in Fig. 7.

I checked that the sampled residuals form a realistic time series by comparing their ACFs against those of the historical data (Fig. 8). The simulated residuals appear to have similar ACFs to the actuals. The ACFs of the simulated residuals are lower than those of the actuals, as expected, but to an acceptable degree.

4.4. Hierarchical reconciliation

Once a demand simulation has been created, it is necessary to reconcile all of the time series in the hierarchy. Here, I test several methods of accomplishing this.

4.4.1. Choosing weights for reconciliation

Three methods for reconciling the hierarchy were tested. The first involved using only the summing matrix **S**, as

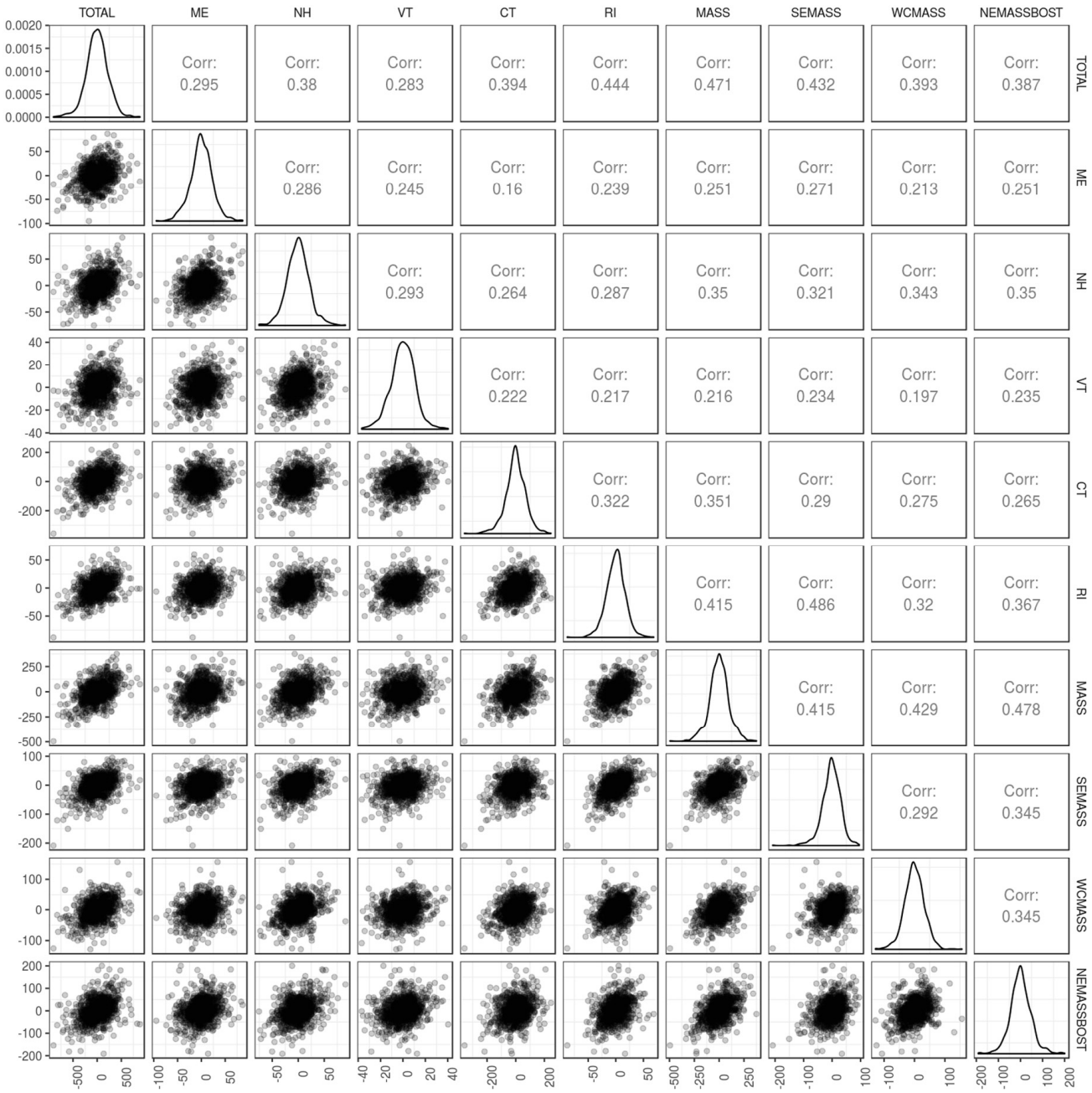

Fig. 7. Correlations of zone residuals based on 1000 points sampled from the hierarchy. Positive correlations are observed between all zones.

per Eq. (2), while the other two were based on specifying different weight matrices in Eq. (4).

As has been mentioned, the entities of $\mathbf{W}$ can be calculated based on the variances of ϵ_h. This works well with time series models where each fitted value is already a one-step-ahead forecast, such as ARIMA and exponential smoothing, but this is not the case for our model. Computing one-step ahead forecasts for the historical data would require the model to be refitted at each step, which is computationally prohibitive. As an alternative, I propose two different weight matrices: the first based on the mean values of each zone's demand and the second calculated from the variance of the residuals. The inverse matrices are specified as

$$\mathbf{W}_{\text{mean}}^{-1} = \frac{1}{\sum_k \bar{y}_k} \cdot \text{diag}\left(\{\bar{y}_k\}_{k=1}^{K}\right),$$

$$\mathbf{W}_{\text{var}}^{-1} = \text{diag}\left(\left\{\sigma_k^2\right\}_{k=1}^{K}\right),$$

where $\text{diag}\left(\{x_k\}_{k=1}^{K}\right)$ represents a diagonal matrix with elements $x_1, x_2, \ldots, x_K$; K is the total number of zones; and σ_k^2 is the variance of the residuals for zone k.

The intuition behind these weights follows from our goal of shifting the more accurate forecasts less than the inaccurate forecasts when reconciling. In the absence of one-step-ahead forecasts, the variance of residuals should serve as a useful proxy. Since residuals and demand are correlated, the mean weight matrix may also prove useful.

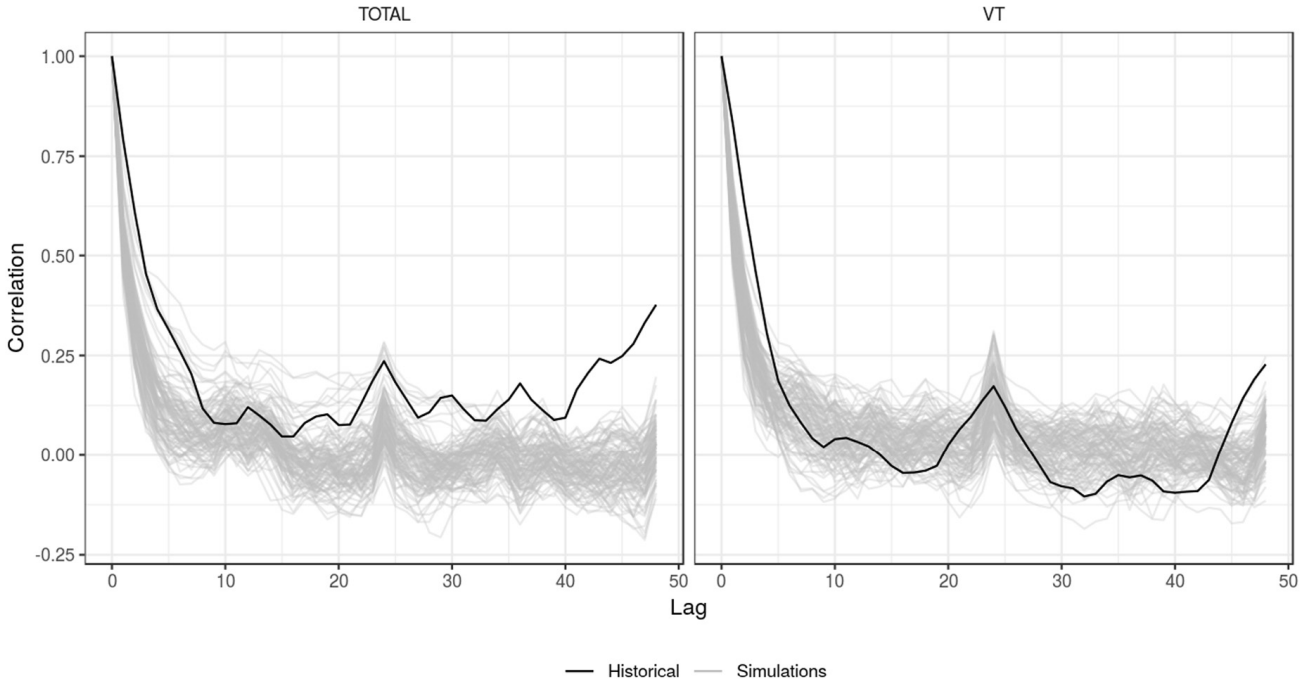

Fig. 8. Autocorrelation functions for historical residuals and simulated residuals in February 2017 for one aggregated zone and one bottom-level zone.

5. Discussion

5.1. Reconciliation results

Monthly RMSE scores for each hierarchical reconciliation method are shown in Fig. 9 and Table 1. All 12 test data sets have been used to produce these results. Overall, the WLS approach using $\mathbf{W}_{\text{var}}$ gives the best performance, while the WLS approach using $\mathbf{W}_{\text{mean}}$ performed slightly worse. The OLS approach performed the worst of the three reconciliation methods.

We investigate why the OLS approach performs worse than the WLS approaches by comparing the forecasts. Fig. 10 shows the base demand forecasts and reconciled forecasts of one of the simulations for one day in the forecast period. While the aggregated zone's reconciled forecasts look reasonable, the bottom-level forecast has severe variance introduced by the use of the OLS methodology. This variance appears in bottom-level zones that have only one parent zone (Total). It so happens that OLS adjustments made to the base forecasts for these bottom-level zones are of a magnitude comparable to that of the adjustments made to the Massachusetts aggregated zone, whereas the bottom-level zones that make up Massachusetts receive significantly smaller adjustments. This discrepancy in adjustments appears to be caused by the unbalanced structure of the hierarchy.

Given these results, the WLS reconciliation method using $\mathbf{W}_{\text{var}}$ as the weight matrix was chosen for reconciling the base forecasts.

5.2. Quantile forecast results

Quantile forecasts were produced based on the WLS reconciliation approach using $\mathbf{W}_{\text{var}}$ weights and L1-regularization, as this model appeared to perform best.

Table 1
RMSE scores for each reconciliation method, averaged across all zones.

Month	Unreconciled	OLS	WLS (mean)	WLS (residual variance)
May 2016	178.1	170.6	138.8	124.5
June 2016	175.4	165.7	135.5	127.9
July 2016	238.0	226.4	188.8	178.7
August 2016	204.0	191.6	160.4	154.7
September 2016	171.6	164.7	144.5	139.8
October 2016	117.5	111.0	107.2	115.0
November 2016	133.4	125.2	113.2	114.4
December 2016	204.4	199.2	178.5	170.7
January 2017	140.6	132.0	114.7	112.0
February 2017	150.5	144.8	131.3	128.0
March 2017	178.1	172.3	165.6	169.4
April 2017	166.5	158.8	160.3	171.1

Examples of the quantile forecasts for one aggregated zone and one bottom-level zone are shown in Fig. 11. An inspection suggests that the quantile forecasts capture the variance in the actuals well. Benchmarking is carried out against the Vanilla model to improve our understand of how well the model is performing.

5.3. Benchmarking against the Vanilla model

The pinball loss scoring function can be used to assess quantile forecasts (Gneiting, 2011). For a given probability level τ, the pinball loss function is defined as

$$L_\tau(y, q_\tau) = \begin{cases} \tau(y - q_\tau) & \text{for } y \geq q_\tau, \\ (1-\tau)(q_\tau - y) & \text{for } q_\tau > y. \end{cases}$$

A lower expected pinball loss score indicates better performance. The expected pinball loss for each model and zone can be estimated by taking the mean of all observed

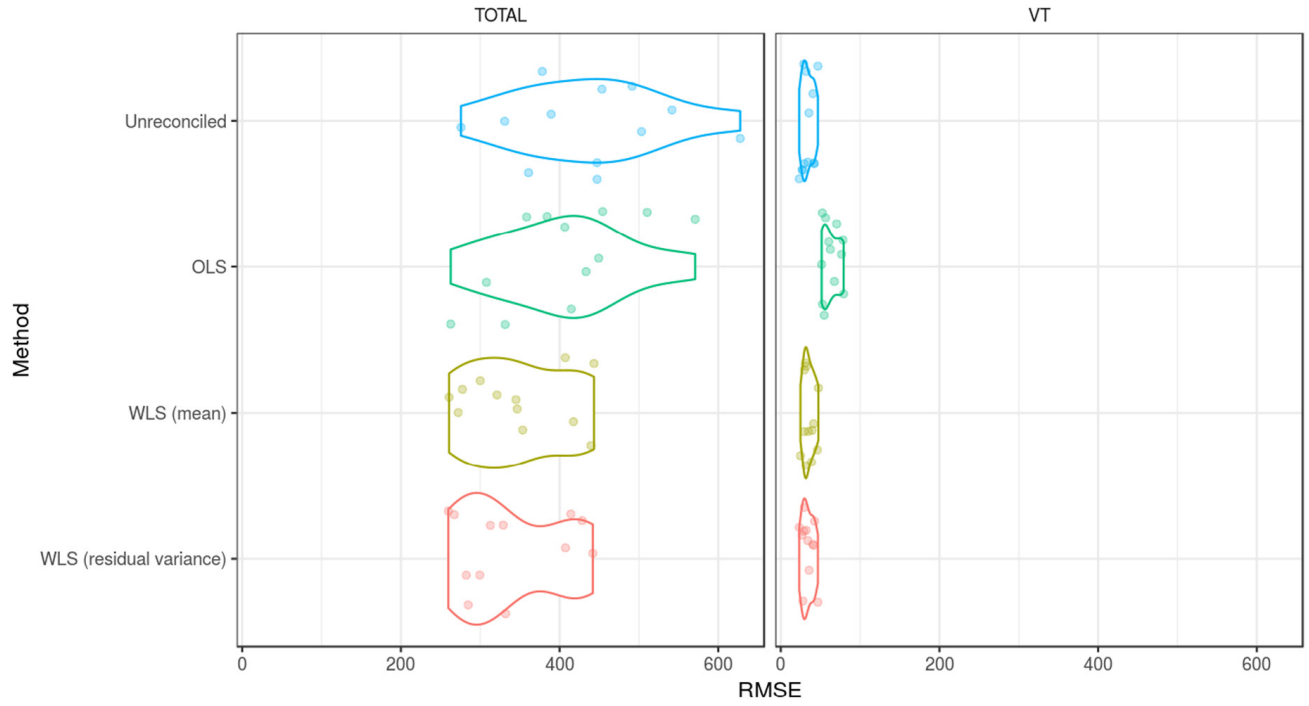

Fig. 9. Hierarchical reconciliation results. The RMSEs for each forecast month from May 2016 to April 2017 are plotted. The contoured lines are violin plots, and represent the density. Similar results are observed in all other zones.

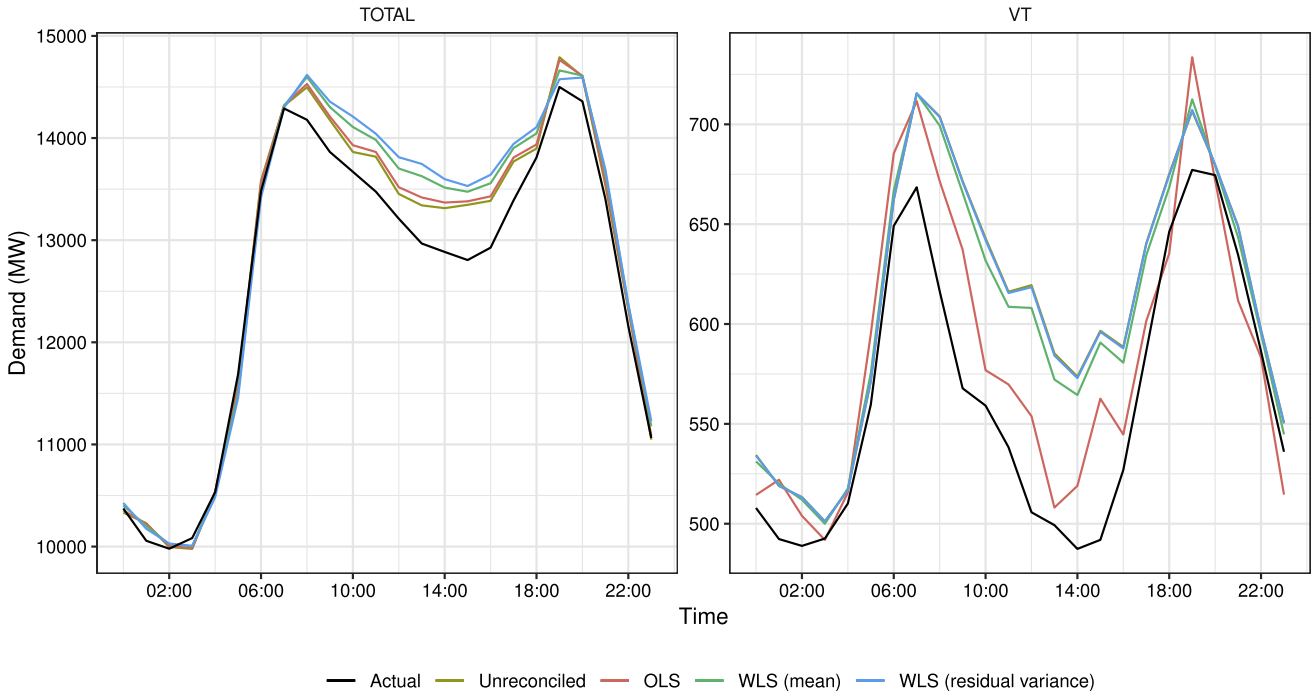

Fig. 10. Original and reconciled forecasts using different weights. Note that I have deliberately chosen a day where over-forecasting occurs in order to better show how the OLS reconciliation method introduces variance.

$L_\tau\,(y_{kt}, q_{kt\tau})$, where $q_{kt\tau}$ is the quantile forecast at probability level τ for zone k at time t.

A comparison of the Vanilla and boosted[4] models is provided in Tables 2 and 3. The boosted model almost always outperforms the Vanilla model. The only exception is for August 2016, when both models appear to perform poorly relative to other months.

5.4. Future research

The performance of the boosted algorithm has been explored here in one context. However, it could potentially be interesting to see how such a model might perform

[4] WLS reconciliation approach using $\mathbf{W}_{\text{var}}$ weights and L1-regularization.

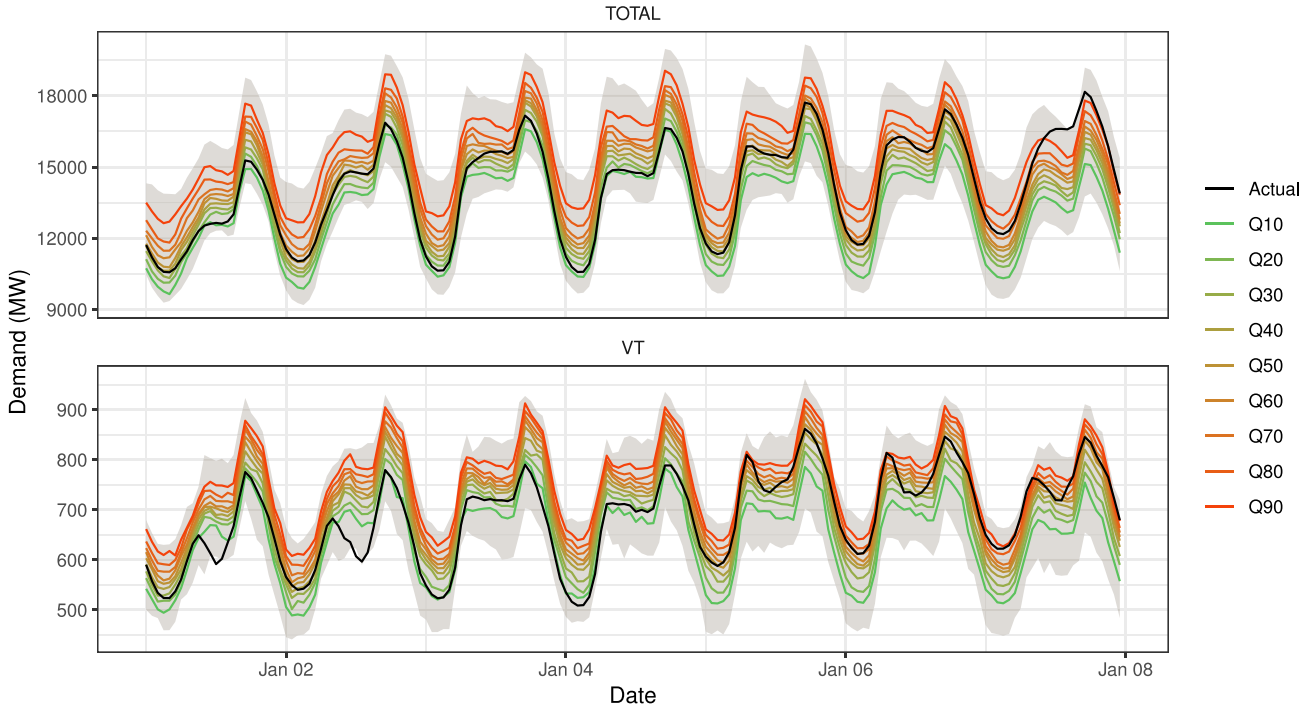

Fig. 11. Actuals and quantile forecasts in the first week of January 2017. The shaded areas show the maximum and minimum simulated demand values.

Table 2
Expected pinball loss scores for each zone averaged across all 12 test sets. Lower values indicate better performances.

Zone	Boosted	Vanilla	Percentage improvement
CT	99.9	108.3	7.8%
MASS	159.6	185.9	14.2%
ME	19.8	22.7	12.6%
NEMASSBOST	69.6	81.6	14.7%
NH	30.8	32.6	5.5%
RI	25.2	27.5	8.4%
SEMASS	47.9	55.4	13.5%
TOTAL	330.5	375.2	11.9%
VT	15.0	18.6	19.4%
WCMASS	48.3	55.0	12.1%

Table 3
Expected pinball loss scores for each forecast month averaged across all zones. Lower values indicate better performances.

Month	Vanilla	Boosted	Percentage improvement
May 2016	74.3	53.5	28.1%
June 2016	75.8	72.7	4.2%
July 2016	160.5	149.2	7.1%
August 2016	168.0	175.2	−4.3%
September 2016	128.7	119.3	7.3%
October 2016	47.4	33.5	29.2%
November 2016	62.8	39.1	37.8%
December 2016	77.3	66.4	14.1%
January 2017	102.4	87.3	14.8%
February 2017	101.8	83.5	18.0%
March 2017	94.9	84.8	10.7%
April 2017	60.0	49.6	17.3%

when forecasting over different horizons. Another area that might be of interest is focusing on other methods for dealing with unbalanced hierarchies, for example adding artificial nodes to balance the hierarchy. Both of these topics are left for future research.

6. Conclusion

This paper has presented a methodology for producing probabilistic hierarchical forecasts. A demand model based on linear gradient boosting has been shown to outperform a commonly-used benchmark model. In addition, both L1 and L2 regularization have been found to improve the model fit. The best performance was observed using a sufficiently large L1 penalty.

Weather simulations were produced by shifting the weather history back and forth by up to four days. Residual simulations used a variable-length block bootstrapping approach. Forecast reconciliation between nodes of the hierarchy was carried out using several different methods, and it was found that using a weight matrix based on the variance of residuals performed best. The advantages of this approach are that the bottom-level zonal forecasts sum correctly to the aggregated zonal forecasts and the forecast accuracy is improved compared to unreconciled models.

Finally, the quantile forecasts produced by the gradient boosted model outperformed those of a commonly-used baseline model. Quantile forecasts were assessed using the pinball loss function. The gradient boosted model performed better in all zones in the hierarchy over a year of monthly forecasts.

Acknowledgments

I am grateful for the financial support of Buildings Alive (Fund: 1752291; Grant: 512031) and to its CEO Craig Roussac. I would also like to acknowledge the invaluable support of my PhD supervisors Rob Hyndman and Souhaib Ben

Taieb. Finally, I would like to thank the referees of the paper for their time and thoughtful feedback.

References

Ben Taieb, S., & Hyndman, R. J. (2014). A gradient boosting approach to the kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382–394.

Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2017). Hierarchical probabilistic forecasting of electricity demand with smart meter data. URL https://robjhyndman.com/papers/HPFelectricity.pdf, Working paper.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). New York, USA: ACM.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2017). xgboost: Extreme gradient boosting. URL https://CRANR-project.org/package=xgboost, R package version 06-4.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407.

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting*, *27*(2), 197–207.

Hong, T. (2010). Short term electric load forecasting (Ph.D. thesis), North Carolina State University.

Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, *32*(3), 896–913.

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, (in preparation).

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589.

Hyndman, R. J., & Fan, S. (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, *25*(2), 1142–1153.

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, *97*, 16–32.

Koren, Y. (2009). The BellKor solution to the Netflix grand prize. In *Netflix prize documentation*.

Kuhn, M. (2017). caret: Classification and Regression Training. URL https://CRANR-project.org/package=caret, R package version 6.0-7.6.

R Core Team (2017). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. URL https://www.R-project.org/.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2015). *Forecasting hierarchical and grouped time series through trace minimization*: *Working Paper 15/15*, Department of Econometrics and Business Statistics, Monash University.

Xie, J., & Hong, T. (2018). Temperature scenario generation for probabilistic load forecasting. *IEEE Transactions on Smart Grid*, *9*(3), 1680–1687.

Ziel, F., & Liu, B. (2016). Lasso estimation for GEFCom2014 probabilistic electric load forecasting. *International Journal of Forecasting*, *32*(3), 1029–1037.

Cameron Roach received his M.Sc. in Mathematics and Statistics at the University of Melbourne. He is currently a Ph.D. student in the Department of Econometrics and Business Statistics at Monash University (Australia).

Chapter 3

Estimating electricity impact profiles for building characteristics using smart meter data and mixed models

The following paper was published in the journal *Energy and Buildings*, Volume 211, March 2020.

All code to reproduce the paper was made available at https://github.com/camroach87/1801-mmme.

Energy & Buildings 211 (2020) 109686

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild

Estimating electricity impact profiles for building characteristics using smart meter data and mixed models

Cameron Roach*

Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria 3800, Australia

ARTICLE INFO

Article history:
Received 23 November 2018
Revised 27 September 2019
Accepted 8 December 2019
Available online 24 December 2019

Keywords:
Smart meters
Energy consumption
Mixed effects models
Multimodel inference
Office spaces

ABSTRACT

Understanding the impact of building characteristics on electricity demand is important for policy and management decision making. Certain building characteristics and equipment may increase or decrease electricity consumption. Due to different operating practices, these impacts on electricity consumption may vary both across the day and across seasons. Quantifying the magnitude and statistical significance of these impacts will help managers and policy makers make better informed decisions. Here we present a mixed effects model to assess the importance of several variables on building electricity consumption. We use smart meter and building attribute data for 129 commercial office buildings. Our building attribute data includes information on installed equipment and meter characteristics of each building. To account for uncertainty in both variable significance and model selection we follow a multimodel inference approach. Demand impact profiles that show the expected change in electricity demand when a characteristic is absent or present are produced for each season. A discussion of the commercial office building characteristics we use and their impact on the daily profile of electricity demand is presented. Our approach has the advantage of only requiring building level demand and characteristic data. No equipment level sub-metering is required. Furthermore, our approach can also be used to quantify changes in electricity consumption caused by other factors that do not directly draw electricity from the grid, such as management decisions or occupant behaviour. We conclude with a discussion of applications for our methodology and future research directions.

1. Introduction

There is an increasing need to focus on the composition of electricity demand. Whereas in the past aggregated demand was sufficient for decision making, we are now often required to delve deeper and understand what underlying factors influence demand. Doing so can give a clearer picture of which building characteristics, occupant behaviours and policies are best able to reduce electricity consumption. For example, at a state or country level we may be interested in measuring the impact of solar or battery power on a typical power demand profile. At a building level facility managers may be interested in which equipment or building characteristics improve energy efficiency and at what points of the day they typically draw demand. Policies with the greatest efficacy can be identified and further promoted. Equipment and building management practices that are most efficient can help guide management and retrofitting decisions.

In this paper we focus on estimating the impact of several building characteristics on daily electricity demand. Electricity demand behaviour is typically presented in the form of power demand profiles which show the expected demand across a day. We define *demand impact profiles* as the observed change in power demand profiles when a particular characteristic is or isn't present. Demand impact profiles show how a characteristic increases or decreases demand over an entire day. Separate profiles are produced for each season due to the strong seasonality of some characteristics (e.g. heating and cooling equipment). Our models are trained on smart meter data from commercial office buildings across Australia and building characteristic data which describes which characteristics are present for each building. As we only require time series data and metadata on individual buildings our methodology can be repeated for any scenario where similar data is available. All code and a walk-through have been made available online to allow for easy implementation (see Section 3.5).

Several approaches to disaggregating smart meter data and quantifying the consumption of appliances exist. Most of these approaches differ to ours. They attempt to reconstruct time series for the integrants making up total metered demand whereas

* Corresponding author.
E-mail address: cameron.roach@monash.edu

https://doi.org/10.1016/j.enbuild.2019.109686
0378-7788/

we attempt to understand the average impact across the day during a particular season. Dinesh et al. [9] use Karhunen Loeve expansion to decompose low frequency smart meter measurements into the appliance level electricity demand. Kalluri et al. [15] examine time series subsequences to study appliance loads. Reinhardt et al. [27] and Weiss et al. [32] use several classification approaches that focus on identifying different appliances at a residential level. Disaggregation of household smart meter data to air-conditioning loads is explored by Perez et al. [24] and validated on 19 buildings. Load disaggregation using supervised classification techniques are explored in Chahine et al. [7] and Liao et al. [17]. Guo et al. [13] propose an algorithm for modelling appliance level consumption when only aggregated data is present and validate on a synthetic dataset. A load disaggregation approach for commercial buildings is proposed by Norford and Leeb [23].

A drawback of these approaches is they can only assess the impact of equipment that draw electricity directly from the grid. They cannot assess other factors such as solar energy generation or the impact of building facade properties such as glazing. Behavioural factors can not be quantified either. Furthermore, uncertainty can not be assessed whereas we construct confidence intervals for our coefficient estimates.

A key advantage of other approaches is that some can create demand time series for certain loads, but we do not focus on achieving the same goal with our method. Our goal is to produce power demand profiles for characteristics and equipment of interest. We feel this is a reasonable aim. A decision maker is not interested in individual time series from a large number of buildings. Rather, they require a summary that clearly communicates how a characteristic of interest influences demand. Sub-metering may play a role in automating systems and diagnosing problems with individual pieces of equipment, but a statistical approach allows us to circumnavigate the time and costs associated with sub-metering to arrive at the same destination - a summary of how electricity usage is typically affected. In fact, the statistical approach goes further by allowing us to assign a degree of uncertainty to our findings. We can obtain an estimate for the expected change in electricity demand and a measure of confidence with which we can accept or reject our findings. In cases where inference is to be conducted we propose our multimodel inference approach using mixed models over traditional frequency analysis approaches.

Some papers have focused on calculating power demand profiles for building equipment. Gunay et al. [12] and Mahdavi et al. [19] used a data driven approach to calculate power demand profiles for office equipment. Plug loads were recorded for several pieces of office equipment and a predictive model was then trained off this metered data. A similar approach was used by Christiansen et al. [8] to assess the energy consumption of medical equipment in hospitals. Menezes et al. [20] offered two approaches to creating power demand profiles. The first relies on sampling from a database of monitored loads for equipment of interest. While it does allow for the calculation of confidence intervals, a drawback is its dependence on the quality of the metered data for each piece of equipment. Another matter is that it can only assess plug loads that have previously been recorded in a database. A second approach that did not rely on a database of metered data was also proposed, but was dependent on knowing or assuming the operational schedules of small office equipment which can sometimes be difficult. Our methodology has a similar aim to these studies, but can produce estimated power demand profiles without relying on a database of metered data for equipment or assuming operational schedules of equipment.

There are many building characteristics and items of equipment that can affect electricity demand. To avoid a naive data dredging analysis we approached building engineers to identify factors they suspected were influencing demand or were of interest to them. Our data included building attributes such as the type of electric equipment installed, building use and building meter characteristics (did the metered demand contain tenant usage?). The characteristics they identified are discussed in more detail in Section 2.2. We then used multimodel inference to test if these were statistically significant predictors and estimate how they influenced electricity demand over the day.

Multimodel inference is an information theoretic approach to model selection that relies on fitting multiple models with different combinations of predictors and then averaging the best performing models based on a suitable weight metric [5]. It is commonly used in ecology [30] but is not often used in the energy sector. To our knowledge the only example is So and Richman [29] which used multimodel inference to create candidate models for the disaggregation of combined meter data for university campus buildings. It is a standard approach when there is no model specification that clearly outperforms others or when there is some uncertainty around which predictors to include. In our case it was necessary as using an all subset approach failed to produce a clear best model. Furthermore, simply picking the best performing model after an all subset analysis fails to take into account model selection uncertainty and often leads to inflated p-values [5]. Multimodel inference allows us to account for both model selection and parameter estimation uncertainty, thereby giving more reliable estimates of variable importance.

Electricity meter data is correlated within buildings. Some buildings will consistently have high demand, and others low. Mixed effects models allow us to account for this *within-subject* correlation by treating each building as a random effect. The random effect size indicates how much the mean of each individual's response variable differs from the sample population's mean. By modelling each building as a random effect, each building is treated as a random sample from a population of buildings with a specific distribution. Instead of model residuals being the only random component of our model, the random building selection is also taken into account. Mixed effects models also include fixed effects which are non-random quantities. In our case, fixed effects are the different building characteristics that we wish to create demand impact profiles for.

The main contribution of this paper is a methodology for generating demand impact profiles of various building characteristics at different times of the year. In addition to estimating the conditional mean we show how to calculate confidence intervals for our coefficient estimates that include both model and parameter selection uncertainty. We apply our approach to a real world dataset consisting of 129 commercial office buildings. In summary, the key advantages of our proposed methodology are:

- Only smart meter and building characteristic data are required to create demand impact profiles.
- Confidence intervals that include both parameter estimation and model selection uncertainty are produced.
- The impact of building characteristics other than plug loads can be quantified.

The paper is structured as follows. Section 2 discusses the time series data and metadata that motivates our research. Our mixed effects model and the multimodel inference approach are introduced in Section 3. Section 4 presents our estimates for building characteristic demand impact profiles and discusses applications and future research. Concluding remarks are made in Section 5.

2. Data

Data has been provided by Buildings Alive. Metered electricity consumption and building characteristic data is available for 129 buildings across Australia. Several years of data are available for

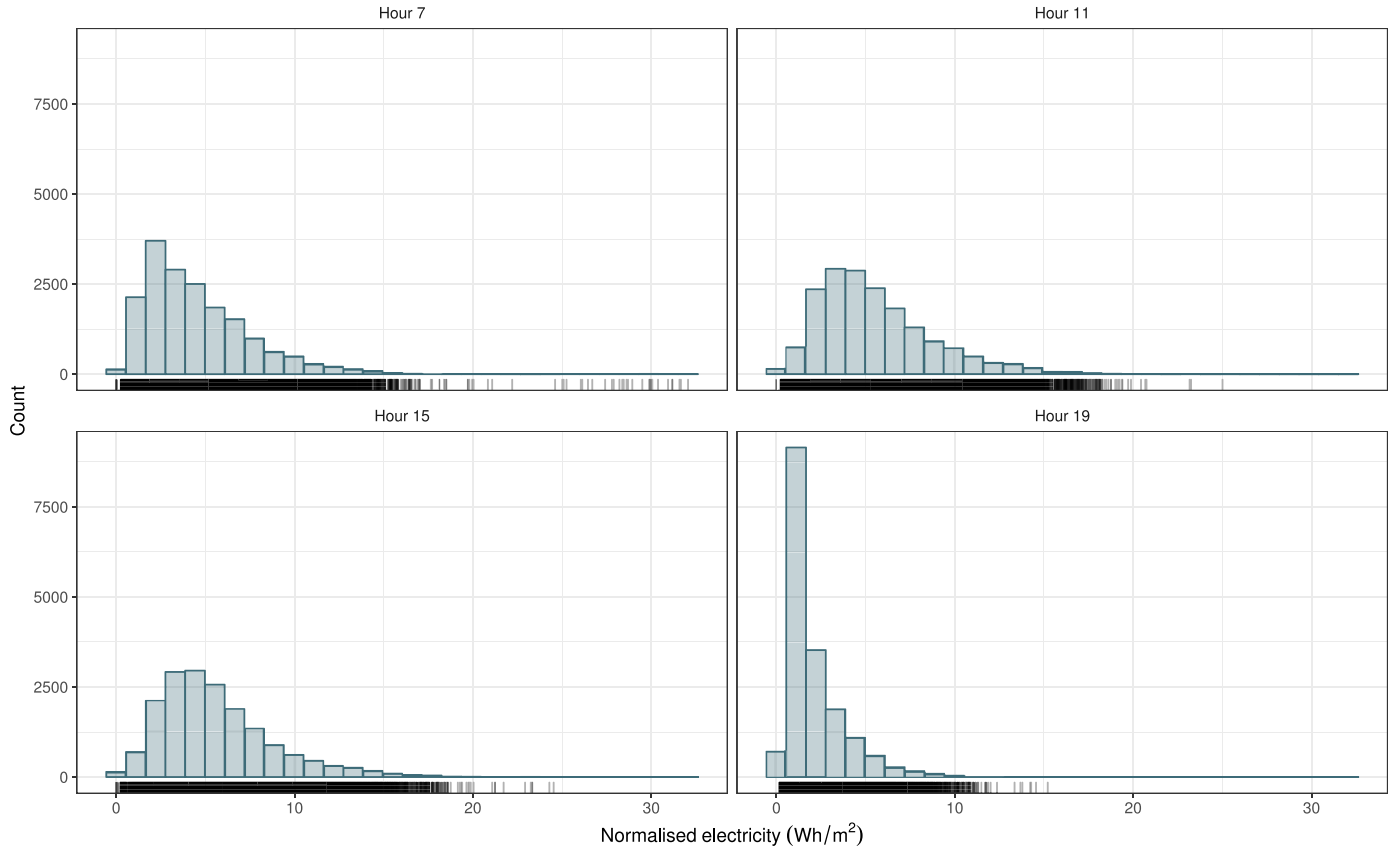

Fig. 1. Histogram of normalised electricity consumption (Wh/m^2) for all commercial office buildings. A sample of 5% of meter readings at four different times of the day are shown. The readings are positive and right-skewed. Similar distributions occur for the remaining hours of the day.

each building. In general, data is available from 1 March, 2015 to 1 March, 2018 for most buildings, although some buildings have slightly less data. Only working days are included when fitting our model as electricity usage is dramatically different on non-working days (weekends and public-holidays). Electricity consumption tends to be significantly lower on non-working days due to equipment not being in use.

2.1. Time series data

We have 15-min electricity consumption data available for each building. Time series plots of the raw data is shown for several buildings in Fig. 3. In our analysis we divide the 15-min electricity consumption values by each building's net lettable area to obtain *normalised* electricity consumption. These values are always positive and right-skewed (Fig. 1) suggesting the use of a log-transform when modelling.

2.1.1. Hourly grouping

In our analysis we group the 15-min interval readings into hours. While this reduces the granularity of the profile somewhat, it still allows us to assess the overall behaviour across a day while reducing the amount of variance in our coefficient estimates. It also means that each model can be fit to more data than would be available if 15-min models were used. Working with 15-min models or other temporal groupings, such as business and non-business hours, is also possible and can be chosen based on an analyst's needs. For our paper we prefer hourly models as they provided useful demand profile visualisations while also allowing sufficient training data for each model.

2.1.2. Outlier filtering

Smart meter data from buildings can be very noisy and typically contain outliers as seen in Figs. 2 and 3. To avoid having our results adversely impacted by outliers we removed them from our analysis using a simple approach. For each hour, season and building the bottom (1st) quantile and top (99th) quantile were trimmed before fitting models. While this may have resulted in some valid values being excluded, it offered a quick way to remove the worst outliers.

2.2. Building attributes

We were motivated in our research by engineers that wished to statistically assess the relevance and importance of different building characteristics on electricity demand. Several characteristics were available but, to avoid overfitting and missing value issues, we limited ourselves to a subset that was of interest to their company and had high data quality. Our main research goal was to understand how each of these selected characteristics affected electricity demand. This is one of the contributions of our methodology - it allows us to assess the statistical significance and effect size of whatever characteristics we are presented with using only smart meter data and attribute data.

Three characteristic types are considered. *Direct* characteristics consume electricity and include items such as cooling equipment and electric heating equipment. *Behavioural* characteristics include tenant and management practices. *Indirect* characteristics are those that affect electricity demand but do not themselves use electricity. Examples can include glazing, insulation and gas heating equipment.

The variables we consider for modelling are described below.

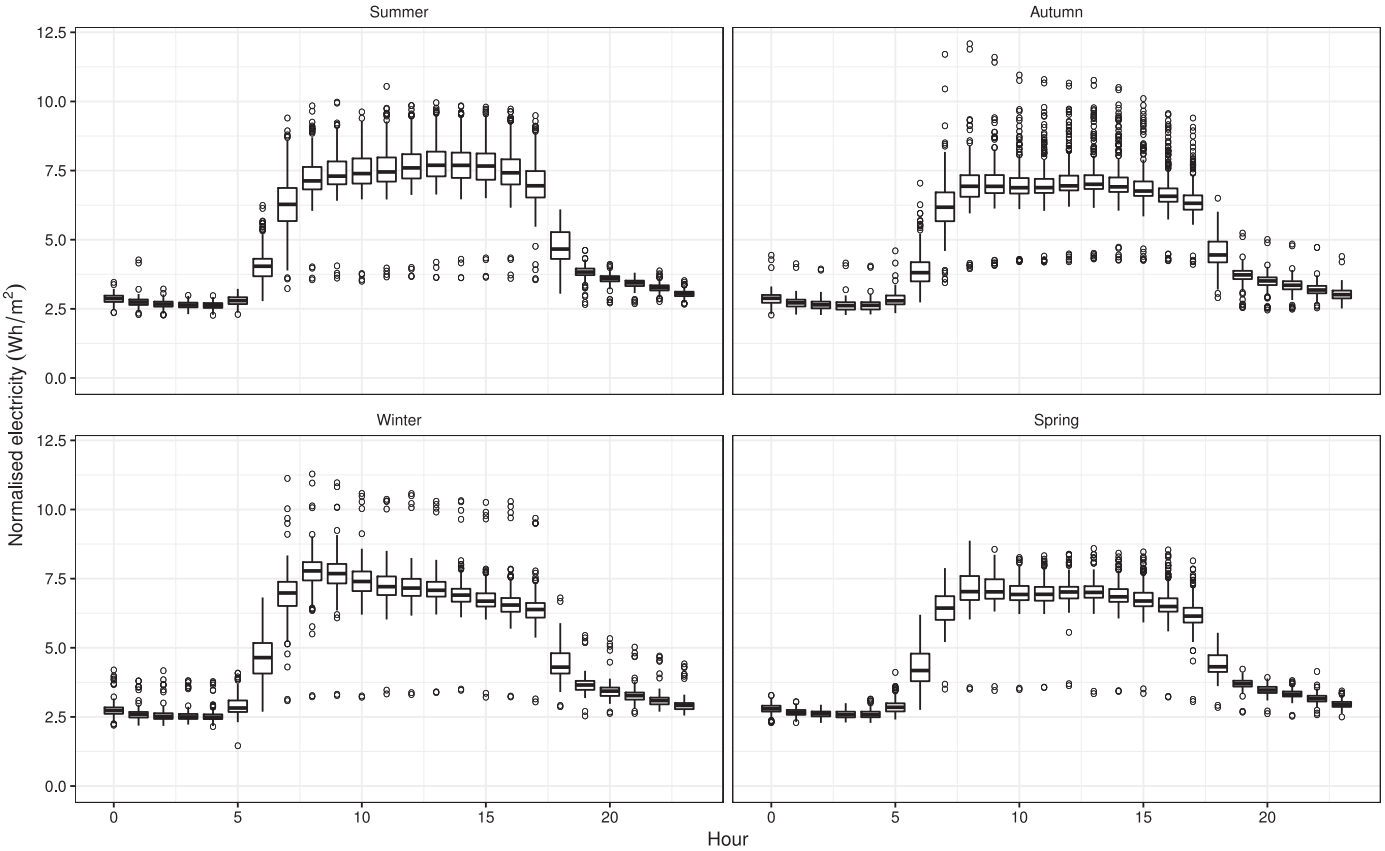

Fig. 2. Hourly boxplots of normalised electricity consumption for building BID0212. Outliers are shown by circles. We see that there have been periods of low occupancy during business hours which result in outliers.

Table 1
Building attributes for six of the 129 buildings.

	Building					
Attribute	BID0045	BID0061	BID0123	BID0210	BID0717	BID0720
Central Dist	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE
DXSystem	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
Electric Element Heating	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE
Gas Fired Boiler	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE
Tenant Feed	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
Water Cooled Condenser	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE

2.2.1. Tenant feed

(behavioural) Some metered data only included electricity demand from the base building and excluded tenant usage. This variable identified if tenant consumption was included in the metered demand. Tenant data may include electricity demand from plug loads (e.g. computers, air conditioning) and lighting.

2.2.2. Water cooled condenser

(direct) A water cooled condenser discharges heat by transferring stored heat energy from a refrigerant to running water. The heated water may then be cooled in a cooling tower.

2.2.3. DX system

(direct) Direct expansion (DX) systems cool air. We were initially presented with two variables, DX system and chiller system, that had strong negative correlation. In other words, buildings would typically be equipped with one or the other. DX systems cool air whereas chiller systems cool water. We chose to work with DX system as one of our variables, though the choice was somewhat arbitrary as the main goal was to avoid multicollinearity. Reverse DX was another available variable, however this was omitted as only five buildings had reverse DX and there was obviously strong correlation with the DX system variable.

2.2.4. Gas fired boiler

(indirect) Gas fired boilers provide heating. They are fuelled by natural gas or propane and do not directly impact electricity demand. If a gas boiler is installed there will be no need to use electric heating to warm the building. Hence, they have an indirect effect as they offset the electricity that would otherwise have been required by an electric system.

2.2.5. Electric element heating

(direct) These heating systems are powered by an electrical source. A current is passed through metallic heating elements causing the elements to heat due to their resistance.

2.2.6. Centralised distribution

(direct) We were initially provided with *centralised distribution* and *per-floor distribution* variables. These had strong negative correlation (buildings typically had one or the other) and so we chose to use only centralised distribution in our model. Centralised distribution systems generate all cooling from one location and rely

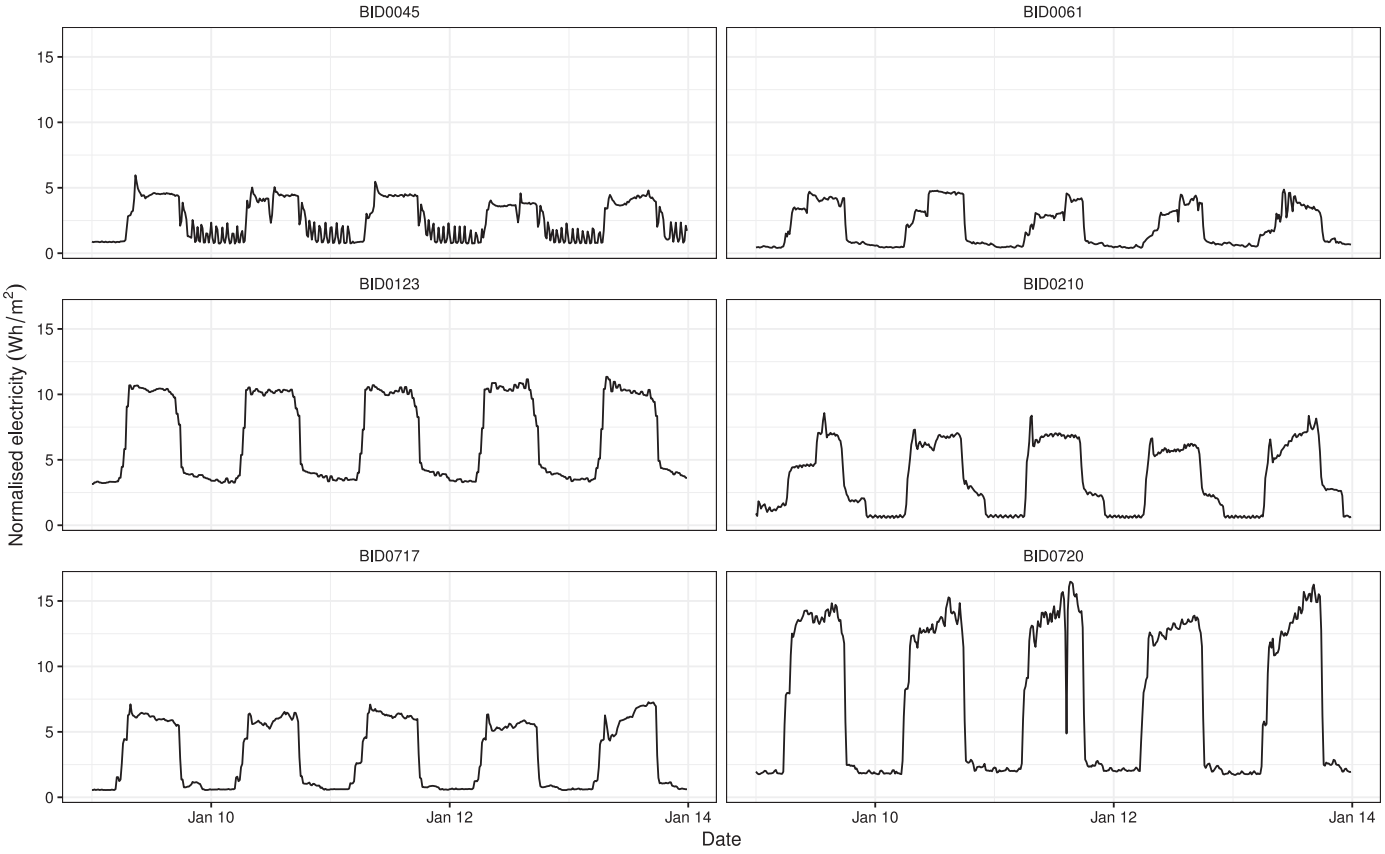

Fig. 3. Time series plots of commercial office building smart-meter data for six of the 129 buildings. Each building has distinct business hour and non-business hour behaviour. The shape and volatility of the readings differ between each building. It is difficult to observe a clear relationship between the attributes listed in Table 1 and electricity demand using these visualisations alone. This motivates the development of our proposed approach that allows us to statistically test and quantify the relationship of each attribute and electricity demand.

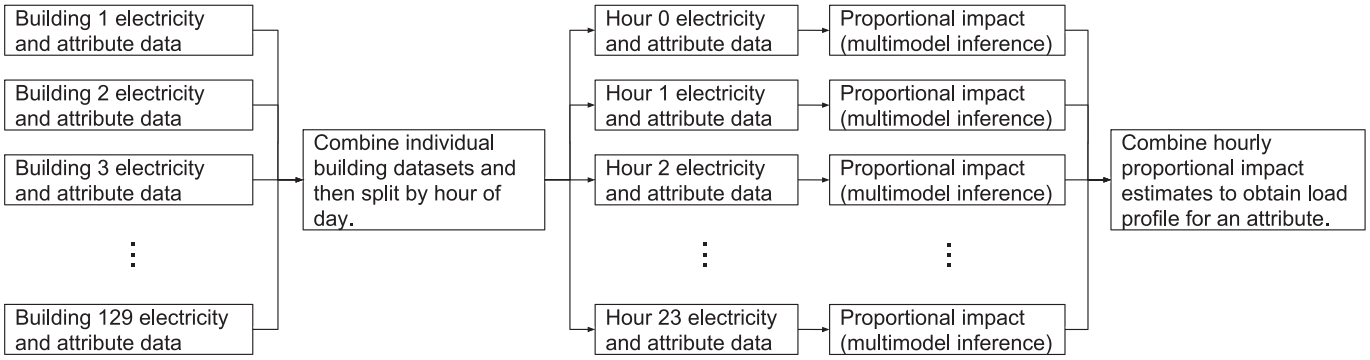

Fig. 4. Flowchart of the proposed methodology for a given season.

on ductwork for air distribution whereas per-floor distribution has cooling units on each floor.

Examples of the building attributes for six buildings are shown in Table 1. It is difficult to observe a relationship between each of these attributes and normalised consumption by simply comparing them to the electricity consumption plots in Fig. 3. High volatility in each time series, coupled with the large number of buildings and the variation between buildings makes the attribute and demand relationship unclear. This motivates our research into finding an appropriate model to quantify their impact on demand.

3. Methodology

Fig. 4 shows a flowchart of our proposed methodology. All of the available electricity and attribute datasets for each building are combined into one and then split by hour of day. Multimodel inference, parameter estimation and the estimation of proportional impacts is carried out separately for each hour (and season). The final step involves combining these hourly proportional impact estimates to produce a demand profile showing the estimated change in demand when an attribute is or isn't present. A simulation study that examines the capability of our proposed methodology is provided in Appendix B.

3.1. Mixed effects model

We use a linear mixed effects model to describe the relationship between building attributes and electricity consumption. Mixed effects models can capture within-subject correlation which allows us to estimate parameters common to an entire popula-

tion (fixed effects) and subject-specific parameters (random effects) [25]. If we consider each building as a subject we have within-subject correlation in the electricity meter readings. Hence, we treat each building as a random effect. Building characteristics can be interpreted as population parameters and so they are modelled as fixed effects.

We create separate models for each season and hour of the day to allow for changes in building behaviour that typically occur over the course of the day and year. For instance, non-business hours will often have significantly less consumption than business hours. Peaks can occur in the morning due to pre-heating or pre-cooling. Heating equipment will be more important in winter and cooling equipment in summer. Distinct models for each hour and season allow for a comparison of their estimated fixed effect coefficients at different times. Plotting these estimates and their confidence intervals provide a clear overview of how each attribute's impact on demand evolves.

For a particular season and hour of the day the electricity consumption over 15-min intervals, y_{ij}, for building j and observation $i \in \{1, 2, \dots n_j\}$ is given below. The Boolean variables x_{hij} are equal to one if the hth building attribute is present and zero otherwise. We use a log-transform on the consumption data as it is positive and right-skewed (see Fig. 1). The mixed model is

$$\begin{gathered}\log y_{ij} = \beta_0 + \textstyle\sum_{h=1}^{p} \beta_h x_{hij} + u_{0j} + u_{1j} t_{ij} + \epsilon_{ij}, \\ \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u), \quad \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}, \\ \epsilon_{ij} \sim N(0, \sigma^2_\epsilon),\end{gathered} \tag{1}$$

where β_h is the coefficient for attribute indicator x_{hij}, u_{0j} is the random intercept for building j, u_{1j} is the random slope coefficient for year t_{ij} and ϵ_{ij} are the residuals. This linear mixed effects model allows the intercept to vary with each building. A random slope for year has been added to capture any trends in building performance. We include a covariance term σ_{u01} in the variance-covariance matrix to allow for correlation between the random intercept and random slope. This is chosen as we tend to observe negative correlation between the random intercept and slope.

We also attempted to model residuals using an autoregressive correlation structure of order 1. While this did reduce autocorrelation in the standardised residuals [25] it had almost no impact on our final fixed effect coefficient estimates and confidence intervals. Furthermore, the $AR(1)$ correlation structure resulted in a considerable increase in computation time and occasionally convergence issues. Given this, we chose to model residuals as in Eq. (1).

3.2. AIC for mixed effects models

A means to assess each model for goodness of fit and complexity is required when conducting model selection. Complicated models may fit data better, but fail to generalise to new datasets. This is indicative of over-fitting rather than a well specified model. Information criteria such as the Akaike information criteria (AIC; Akaike [1]) and Bayesian information criteria (BIC; Schwarz [28]) score models on how well they fit data while also penalising them for complexity. Choosing a suitable information criteria for a linear mixed effects model is typically more complicated than in linear regression due to issues arising from the selection of covariance structures and positive semidefinite constraints on the covariance matrix [21].

For our model selection we use the marginal AIC (mAIC) which is the most widely used information criteria for mixed effects models. Vaida and Blanchard [31] define the mAIC as

$$mAIC = -2\ell\left(\hat{\beta}\right) + 2(p+q), \tag{2}$$

where $\ell\left(\hat{\beta}\right)$ is our log-likelihood function, p is the number of fixed effects, q is the number of random effects. We choose this criterion as it is both simple to understand and has been used in many studies [21].

3.3. Multimodel inference

Despite the effort taken to carefully determine appropriate building characteristics for our model there is still uncertainty regarding the importance of each proposed variable. Multimodel inference allows us to incorporate this model selection uncertainty into our parameter estimates. Instead of simply conducting stepwise variable selection or best subset selection and then choosing the best model, we instead use a candidate set of models on which we base our inference. This model averaging approach has merit as sometimes the best model will only offer a small improvement over other models based on a quality score such as AIC. Had a different dataset sample been present it may have resulted in another model being selected as the best [5]. Furthermore, different models will sometimes show the same variable being significant or insignificant [30]. The overall conclusion is that model selection uncertainty needs to be taken into account when conducting inference in situations such as ours. If we only focus on parameter estimation uncertainty without considering model selection uncertainty we will likely underestimate the size of our confidence intervals.

Multimodel inference focuses on selecting a subset of models on which to base inference. Note that effort should be taken to avoid data dredging (where the computer is left to select the best variables with no prior hypothesis on the researcher's part). Burnham and Anderson [5] suggest using prior knowledge of the situation to determine suitable predictors. Fitting only a subset of all possible models is a sensible approach because fitting all possible models can quickly result in a large computational burden if the number of predictors is large[1]. As discussed in Section 2.2 we have narrowed all available predictors down to a reasonable subset based on advice from domain experts.

Given our subset of predictors, multimodel inference fits a separate mixed effects model to every combination of predictor variables. This gives 2^p models each with an *mAIC* score. The best performing models form our candidate set on which we conduct inference.

3.3.1. Candidate sets

To construct our candidate set we must first determine the probability of each candidate model. To do so we use Akaike weights. Given R candidate models the Akaike weight for model g_i is

$$w_i = \frac{\mathcal{L}(g_i|\mathbf{x})}{\sum_{r=1}^{R} \mathcal{L}(g_r|\mathbf{x})} = \frac{e^{-\frac{1}{2}\Delta_i}}{\sum_{r=1}^{R} e^{-\frac{1}{2}\Delta_r}}, \tag{3}$$

where $\mathcal{L}(g_i|\mathbf{x})$ is the likelihood of model g_i given data $\mathbf{x}$ and $\Delta_i = mAIC_i - mAIC_{min}$ is referred to as the AIC difference. We use *mAIC* in our analysis as we are dealing with mixed effects models. Other information criteria for mixed effects models such as the conditional AIC [11,31] may also be used.

We can use the AIC differences and Akaike weights to select a subset of models for inference. We refer to this subset as the candidate set. For our analysis we use the common practice of selecting models with the highest weights such that their cumulative sum is just above 95%. As the weights serve as model probabilities we can refer to this as a 95% confidence set.

[1] Given p predictors we would require 2^p models to be fitted.

3.3.2. Parameter estimates

Parameters are estimated by "averaging" the models in our confidence set. There are two common approaches to model averaging. *Natural-model averaging* averages over all candidate models where a parameter of interest occurs. *Full-model averaging* considers all candidate models. If a variable is not selected in one of the candidate models, full-model averaging sets its estimate to zero. Hence, full-model averaging takes into account when a variable has not been selected whereas natural-model averaging does not. Furthermore, simulation studies have found that full-model averaging can help to reduce problems caused by model selection bias towards overly complex models [18]. We use full-model averaging for this reason.

Given a candidate set of R models our full-model averaged coefficients β_h are estimated by

$$\hat{\bar{\beta}}_h = \sum_{i=1}^{R} w_i \hat{\beta}_{hi}, \tag{4}$$

where $\hat{\beta}_{hi}$ is the estimate of β_h based on model g_i. If β_h is not chosen in model g_i then $\hat{\beta}_{hi}$ is defined to equal zero in the above formula.

3.3.3. Unconditional confidence intervals

Once a set of candidate models has been identified we can construct confidence intervals that reflect both parameter and model selection uncertainty. The $(1-\alpha)100\%$ unconditional confidence intervals for the a model averaged coefficient $\hat{\bar{\beta}}_h$ is given by

$$\hat{\bar{\beta}}_h \pm z_{1-\alpha/2} \widehat{\text{ase}}\left(\hat{\bar{\beta}}_h\right), \tag{5}$$

where $\widehat{\text{ase}}\left(\hat{\bar{\beta}}_h\right)$ is the adjusted standard error from Burnham and White [6]. It is given by

$$\widehat{\text{ase}}\left(\hat{\bar{\beta}}_h\right) = \sum_{i=1}^{R} w_i \sqrt{\left(\frac{t_{\text{df}_i,1-\alpha/2}}{z_{1-\alpha/2}}\right)^2 \widehat{\text{var}}\left(\hat{\beta}_{hi}|g_i\right) + \left(\hat{\beta}_{hi} - \hat{\bar{\beta}}_h\right)^2}, \tag{6}$$

where $\bar{\beta}_h$ is the model averaged estimator of β_h, $\widehat{\text{var}}\left(\hat{\beta}_{hi}|g_i\right)$ is the estimated variance of parameter β_{hi} in model g_i, and w_i are weights. The calculation of $\widehat{\text{var}}\left(\hat{\beta}_{hi}|g_i\right)$ for mixed effects models is reasonably complex and is omitted (see Bates et al. [3] for a discussion).

3.4. Estimating power demand profiles

In Eq. (1) we have modelled the log-transform of electricity consumption over 15-min intervals as our response. We use the estimator proposed by Kennedy [16] to calculate the proportional impact, p_h, of X_{hij} on the dependent variable Y_{ij}. Kennedy's estimator is consistent and almost unbiased [10]. For a dummy variable changing from zero to one the estimator is given by

$$\hat{p}_h = e^{\hat{\bar{\beta}}_h - 0.5\widehat{\text{ase}}\left(\hat{\bar{\beta}}_h\right)^2} - 1, \tag{7}$$

where we have replaced the coefficient estimate and variance of Kennedy's original estimator with our full-model averaged counterparts. As this is the proportional impact we can rewrite the above expression in terms of our response variable to obtain

$$Y^*_{ij} = e^{\hat{\bar{\beta}}_h - 0.5\widehat{\text{ase}}\left(\hat{\bar{\beta}}_h\right)^2} \mathbb{E}\left[Y_{ij}|X_{hij}=0\right], \tag{8}$$

where Y^*_{ij} is the new consumption value after our Boolean variable has changed from false to true. Calculating Y^*_{ij} using this formula for each hour of the day gives our power demand profiles. Taking the difference between Y^*_{ij} and $\mathbb{E}\left[Y_{ij}|X_{hij}=0\right]$ gives our estimate of the demand impact profiles. Note that we have chosen to base our profiles off $\mathbb{E}\left[Y_{ij}|X_{hij}=0\right]$ rather than an unconditional mean or median because we are working with proportional impacts. Since we show the impact of a variable switching from false to true it seems reasonable to apply the proportional impact to the mean demand that we observe when an attribute isn't present, hence the use of the conditional expectation. Simply using the mean or median may exaggerate the demand impact profiles.

3.5. Fitting models

We fit our models using the R statistical programming language [26]. The `lme4` package is used to fit our mixed effects model and calculate mAIC scores [4]. Multimodel inference is carried out using the `MuMIn` package [2].

All code used to produce this analysis has been made available at https://github.com/camroach87/1801-MMME https://github.com/camroach87/1801-MMME https://github.com/camroach87/1801-MMME.

4. Discussion

In this section we apply our methodology to assess the impact of building characteristics on electricity demand across the day. We also examine how well our models fit the data and comment on the limitations of our statistical methodology. Several future research directions are put forward.

To validate our approach a simulation study was also conducted using simulated time series data designed to mimic electricity demand in commercial buildings. We found that after simulating electricity consumption for 129 buildings our methodology was able to estimate the coefficients acceptably, with most estimates falling within the 90% confidence intervals. Details of the simulation study are provided in Appendix B.

4.1. Goodness of fit

Prior to analysis of our results it is important to assess if our models actually fit the data acceptably. We use marginal and conditional R^2 values for each season and hourly model to assess the goodness of fit. Adapting the specification of Nakagawa and Schielzeth [22] to our case, the conditional R^2 calculates the proportion of the variance explained by both fixed and random effects and is given by

$$R^2_c = \frac{\sigma^2_f + \sigma^2_u}{\sigma^2_f + \sigma^2_u + \sigma^2_\epsilon}, \tag{9}$$

where $\sigma^2_f = \text{var}\left(\sum_{h=1}^{p} \beta_h x_{hij}\right)$ is the variance of the fixed effects, σ^2_u is the variance of random effects and σ^2_ϵ is the residual variance. As we are working with a random slope model the random effect variance is calculated as described in Johnson [14]. Marginal R^2 considers only fixed effects and is defined as

$$R^2_m = \frac{\sigma^2_f}{\sigma^2_f + \sigma^2_u + \sigma^2_\epsilon}. \tag{10}$$

The goodness of fit statistics are shown in Table 2 and Fig. 5. Overall, it appears that our model formulation gives reasonable fits with the conditional R^2 values averaging close to 0.9. Models close to the start and end of business hours show the worst fits due to the noisiness of data during these times. This likely reflects the different operating schedules for different buildings at these times of the day. Another point to note is our marginal R^2 values are consistently higher during winter business hours compared to other

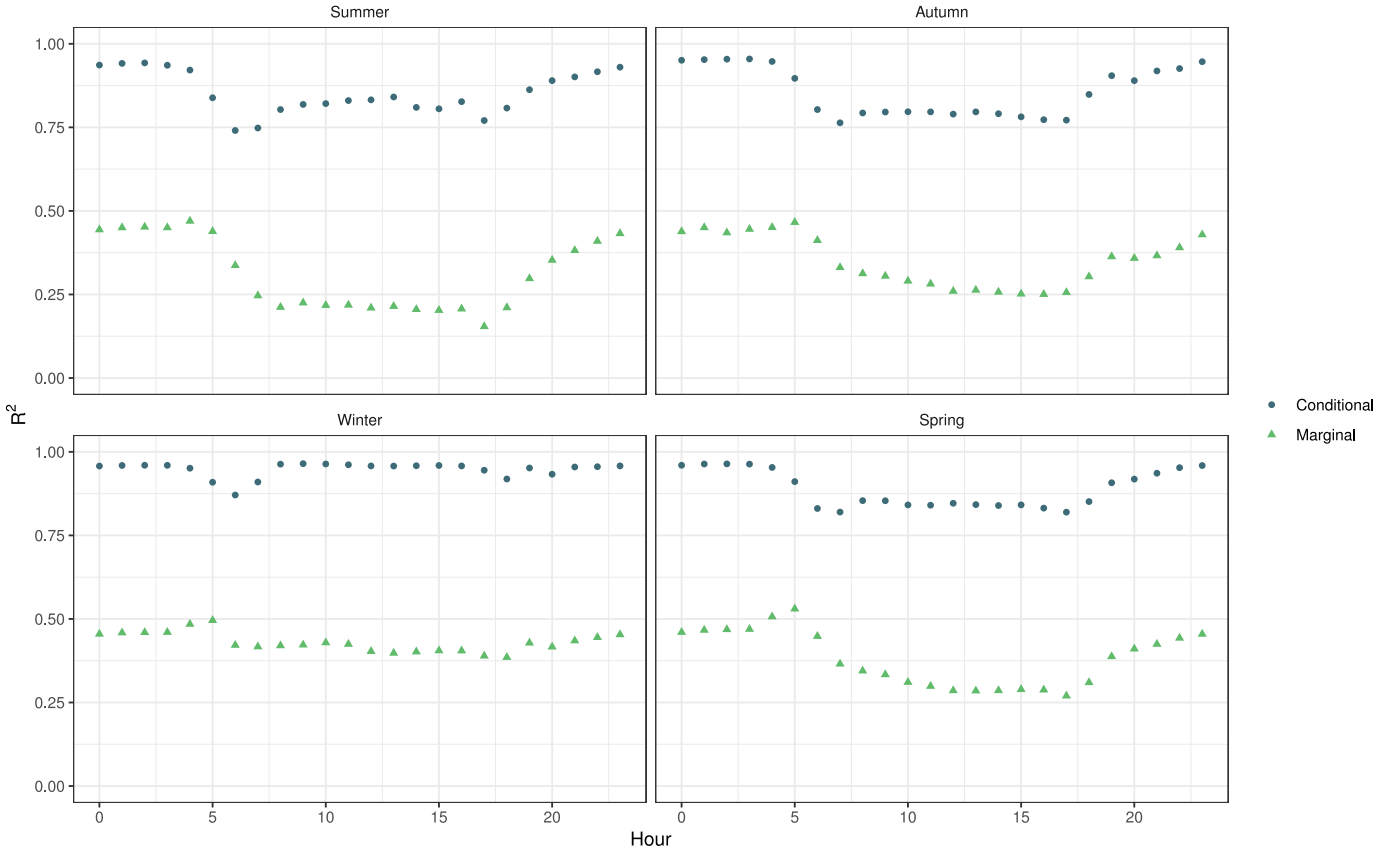

Fig. 5. Conditional and marginal R^2 values for each hour and season's best model after all-subsets selection. The marginal R^2 value represents the variance explained by fixed effects whereas the conditional R^2 value is the variance explained by both fixed and random effects. When we plot these goodness of fit values we see a drop during the working hours which is expected due to the more volatile nature of demand during these times.

Table 2
Conditional and marginal R^2 values for each hour and season's best model after all-subsets selection.

	Summer		Autumn		Winter		Spring	
Hour	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m	R^2_c	R^2_m
0	0.94	0.44	0.95	0.44	0.96	0.45	0.96	0.46
1	0.94	0.45	0.95	0.45	0.96	0.46	0.96	0.47
2	0.94	0.45	0.95	0.43	0.96	0.46	0.96	0.47
3	0.94	0.45	0.95	0.45	0.96	0.46	0.96	0.47
4	0.92	0.47	0.95	0.45	0.95	0.48	0.95	0.51
5	0.84	0.44	0.90	0.47	0.91	0.50	0.91	0.53
6	0.74	0.34	0.80	0.41	0.87	0.42	0.83	0.45
7	0.75	0.25	0.76	0.33	0.91	0.42	0.82	0.37
8	0.80	0.21	0.79	0.31	0.96	0.42	0.85	0.34
9	0.82	0.22	0.80	0.30	0.96	0.42	0.85	0.33
10	0.82	0.22	0.80	0.29	0.96	0.43	0.84	0.31
11	0.83	0.22	0.80	0.28	0.96	0.42	0.84	0.30
12	0.83	0.21	0.79	0.26	0.96	0.40	0.85	0.29
13	0.84	0.21	0.80	0.26	0.96	0.40	0.84	0.29
14	0.81	0.21	0.79	0.26	0.96	0.40	0.84	0.29
15	0.81	0.20	0.78	0.25	0.96	0.41	0.84	0.29
16	0.83	0.21	0.77	0.25	0.96	0.40	0.83	0.29
17	0.77	0.15	0.77	0.26	0.95	0.39	0.82	0.27
18	0.81	0.21	0.85	0.30	0.92	0.39	0.85	0.31
19	0.86	0.30	0.90	0.36	0.95	0.43	0.91	0.39
20	0.89	0.35	0.89	0.36	0.93	0.42	0.92	0.41
21	0.90	0.38	0.92	0.37	0.95	0.44	0.94	0.42
22	0.92	0.41	0.93	0.39	0.96	0.44	0.95	0.44
23	0.93	0.43	0.95	0.43	0.96	0.45	0.96	0.45

seasons. This indicates that the chosen building attributes are better at modelling heating loads than cooling loads.

Marginal R^2 only includes the impact of fixed effects allowing us to assess how much of the variance in electricity consumption is explained by them. This lets us examine the goodness of fit when using building attribute variables only. If fixed effects do not explain any variation in the data our marginal R^2 values will be close to zero. In our case we see marginal R^2 values range between 0.25 and 0.5 whereas the conditional R^2 values range between 0.8 and 0.95. To be able to explain 25–50% of the variance in the data with only a handful of attributes is encouraging and shows that at least some of the building characteristics we are working with have explanatory power. For reference, 0.25-0.5 is not an unreasonable range[2] when compared with ecology studies [22]. Given the building attributes alone are able to account for some of the variation in the data suggests that they do improve the model fit. Once inter-building differences are accounted for by including the random effects we see subjectively good fits based on the conditional R^2. It is recommended that future smart-meter studies using mixed effects models should include both marginal and conditional R^2 values to allow for a discussion on how well their fixed effects model the data. Being able to examine how much variance the fixed effects capture is important to consider and provides interesting information [22]. Should marginal R^2 values ever be close to zero applying our methodology will not prove particularly useful even if high conditional R^2 values are observed, and so it is important to include this information.

We also examined quantile-quantile (QQ) plots to check if model residuals were normally distributed. For the most part the residuals did appear to be normally distributed, though there was some evidence of heavy tails. This was largely due to some remaining outliers in the time series data - sometimes caused by erratic spikes and unexpected drops in demand. Extreme weather

[2] More studies using mixed effects models on smart-meter data need to be conducted before it is fair to conclude that the values presented here are low or high.

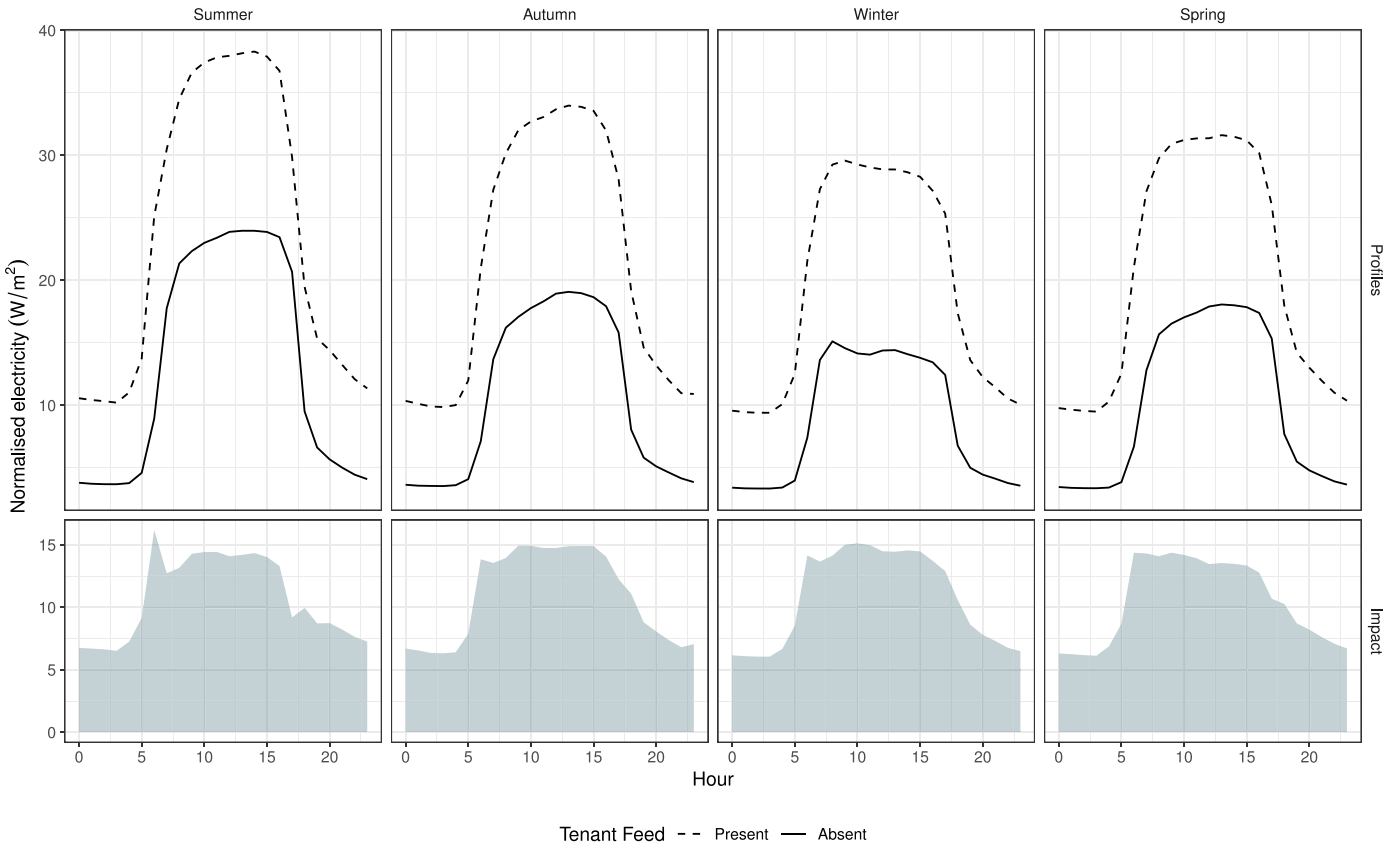

Fig. 6. Profile plots of tenant feed impact. This is an example of a behavioural aspect of electricity demand. Here we can directly assess how tenants typically use electricity during each of the seasons. One notable feature of the plots is the large spike at 6:00 a.m. during summer mornings, possibly caused by precooling of offices. We can not know this for sure without collecting further attribute data on tenant cooling equipment, but does serve to motivate further investigation.

and unmarked local holidays in the data could each be responsible for some of the noise. Including the relationship between temperature and electricity demand would likely improve our fits. However, due to the added complexity[3] and a desire to keep our mixed effects model clear for explanatory purposes we leave this for future research. Considering the noisiness of the data we were satisfied that the residuals were close enough to being normally distributed for our analysis.

4.2. Demand impact profiles

Plots of the power demand profiles and demand impact profiles for tenant consumption, electric element heating and DX systems are shown in Figs. 6–8, respectively. Power demand profiles with and without the attribute in question are shown in the top panels; whereas the bottom panels show the demand impact profiles. These three plots show examples of behaviour impact, caused by occupants; and equipment impact, caused by equipment that directly draws electricity demand. Note that normalised consumption over 15-min intervals was used as the response variable in our mixed effects model. To aid in interpretation of results the fixed effect estimates are converted from units of energy to units of power[4] when plotting.

As we have data for buildings with and without tenant consumption data we can estimate the expected behaviour of tenants (Fig. 6). This is an example of assessing occupant behaviour with our model. As expected we see a large contribution to electricity demand by tenants, ranging from 6 to 15 W/m^2 across all seasons. Of particular interest is the large morning spike we see at 6:00 a.m. in summer, possibly caused by tenants attempting to pre-cool their offices to ensure occupant comfort over the course of the day. Observing tenant behaviour allows us to quantify and view when demand from tenants is occurring. This provides some guidance on how much potential savings there might be (as an upper bound), or how the shape of the profile might be modified through intervention. One useful application of this is to assess if a demand management initiative is successful. Including it as a fixed effect will allow for demand impact profiles to be created. Decision makers may then assess statistically if said initiative was successful or not.

Profiles for electric element heating are shown in Fig. 7. Here we assess a building characteristic that directly uses electricity and where the behaviour is known. As would be expected with heating equipment we see only a small change in demand during summer. Slightly higher demand is seen in autumn and spring, likely due to cold days during the months that border winter. Winter sees a large increase in electricity demand due to heating required during business hours.

As a final example, the subplots in Fig. 8 highlight that buildings with DX systems tend to use more electricity than those without. Given there was a strong negative correlation with chiller systems (see Section 2.2) it seems reasonable to conclude that chiller systems offers a more energy efficient alternative based on this dataset. However, this conclusion should be tempered by inspecting the confidence intervals in Fig. 9. The coefficient estimates are,

[3] The electricity temperature relationship in each building can be quite nonlinear and varies across the day.

[4] This is a simple matter of dimensional analysis where 1 Wh of energy consumed over a duration of 15-min is equivalent to 4 W:

$$\frac{1\ \text{Wh}}{15\ \text{min}} = \frac{1\ \text{Wh}}{1/4\ \text{h}} = 4\ \text{W}.$$

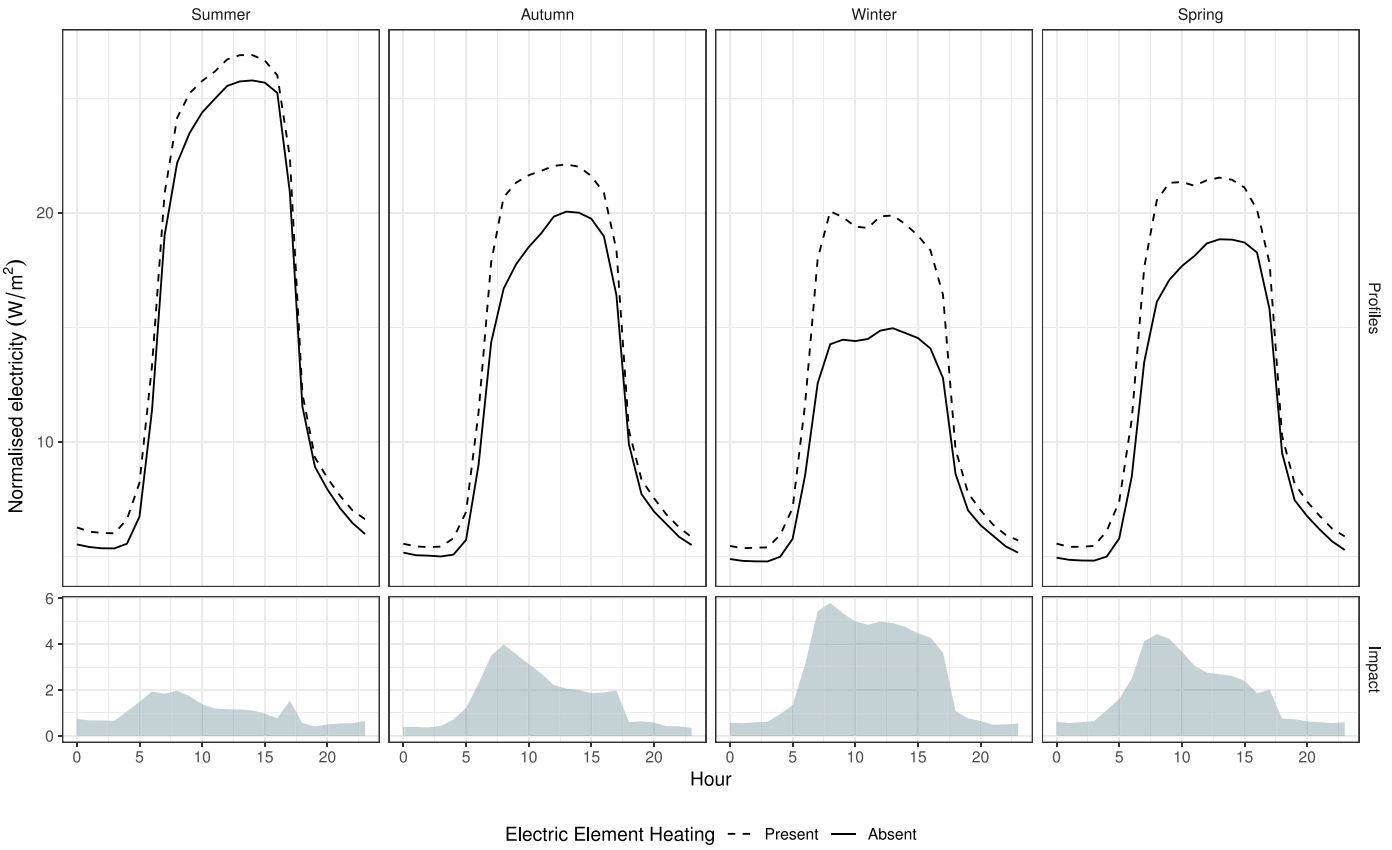

Fig. 7. Profile plots of electric element heating impact. As expected the heating demand mainly plays a role in winter. The demand in autumn and spring is likely caused by cold days during the shoulder months of winter.

for the most part, only significant ($\alpha = 0.1$) during winter. It is possible that this result captures the impact of reverse DX systems being used for heating during winter months.

Centralised distribution and gas fired boilers did not show statistically significant ($\alpha = 0.1$) effects at any time of the year. Water cooled condensers were, for the most part, not statistically significant. Only one model for water cooled condensers (autumn at 5:00 p.m.) had a statistically significant impact, but did not provide any useful or interesting conclusions. For reference, a discussion of the coefficients and confidence intervals of all attributes and seasons is provided in Appendix A.

In summary, for our building stock we can conclude that:

- We do not observe a statistically significant difference between centralised and per-floor distribution systems. Hence, we can not conclude that one is more or less efficient than the other based on this dataset and controlling for the variables we have chosen to model.
- We observe DX systems use more energy than chiller systems during winter, though this is possibly due to reverse DX systems being used for heating.
- Electric element heating increases the expected electricity demand by as much as 6 W/m^2 during the winter. The largest draw occurs in the morning between 7:00 a.m. and 9:00 a.m.
- Gas fired boilers did not make a statistically significant impact on electricity demand. There was an expectation that their presence may result in offsetting electricity demand, but this did not show in the final results. This could perhaps indicate issues with data or building operation.
- Unsurprisingly, tenant feed had a significant impact for all seasons and at all times of the year. Large increases in tenant demand of approximately 8 W/m^2 are seen between the hours of 6:00 a.m. to 4:00 p.m. Tenant demand gently tapers off after 4:00 p.m.during autumn, winter and spring, but has a sharp drop after 4:00 p.m.in summer. There is a large spike in tenant electricity demand at 6:00 a.m. in summer which indicates that there may be opportunities within our building stock to spread out demand at this time to reduce peaks.
- The presence of water cooled condensers in buildings does not appear to have a significant impact on electricity demand.

Note that these findings only apply to our building stock when controlling for the variables we have used in our model. Each conclusion is subject to idiosyncrasies in the data or perhaps be indicative of unusual building behaviour.

Determining the expected consumption of tenants in buildings may allow for targeted energy saving measures. For example, if there are large peaks at certain times of the day, facility managers may make tenants aware of the need to reduce demand during these times. The early morning spike we observe in summer could be deemed one such time. This peak possibly corresponds to early morning cooling to ensure indoor environments are comfortable when tenants arrive. If this was automated to spread cooling across a longer period the peak could be reduced. While this wouldn't necessarily decrease the overall consumption, the lower peak may reduce costs for peak power consumption[5] and be useful in grid management.

Using these results, the company that provided data were able to determine which attributes were important and how heavily they should be weighted when assessing building similarity for electricity demand. This allowed for benchmarking and target set-

[5] Especially if time of use tariffs are in place and we observe peaks during expensive times of the day.

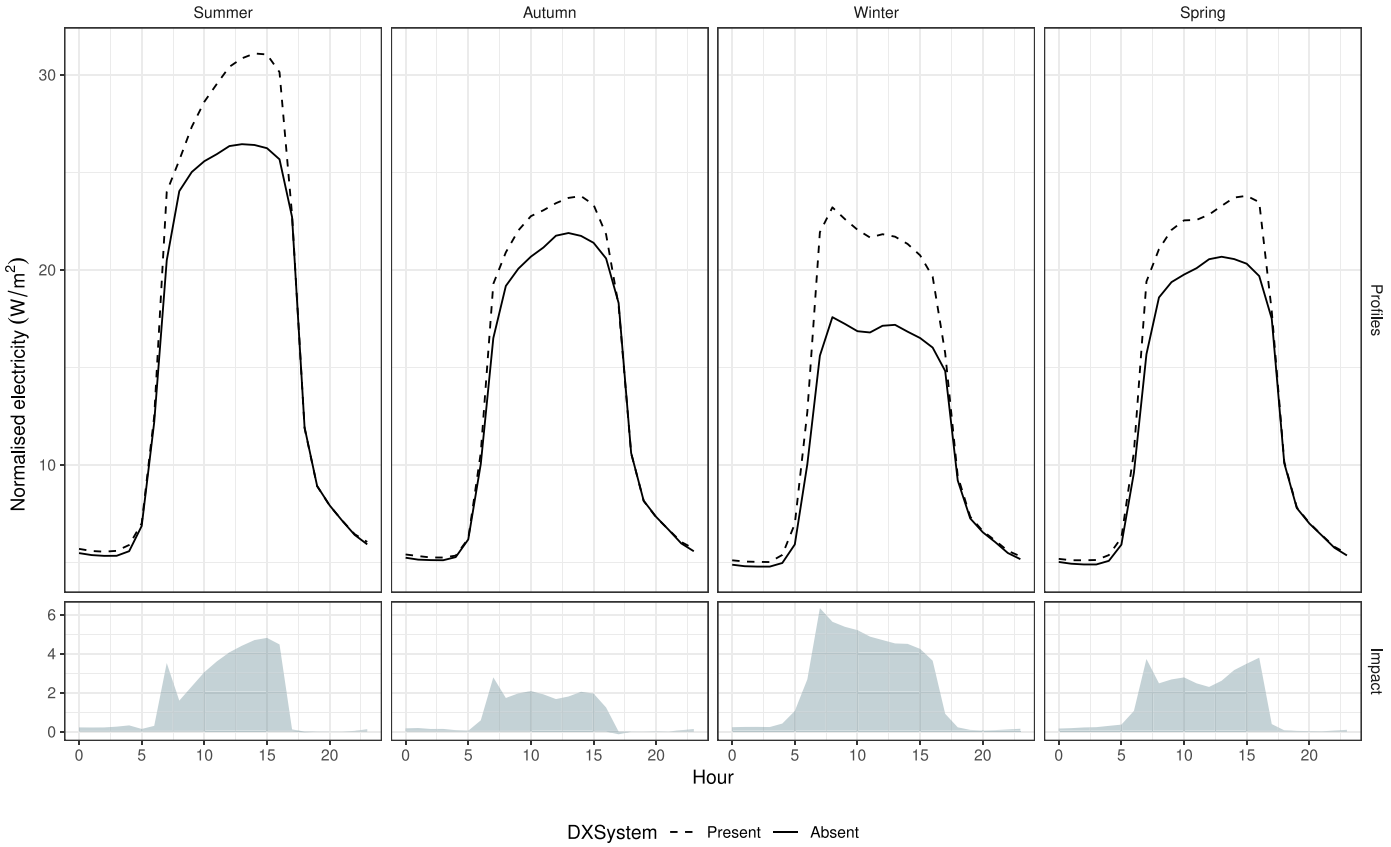

Fig. 8. Profile plots of DX system impact. Buildings fitted with DX systems use more electricity during winter months, which may be representative of reverse DX systems being used for heating. Confidence intervals for the estimated coefficients show the winter results are statistically significant during business hours.

ting for facility managers based not only on a single building's historical demand, but also on other buildings with similar characteristics.

4.3. Modelling limitations

As with any statistical approach there are certain limitations. While we may find that some variables are statistically significant we can not necessarily comment on causality. Confounding variables and idiosyncrasies of a given dataset may be responsible for some results. Furthermore, if a large number of fixed effects are being explored then it is important to have a large number of buildings in the dataset to avoid the curse of dimensionality. Despite these matters, it is still useful to conduct analyses such as this to better understand underlying behaviour and identify possible anomalies that are worthy of further investigation.

One difficulty of drawing conclusions from mixed effects models is that technical data describing installed equipment is conflated with operational practice. For instance, our centralised distribution variable merely indicated if a building had centralised distribution and not the model, size or how it is typically operated. Larger systems tend to be more efficient than smaller ones and centralised systems tend to have lower cooling loads then per-floor systems as they cannot be influenced (easily) by tenants. While it is theoretically possible to work around this by adding new attribute variables, it quickly increases the dimensionality of the model which then requires more buildings when fitting. More building data and detailed attribute descriptions are required to facilitate fine-grain analysis such as this.

When using statistical models any analysis is limited by the quality and breadth of the available data. Survey results or building characteristics are needed for multiple buildings before any models can be fit. It is important to be aware that when statistical findings do not match with expected behaviour further investigation should be carried out.

4.4. Further research

Including more buildings will allow for more confident conclusions to be reached and more characteristics to be investigated. Of particular interest could be applying our approach to the residential sector to quantify the impact of solar generation, batteries or other household items. Given the prevalence of smart meter data available to both government and distributors, only a survey of household goods would be required. One interesting research direction involves assessing the impact of policies on electricity consumption. If separate models are fit for each year, we can observe how the demand impact profiles are changing between each yearly model. If a policy is introduced to reduce demand for a particular building attribute, we should see a decrease in the profiles between the yearly models. This may help us to determine if polices are proving to be effective or not.

Our mixed effects approach may also be adapted to estimate demand profiles for other building portfolios or energy sources. As long as appropriate goodness of fit measures are calculated and the statistical significance of variables are tested, most modelling pitfalls and erroneous conclusions should be avoided. Different settings will require consideration of which building characteristics to include and possible modifications to the formula we presented. However, the key idea of splitting data into buckets across the day (hours in our case) and fitting mixed effects models to each should

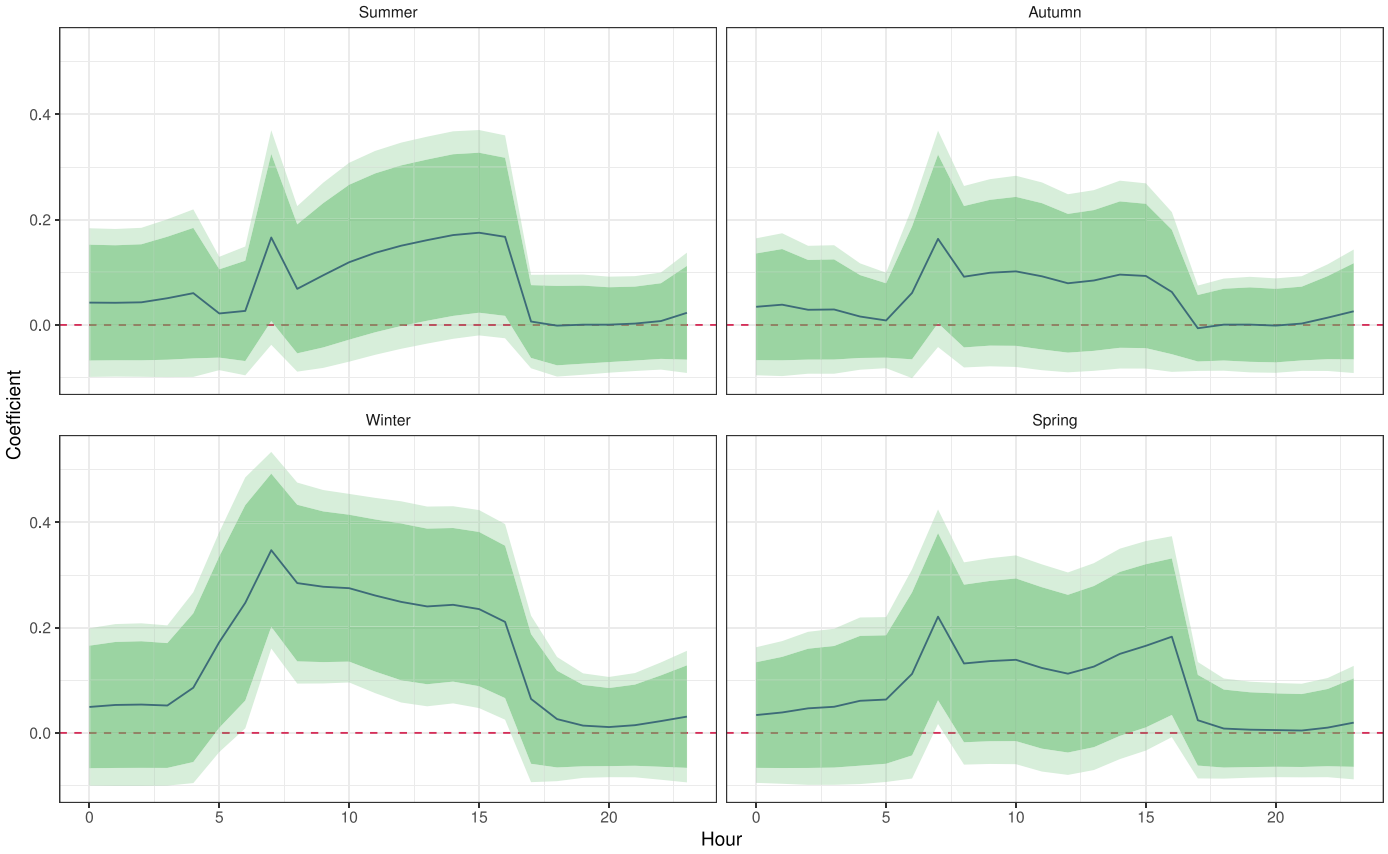

Fig. 9. Coefficient profile plots for DX system coefficients. The solid line shows the estimated coefficients, β_h, for each hourly model described in Eq. (1). The 80% and 90% confidence intervals are indicated by the shaded ribbons. Coefficient estimates for winter business hours are significant.

still be applicable. Doing so is beyond the scope of this paper and is left as a future research direction.

Including the relationship between demand and weather variables may offer further improvements to our mixed effects modelling. Despite pursuing a simple formulation throughout our analysis and only adding a trend term to capture yearly changes in climate, we do expect including weather variables would further improve model fits. Doing so also removes them as potential confounding variables.

5. Conclusion

This paper presents an approach to estimating the impact different building characteristics may have on electricity demand of commercial office buildings. Our approach only requires that smart meter data and attribute data is available for several buildings. The impact that each building characteristic has on energy demand can be presented in the form of demand impact profiles which show the expected change in demand should a characteristic be absent or present. Key advantages of this approach are that it does not require equipment sub-metering and that it can be used to estimate the impact of indirect and behavioural factors.

In addition to providing demand impact profiles, our methodology also produces confidence intervals based on a statistical framework. This allows us to assess the statistical significance of each building characteristic. Mixed effects models are used to account for the correlation within each building's meter readings. A multi-model inference approach, which fits multiple models and weights coefficient estimates by the strength of each model fit, allows us to take both parameter estimation and model selection uncertainty into account when calculating confidence intervals. Hence, we can produce estimates of the effect size and the statistical significance of each characteristic's effect on electricity demand.

To justify the validity of our approach we have conducted a simulation study which shows our model estimates latent variables well. Furthermore, we also describe a case study using 129 commercial office buildings from across Australia. We were provided with several building attributes that engineers wished to assess statistically. Applying our methodology allowed us to understand how tenant behaviour and equipment behaviour affected electricity demand across the day at different times of the year. Furthermore, it also highlighted that some characteristics did not appear to have a statistically significant impact on energy demand.

It is our hope that this approach will open the door to further research in this area. Combining smart meter data from multiple buildings with additional descriptive data sources will allow us to quantify electricity consumption patterns. This has potential applications in benchmarking and assessing the efficacy of demand management programs. Such research can play a role in helping us understand the potential of energy efficient practices by highlighting which equipment or design choices result in lower demand. This in turn may help managers and policy makers to enact better, data-driven decisions.

Acknowledgements

This research project was supported by funding from Buildings Alive (Fund: 1752291; Grant: 512031). I would like to thank Buildings Alive and its CEO Dr. Craig Roussac for making data available and their guidance in understanding commercial building equipment and behaviour. I would also like to thank Professor Rob Hyndman and Dr. Souhaib Ben Taieb for their continued support and

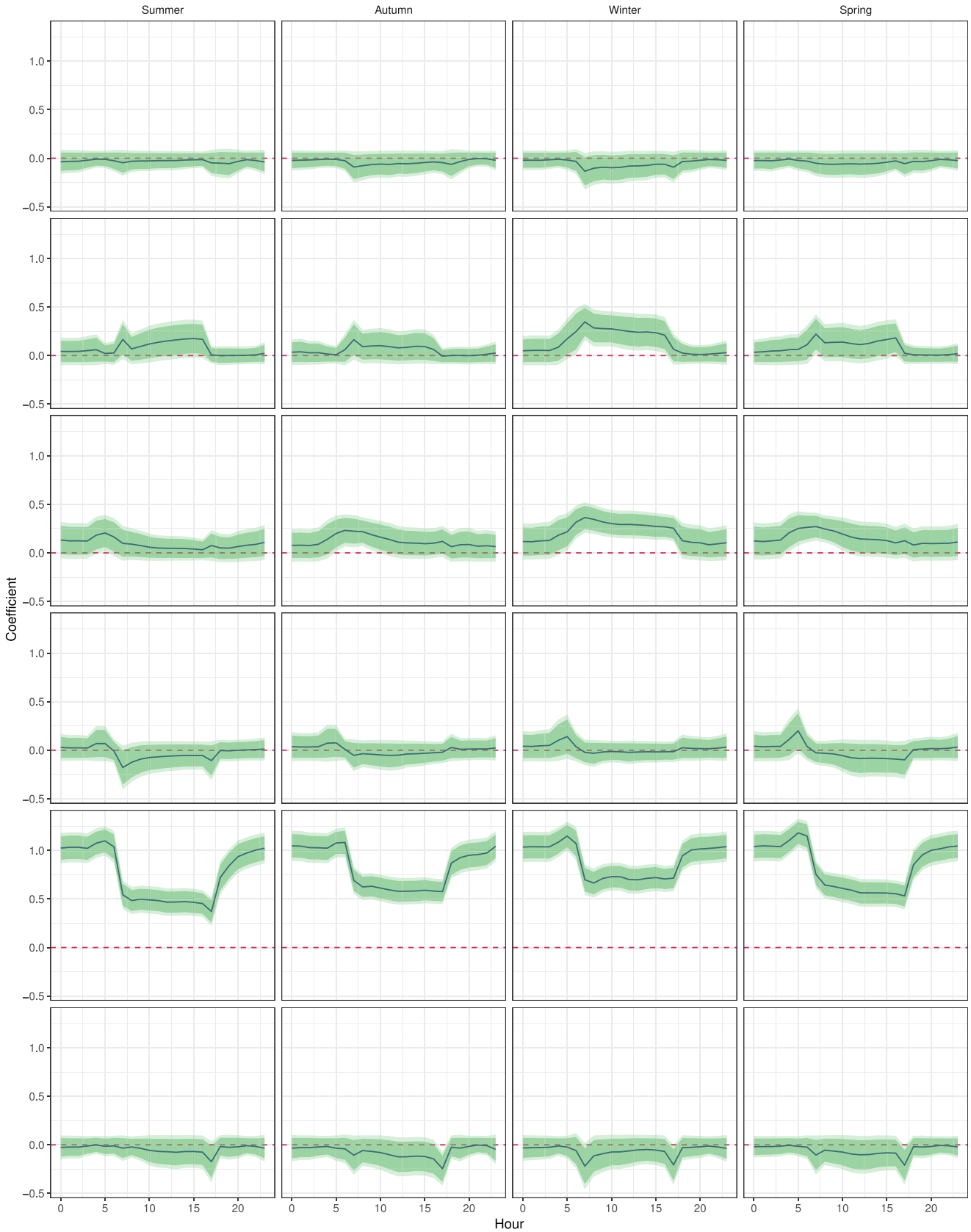

Fig. 10. Coefficient profile plots for all attribute coefficients. The solid line shows the estimated coefficients, β_h, for each hourly model described in Eq. (1). The 80% and 90% confidence intervals are indicated by the shaded ribbons.

advice throughout the research process. Finally, I would like to thank the reviewers for their thorough and constructive feedback.

This research was supported by use of the Nectar Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS).

Appendix A. Profiles of all attributes

Hourly coefficient profiles for each attribute discussed in Section 2.2 are shown in Fig. 10. These are produced using Eq. (1) and the multimodel inference approach. The coefficients represent the proportional change in demand when an attribute is or isn't present. Unconditional confidence intervals are also shown which allow us to gauge how statistically significant each attribute's impact is. Note that these coefficients and confidence intervals are calculated separately for each hour of the day to take the changing building dynamics into account.

We can draw several conclusions from Fig. 10. For example, some of the attributes only have a statistically significant impact at certain times of the year. For example, electric element heating is not significant during summer, but does have a statistically significant impact during business hours in winter. A similar result is observed for DX systems. Interestingly, centralised distribution does not appear to be statistically significant at any time of the year. This indicates that in the buildings used in our analysis, the normalised demand does not appear to be affected by whether centralised or per-floor distribution is present. Water cooled condenser and gas fired boiler variables do not appear to play any statistically significant role in energy demand during any of the seasons. We can always see significant coefficient estimates for tenant feed, which is unsurprising given the amount of energy tenants are using.

Appendix B. Simulation results

To validate our approach we conduct a simulation with known fixed effect coefficient values. If we assume that each of the estimated coefficients shown in Fig. 10 are correct, we can simulate time series for multiple buildings and test if our proposed methodology can correctly estimate these effect sizes.

Using the estimated winter coefficients shown in Fig. 10 we simulate demand for building j at time t using

$$\log y_{tj} = \sum_a X_{ja} A_{ah} + \epsilon_{tj}, \quad \epsilon_{tj} \sim ARMA(1,1), \tag{11}$$

where

- $A_{ah} \sim \mathrm{N}(a_{ah}, \sigma_a)$ is the assumed known impact of attribute a during hour h,
- a_{ah} is the estimated coefficient,
- σ_a is the standard error of the fixed effect coefficients, and
- $X_{ja} \sim \mathrm{Bernoulli}(p_a)$ is a random variable indicating if attribute a is present for building j.

The probabilities that an attribute was present for a building, p_a, were chosen based on the relative frequency of attributes in the 129 commercial office buildings. The expected impact of an attribute is a_{ah} and is assumed to be equal to the estimates we produced in our analysis using real data. The variance σ_a allows for the impact of attribute a to vary between buildings and is chosen to be close to the estimated standard error for the fixed effects. Most fixed effects had standard error close to 0.3 and so for simplicity we chose $\sigma_a = 0.3$ for all attributes. As we are working with time series data we allow for autocorrelation in the model errors by treating them as an $ARMA(1,1)$ process with $\phi_1 = 0.6$, $\theta_1 = 0.6$ and $\sigma_\epsilon = 0.05$. Doing so produced residuals with similar standard

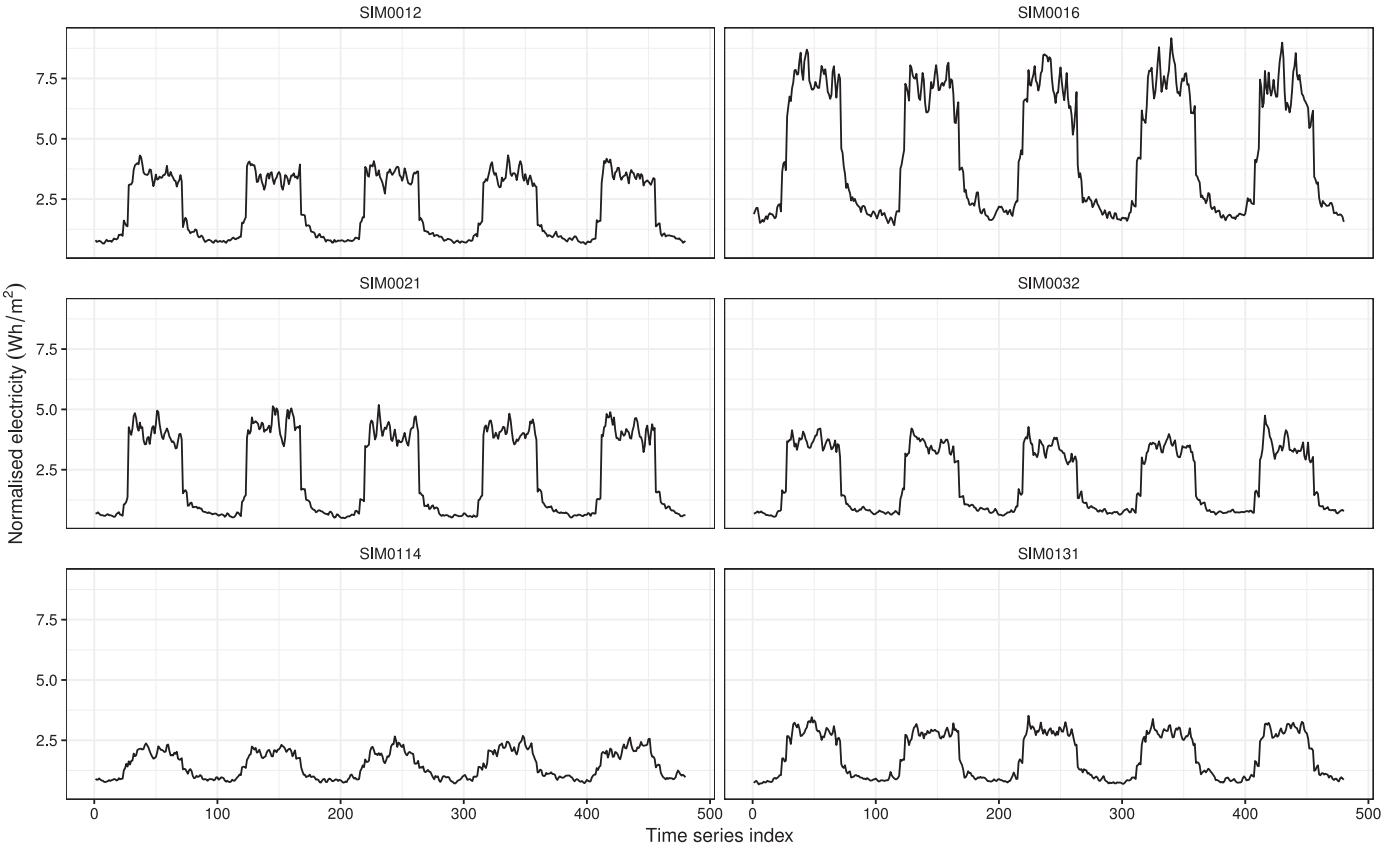

Fig. 11. Time series plots of simulated smart-meter data for six of the 150 simulations. As observed with the real data, each building has distinct business hour and non-business hour behaviour and the load profiles differ between each building.

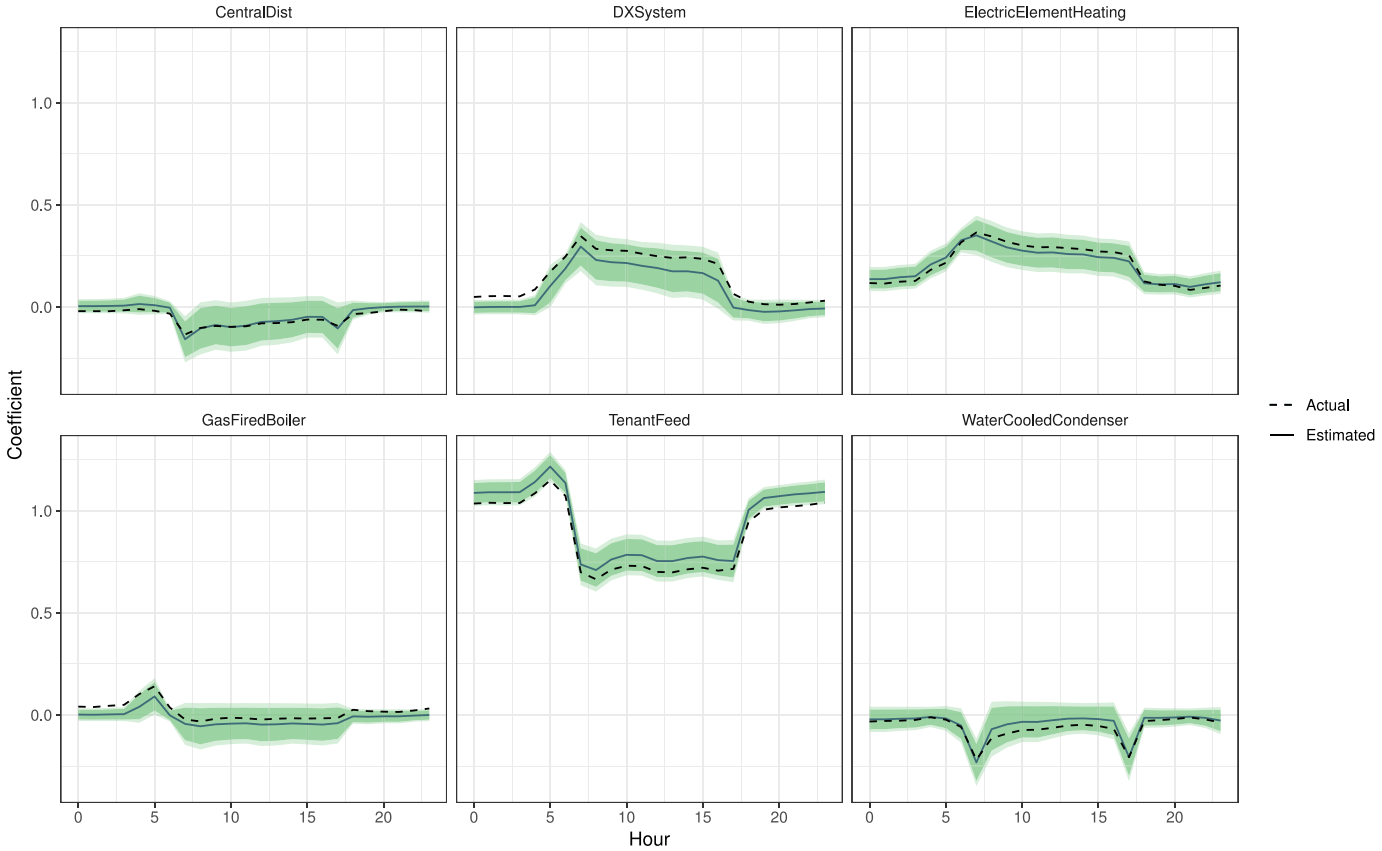

Fig. 12. Coefficient profile plots based on simulated data for all attribute coefficients. The solid line shows the estimated coefficients for each hourly model and the dashed line shows the known fixed effect values that were chosen. 80% and 90% confidence intervals are indicated by the shaded ribbons. We see that our methodology is able to provide good estimates of the fixed effects.

deviation to those observed in the real data. As a final check, we plotted the simulated time series as shown in Fig. 11 to ensure that they looked reasonable. The simulated time series appear to be realistic when compared to the time series in Fig. 3.

We simulated 90 days of 15-min demand data for 150 buildings using the estimated profiles from winter. Applying our mixed effects multimodel inference approach to this data produced the fixed effect estimates and confidence intervals shown in Fig. 12. We see that our estimates and confidence intervals perform quite well in estimating the actual values.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.enbuild.2019.109686.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: Selected Papers of Hirotugu Akaike, Springer Series in Statistics, Springer, New York, NY, 1998, pp. 199–213. https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15.

[2] K. Bartoń, MuMIn: Multi-Model Inference, 2019. R package version 1.43.3/r450. https://R-Forge.R-project.org/projects/mumin/.

[3] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear Mixed-Effects models using lme4, Journal of Statistical Software, Articles 67 (1) (2015) 1–48. https://www.jstatsoft.org/v067/i01.

[4] D. Bates, M. Maechler, B. Bolker, S. Walker, lme4: Linear Mixed-Effects Models using 'Eigen' and S4, 2019. R package version 1.1-21, https://CRAN.R-project.org/package=lme4.

[5] K.P. Burnham, D.R. Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Springer Science & Business Media, 2003. https://market.android.com/details?id=book-fT1Iu-h6E-oC.

[6] K.P. Burnham, G.C. White, Evaluation of some random effects methodology applicable to bird ringing data, Journal of applied statistics 29 (1-4) (2002) 245–264. https://doi.org/10.1080/02664760120108755.

[7] K. Chahine, K.E.K. Drissi, C. Pasquier, K. Kerroum, C. Faure, T. Jouannet, M. Michou, Electric load disaggregation in smart metering using a novel feature extraction method and supervised classification, Energy Procedia 6 (2011) 627–632. http://www.sciencedirect.com/science/article/pii/S1876610211014834.

[8] N. Christiansen, M. Kaltschmitt, F. Dzukowski, F. Isensee, Electricity consumption of medical plug loads in hospital laboratories: Identification, evaluation, prediction and verification, Energy andBuildings 107 (2015) 392–406. http://www.sciencedirect.com/science/article/pii/S0378778815302073.

[9] C. Dinesh, B.W. Nettasinghe, R.I. Godaliyadda, M.P.B. Ekanayake, J. Ekanayake, J.V. Wijayakulasooriya, Residential appliance identification based on spectral information of low frequency smart meter measurements, IEEE Transactions on Smart Grid 7 (6) (2016) 2781–2792. https://doi.org/10.1109/TSG.2015.2484258.

[10] D.E. Giles, Interpreting Dummy Variables in Semi-Logarithmic Regression Models: Exact Distributional Results, University of Victoria Department of Economics Econometrics Working Paper 1101, 2011. http://web.uvic.ca/~dgiles/downloads/working_papers/ewp1101.pdf.

[11] S. Greven, T. Kneib, On the behaviour of marginal and conditional AIC in linear mixed models, Biometrika 97 (4) (2010) 773–789. https://academic.oup.com/biomet/article-abstract/97/4/773/241321.

[12] H.B. Gunay, W. O'Brien, I. Beausoleil-Morrison, S. Gilani, Modeling plug-in equipment load patterns in private office spaces, Energy Build. 121 (2016) 234–249. http://www.sciencedirect.com/science/article/pii/S0378778816301268.

[13] Z. Guo, Z.J. Wang, A. Kashani, Home appliance load modeling from aggregated smart meter data, IEEE Transactions on Power Systems 30 (1) (2015) 254–262. https://doi.org/10.1109/TPWRS.2014.2327041.

[14] P.C. Johnson, Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models, Methods inEcology and Evolution / British Ecological Society 5 (9) (2014) 944–946. https://doi.org/10.1111/2041-210X.12225.

[15] B. Kalluri, A. Kamilaris, S. Kondepudi, H.W. Kua, K.W. Tham, Applicability of using time series subsequences to study office plug load appliances, Energy Build. 127 (2016) 399–410. http://www.sciencedirect.com/science/article/pii/S0378778816304546.

[16] P. Kennedy, Estimation with correctly interpreted dummy variables in semilogarithmic equations, The American economic review 71 (4) (1981). https://EconPapers.repec.org/RePEc:aea:aecrev:v:71:y:1981:i:4:p:801.

[17] J. Liao, G. Elafoudi, L. Stankovic, V. Stankovic, Non-intrusive appliance load monitoring using low-resolution smart meter data, in: 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), 2014, pp. 535–540. https://doi.org/10.1109/SmartGridComm.2014.7007702.

[18] P.M. Lukacs, K.P. Burnham, D.R. Anderson, Model selection bias and Freedman's paradox, Annals of the Institute of Statistical Mathematics 62 (1) (2010) 117. https://link.springer.com/article/10.1007/s10463-009-0234-4.

[19] A. Mahdavi, F. Tahmasebi, M. Kayalar, Prediction of plug loads in office buildings: Simplified and probabilistic methods, Energy Build. 129 (2016) 322–329. http://www.sciencedirect.com/science/article/pii/S0378778816307071.

[20] A.C. Menezes, A. Cripps, R.A. Buswell, J. Wright, D. Bouchlaghem, Estimating the energy consumption and power demand of small power equipment in office buildings, Energy Build. 75 (2014) 199–209. http://www.sciencedirect.com/science/article/pii/S0378778814001224.

[21] S. Müller J.L. Scealy A.H. Welsh Model Selection in Linear Mixed Models Statistical Science 28 2013 135 167 URL: https://projecteuclid.org/euclid.ss/1369147909.

[22] S. Nakagawa, H. Schielzeth, A general and simple method for obtaining R2 from generalized linear mixed-effects models, Methods in Ecology and Evolution / British Ecological Society (2013). http://onlinelibrary.wiley.com/doi/10.1111/j.2041-210x.2012.00261.x/full.

[23] L.K. Norford, S.B. Leeb, Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms, Energy Build. 24 (1996) 5l–64. https://pdfs.semanticscholar.org/d6b1/43ad1c53a9c5263b62e5534ef899425a23bd.pdf.

[24] K.X. Perez, W.J. Cole, J.D. Rhodes, A. Ondeck, M. Webber, M. Baldea, T.F. Edgar, Nonintrusive disaggregation of residential air-conditioning loads from sub-hourly smart meter data, Energy Build. 81 (2014) 316–325. http://www.sciencedirect.com/science/article/pii/S0378778814005131.

[25] J.C. Pinheiro, D.M. Bates, Mixed-Effects Models in S and S-plus, Statistics and Computing, Springer, 1978.

[26] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2018.

[27] A. Reinhardt, P. Baumann, D. Burgstahler, M. Hollick, H. Chonov, M. Werner, R. Steinmetz, On the accuracy of appliance identification based on distributed load metering data, in: 2012 Sustainable Internet and ICT for Sustainability (SustainIT), 2012, pp. 1–9. https://ieeexplore.ieee.org/abstract/document/6388037/.

[28] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (2) (1978) 461–464. https://projecteuclid.org/euclid.aos/1176344136.

[29] N. So, R. Richman, A high level method to disaggregate electricity for cluster-metered buildings, Energy Build. 111 (2016) 351–368. http://www.sciencedirect.com/science/article/pii/S0378778815303844.

[30] M.R.E. Symonds, A. Moussalli, A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion, Behavioral Ecology and Sociobiology 65 (1) (2011) 13–21. https://link.springer.com/article/10.1007/s00265-010-1037-6.

[31] F. Vaida, S. Blanchard, Conditional akaike information for mixed-effects models, Biometrika 92 (2) (2005) 351–370. https://academic.oup.com/biomet/article709abstract/92/2/351/233128.

[32] M. Weiss, A. Helfenstein, F. Mattern, T. Staake, Leveraging smart meter data to recognize home appliances, in: 2012 IEEE International Conference on Pervasive Computing and Communications, 2012, pp. 190–197. ieeexplore.ieee.org, https://doi.org/10.1109/PerCom.2012.6199866.

Chapter 4

Subject-specific curves for time series forecasting of smart meter demand

The following paper was submitted to the *Journal of Forecasting*.

All code to reproduce the paper will be made available at https://github.com/camroach87/1901-sscts after publication.

Subject-specific curves for time series forecasting of smart meter demand

Cameron Roach
Monash University
Email: cameron.roach@monash.edu
Corresponding author

Rob Hyndman
Monash University

Souhaib Ben Taieb
University of Mons

6 April 2020

JEL classification: C10,C14,C52

Subject-specific curves for time series forecasting of smart meter demand

Abstract

Buildings are typically equipped with smart meters to measure electricity demand at regular intervals. Smart meter data for a single building have many uses, such as forecasting and assessing overall building performance. However, when data are available from multiple buildings, there are additional applications that are rarely explored. For instance, we can explore how different building characteristics influence energy demand. If each building is treated as a random effect and building characteristics are handled as fixed effects, a mixed effects model can be used to estimate how characteristics affect energy usage. In this paper we demonstrate that producing one-day ahead demand predictions for 123 commercial office buildings using mixed models can improve forecasting accuracy. We experiment with random intercept, random intercept and slope, and subject-specific curve mixed models. The predictive performance of the mixed effects models are tested against naive, linear and nonlinear benchmark models fitted to each building separately. Having justified the use of a mixed model framework, we provide an example showing how mixed model frameworks can, when combined with smart meter data and building attributes, be used to carry out scenario analysis. We demonstrate how expected electricity consumption may increase or decrease given a change in building attributes. This research justifies using mixed models to improve forecasting accuracy and to quantify changes in energy consumption under different building configuration scenarios.

Keywords: time series forecasting, mixed-effects models, smart meters, energy, electricity

1 Introduction

Several papers have examined forecasting electricity demand for buildings by fitting separate models to each building (Ghofrani et al. 2011; Gajowniczek & Ząbkowski 2014; Arora & Taylor 2016; Ben Taieb et al. 2016). While some have attempted to improve forecasts by leveraging the hierarchical nature of electricity demand (Ben Taieb, Taylor & Hyndman 2020; Ben Taieb et al.

2017) few, if any, have explored improving forecast accuracy using a mixed effects framework. If buildings behave in a similar manner a well-specified mixed model may produce more accurate forecasts than individual models. Furthermore, a mixed-model framework allows us to quantify differences between buildings which would not otherwise be possible when using a "building-specific" modelling approach. A mixed effects approach opens the door to scenario analyses by allowing us to estimate how demand might change under different equipment or usage scenarios.

This paper explores how electricity forecasting accuracy can be improved by using mixed effects models. We examine if mixed models can produce forecasts as accurately as separate models fit for each subject. We approach the problem in the context of producing one-day ahead forecasts of electricity demand for 123 commercial office buildings in Australia. When working with mixed effects models, each building is treated as a random effect and building characteristics are treated as fixed effects. We attempt to model the relationship between temperature and demand using both linear and spline based methods.

To the author's knowledge few papers have explored using mixed models in an electricity demand forecasting role. Brabec et al. (2008) appears to be closest to this area. In their paper, a nonlinear mixed effects model (NLME) was used to forecast daily gas demand for individual customers. Predictors such as day of week and temperature were treated as random effects. Their NLME model was benchmarked against ARIMAX and ARX approaches. The paper concluded by saying there was no clear winner between the NLME and benchmark models and that both potentially have strengths and weaknesses. Unfortunately, there are few other papers within the energy field that use mixed effects models[1] for forecasting.

Moving away from the energy sector there are more papers to draw from. Ibrahim & L'Ecuyer (2013) compared the performance of fixed effects and mixed effects models when forecasting call center arrivals. Making use of correlation structures within the data was shown to improve forecast accuracy when tested against several benchmark models on real-world data sets. Frees & Miller (2004) explored lottery sales forecasting by postcode using a linear mixed model applied to longitudinal data. They derived best linear unbiased predictors for what they termed longitudinal data mixed models. Random effects were incorporated for each subject and, separately, each time period. When compared against an ordinary regression model (with common intercept between all subjects) and a basic fixed effects model (with a different intercept for each subject), both with $AR(1)$ error structures, the mixed model that used both time and

[1]Some papers claim to use mixed models. However, this term is often applied to cases where a combination of models have been used which is different to mixed effects models in the statistical sense.

subject random effects (two-way error model) was found to be inferior when forecasting on an out-of-sample test set. However, another one-way error components model that only included treated subjects as random effects was found to produce the best forecasts overall. This suggests that mixed models can compete with ordinary *pooled* regression models. However, the question remains as to how well a mixed model would perform when compared to ordinary regression models fit *separately* to each subject. Another paper that focused on call center forecasting (Aldor-Noiman, Feigin & Mandelbaum 2009) used a mixed Poisson process to estimate future arrival counts. Soyer & Tarimcilar (2008) had a similar aim and showed that a Bayesian approach incorporating random effects was superior to a fixed effects model.

These papers all point to the viability of using mixed effects models for forecasting. None explored the possibility of conducting scenario analysis by varying the fixed effects within a model. This is surprising as quantifying the impact of different characteristics between subjects is one of the obvious advantages of moving to a mixed effects framework. Our paper gives a simple illustration of how this may be done.

Few papers have attempted to assess the impact of differences in building characteristics using statistical methods and smart meter data. To the author's knowledge, only a previous paper by Roach (2020) has looked into this using mixed effects models. Whereas that paper focused on estimating demand impact profiles for building attributes at different times of the year, this paper focuses on improving forecast accuracy and producing scenario analyses by estimating the expected change in time series conditional on the building attributes.

Several papers have shown the relationship between electricity demand and temperature are well modelled using nonparametric components such as cubic splines (Hyndman & Fan 2010; Fan & Hyndman 2012). This paper uses a similar approach within a mixed model framework. Other papers that explore semiparametric mixed models include Grajeda et al. (2016); Ugarte et al. (2009); and Durbán et al. (2005). Durbán et al. (2005) is of particular note as it introduces the concept of subject-specific curves using piecewise linear splines for longitudinal data. We build on the idea of subject-specific curves by applying them to time series data and incorporating natural splines.

The main contribution of this paper is to present an approach to forecasting electricity demand for individual buildings using a mixed effects framework. Furthermore, we show how such a model can be used to conduct scenario analysis allowing us to quantify expected energy savings given changes in building attributes. Finally, this paper serves to enrich the literature on forecasting with mixed effects models and smart meter data.

The paper is structured as follows. Section 2 describes the data we are working with. Section 3 gives a detailed description of the models formulations and how they are assessed. Forecasting results and a scenario analysis example are presented in Section 4. Concluding remarks are given in Section 5.

2 Data

We have time series and attribute data for 123 commercial office buildings located across Australia. We focus on business days in our analysis as these are significantly more important than non-business days for energy management. Non-business days typically have far less demand than business days as equipment is non-operational. Note that our approach can be applied to non-business days as well.

2.1 Time series data

Smart meter data recorded at 15-minute intervals for 123 buildings are used when training and validating our models. The electricity demand is normalised by each building's net lettable area (NLA) to ensure demand is comparable between buildings. An example of a day of smart meter readings from six buildings is shown in Figure 1. Temperature data recorded at 15-minute intervals from the closest available weather station are also available for each building.

The relationship between current temperature and electricity demand is shown in Figure 2 for two buildings at midday and midnight. There is a clear difference in this relationship between the business and non-business periods. Furthermore, this relationship varies between each building. Note that when splines are fitted independently for each building they can be very sensitive to outliers or data at domain extremes. This partly motivates our use of mixed effects models where data from multiple buildings can be used to fit these relationships resulting in models that are more robust to outliers or sparse data at extremes.

One issue that arises from only including business days is the gap between Friday and Monday in our demand time series. As we are including one-day lagged demand as one of our predictors, observed values on Monday will use Friday demand values as their one-day lagged demand (and similarly for two-day lagged demand variables). If this weren't done and we instead used non-working days, a very different relationship between Monday's lagged demand variables and current demand would be observed compared to other weekdays.

Unlike lagged demand variables, which are used to capture operational changes in a building, lagged temperature variables are used to capture thermal inertia. Thermal inertia is residual heat energy that remains in a building after a run of warm weather (or conversely for cold weather). For example, if several warm days occur sequentially, the expected demand can increase as more cooling is typically required to maintain indoor environment quality. Hence, lagged temperature variables are based off all days - not just working days. Monday's lagged temperature variables will include temperatures observed over the weekend.

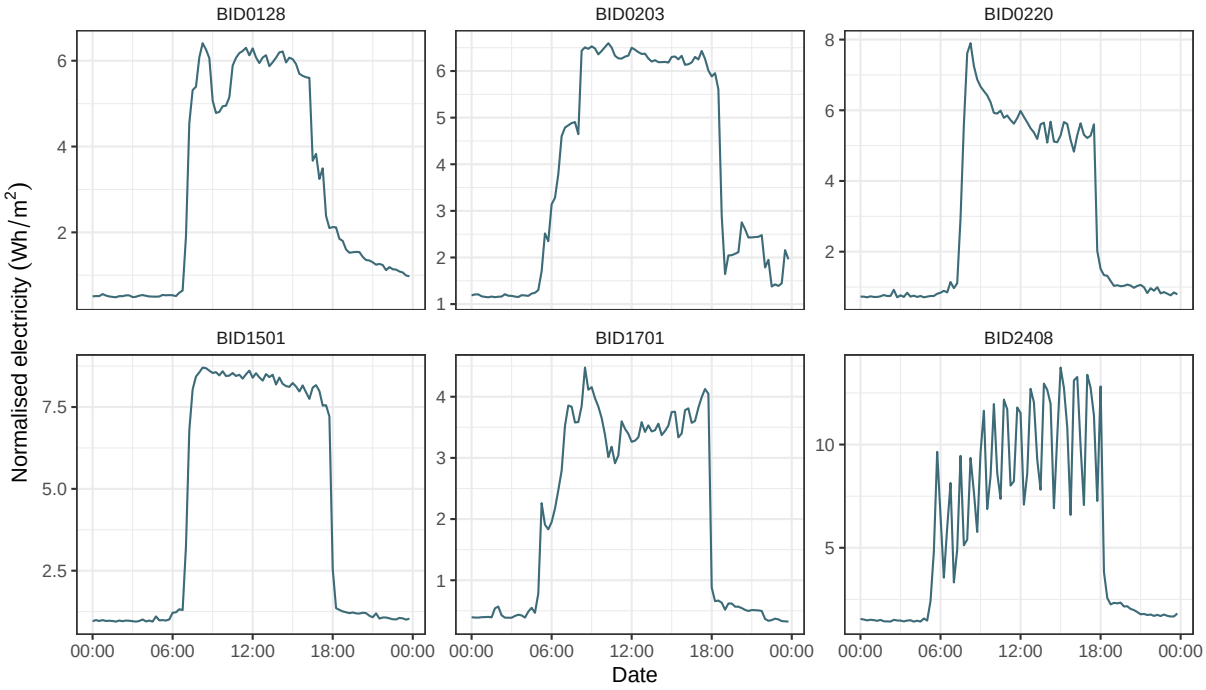

Figure 1: *Normalised electricity demand of six commercial office buildings in Australia on 9 January, 2017. Only one day of data is shown although it is enough to see clear differences in the demand profiles. Both the magnitude and volatility of demand varies greatly between buildings.*

2.2 Attribute data

Building attribute data describes different characteristics of each building. The data are Boolean and indicate if a particular attribute is absent or present. A previous paper (Roach 2020) that examined important drivers of commercial office building demand identified the following attributes as relevant:

- tenant feed
- DX system
- electric element heating
- centralised distribution.

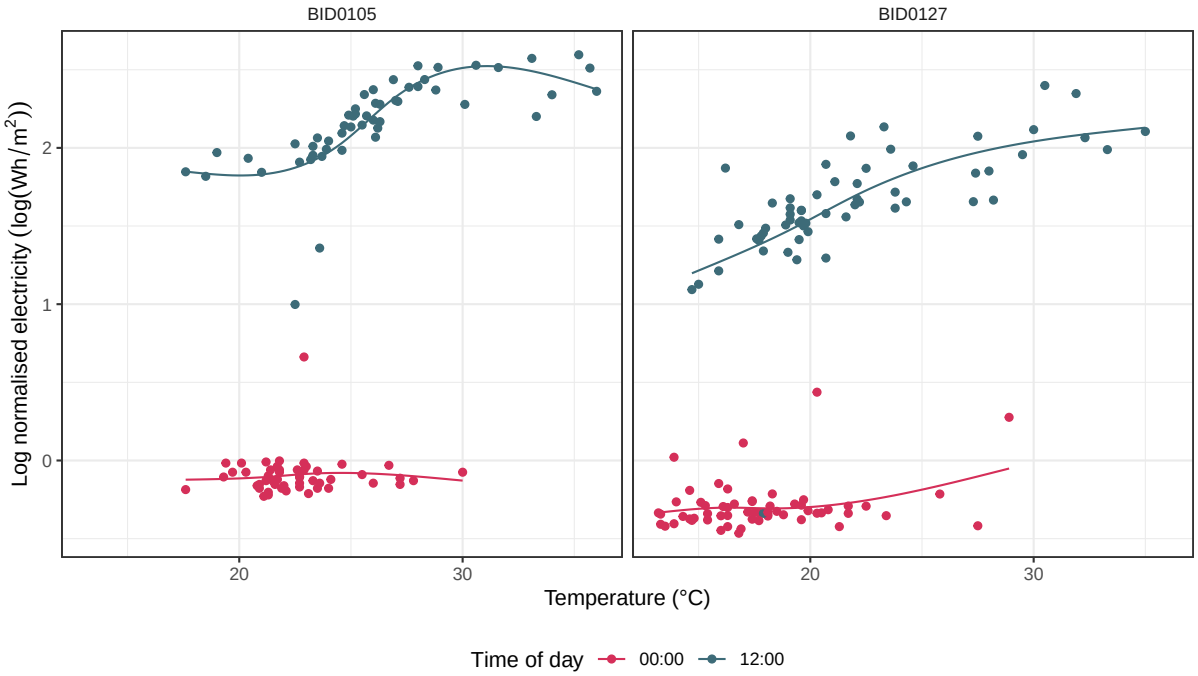

Figure 2: *Relationship between temperature and normalised electricity demand at midday and midnight for two Australian office buildings during Summer. Logged values are shown as we use these as our response variable when fitting models to enforce a positivity constraint. The relationship between temperature and demand is different both between buildings and at different times of the day.*

For a more detailed discussion of each of these attributes refer to Roach (2020).

2.3 Modelling the temperature electricity relationship

The relationship between temperature and electricity demand is shown in Figure 3. The relationship is modelled using three approaches: a linear spline, cubic spline and natural spline. All are fitted with three degrees of freedom. We can see that using the linear spline doesn't capture the smooth sigmoid like curve of the relationship whereas the cubic spline appears to extrapolate in an unrealistic manner at the extreme temperatures. The natural spline has less pronounced movement near the extremes due to the linearity constraint and so it is preferred over the other two methods. Hence, each weather variable and lagged weather variable is modelled using a natural spline with three degrees of freedom. Knots were placed at the 33rd and 67th quantiles.

3 Methodology

Several linear and mixed effects models were tested to determine which produced the most accurate forecasts conditional on selected features. Here we describe the various benchmarking

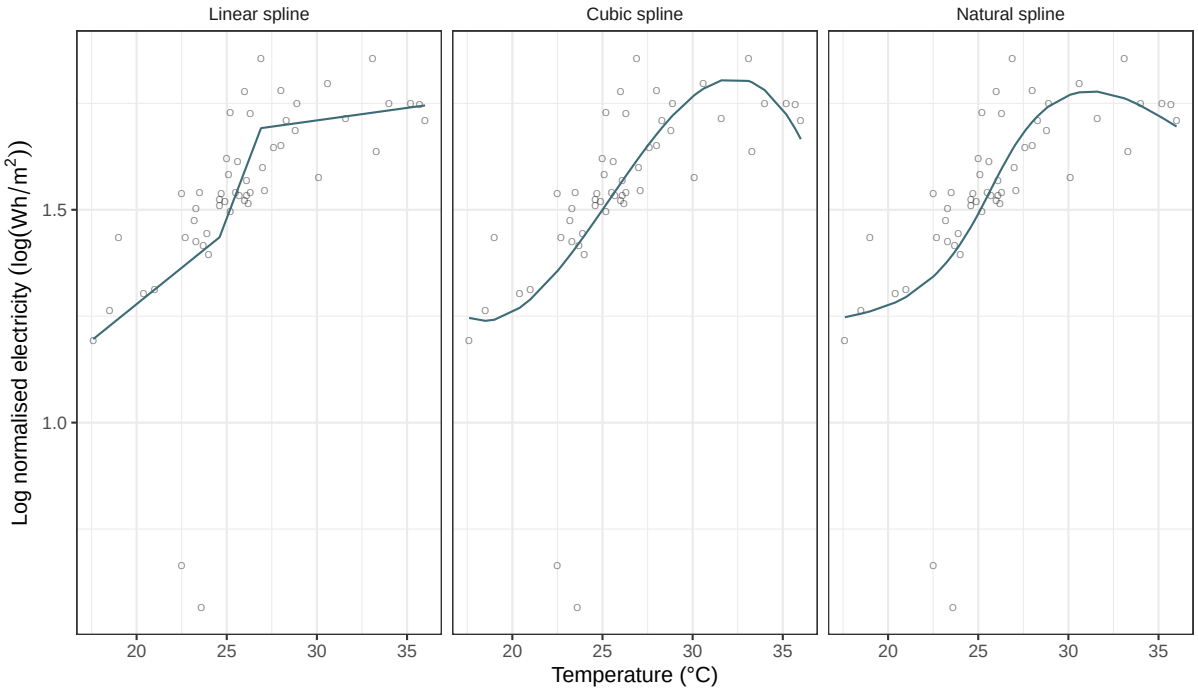

Figure 3: *Linear and cubic splines with three degrees of freedom fit to one building's demand data during Summer at 11:45 am. We observe that the linear spline results in severe kinks in the relationship which seems unrealistic. The cubic spline gives a much smoother fit and appears to be the more reasonable option. The natural spline is better again as it does not have the dramatic dip in predicted demand that the cubic spline has for high temperatures.*

and mixed effects models and their formulations. All analysis was produced using the R statistical programming language (R Core Team 2019). Mixed effects models were fit using the `lme4` package for mixed effects models (Bates et al. 2019).

3.1 Model formulation

To justify our final model that we use for scenario analysis we test several models that can be thought of as simpler versions. Each model has a change introduced and the improvement in performance is used as justification for each. As a starting benchmark, we fit a naive model that uses the previous day's observed values. The second model fit individual linear regression models to each building. The third model also involves fitting a set of individual models but incorporate natural splines to model the temperature and demand relationships. The fourth model is the first to be trained using data from all the buildings and treats each building as a dummy variable. The fifth, sixth and seventh models are random intercept, random slope and subject-specific curve models. Finally, the eighth model is a subject-specific curve model that includes building attributes as fixed effects. This is summarised in Table 1.

Table 1: *Model descriptions*

Model	Abbreviation	Description	Predictor Variables
Naive	Naive	Naive forecasting model	Previous day's demand
Individual linear regression	ILR	Linear regression models fit to each building	Current temperature
Individual natural splines	INS	Natural spline models fit to each building	Current temperature
Pooled regression	PR	Regression model. Used for feature selection	Current temperature, selected features
Random intercept	RI	Mixed effects model with random intercept	Current temperature, selected features
Random intercept and slope	RIS	Mixed effects model with random intercept and slope	Current temperature, selected features
Subject-specific curves	SSC	Mixed effects model with subject-specific curves	Current temperature, selected features
Subject-specific curves with attributes	SSCATTR	SSC with building attributes included as fixed effects	Current temperature, selected features

Due to the evolving nature of energy demand across the day we fit separate models for each 15-minute period of the day. This gives 96 models for each building when fitting individual model formulations and 96 models for each mixed effects formulation.

Throughout our modelling we use natural splines to model the relationship between predictor variables and demand. This differs somewhat from other studies on semiparametric mixed effects models which use piecewise linear splines to model variable relationships (Durbán et al. 2005). However, when we inspect Figure 3 we see that a natural spline gives a more reasonable fit at the sparsely populated extremes when dealing with temperature data.

Predictors are centered and scaled prior to training models. The exact features that are used for models are determined through our feature selection approach (Section 3.2). Models are fit by maximising the log-likelihood criterion.

3.1.1 Individual models

Individual models serve as benchmark models to determine if moving to a mixed models framework improves prediction accuracy. Separate models are fit for each building. Note that a subscript for building has been omitted from each of these individual models to improve clarity.

Naive forecast model The simplest benchmark is a naive forecasting model, where the previous day's values are used. This is often a surprisingly effective forecasting approach (Hyndman & Athanasopoulos 2018). The demand of a building at time t is given by

$$y_t = y_{t-24\text{ hours}} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2).$$

Note that since we have restricted ourselves to business days, $t - 24$ hours is a slight abuse of notation and is used to represent the observed values from the last *business* day. So a forecast for Monday will use observed values from the previous Friday. Using observed values from Sunday would produce a much weaker benchmark due to different demand dynamics on working and non-working days.

Individual linear regression model A simple benchmark model is created by fitting a linear regression model to each building and period of the day. The demand of a building at time t is given by

$$\log y_t = \beta_{0,p} + \beta_{1,p} w_{0,t} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_p^2),$$

where p is the 15-minute period of the day at time t, $w_{0,t}$ is the scaled temperature experienced[2] at time t and ϵ_t is the residual. We call these our "Individual Linear Regression" (ILR) models.

Individual natural spline model A linear relationship between temperature and electricity demand may not be sufficient to adequately capture the relationship between the two. Natural cubic splines allow a more flexible relationship between predictors and the response. In this model the log demand of each building is modelled separately using natural splines. A building's demand based on temperature and other selected predictors is given by

$$\log y_t = f_p(w_{0,t}) + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_p^2),$$
$$f_p(x) = \sum_{k=1}^{K} \beta_{p,k} (x - \kappa_k)_+^3,$$

where f_p is a smooth function modelling the relationship between $w_{0,t}$ and the logged demand for period p. We refer to this set of models as the "Individual Natural Spline" (INS) models.

We use natural splines with three degrees of freedom as our smooth functions. Knot positions κ_k are calculated based on quantiles of the data. Natural splines are chosen over other types as they enforce the constraint of linearity beyond the boundary points, which seems a fair assumption when considering the behaviour of electricity demand consumption in relation to

[2]The 0 subscript denotes no lag and is consistent with Table 2 with the building subscript dropped.

Table 2: *Predictor variables evaluated during feature selection when determining $\mathcal{P}_t$.*

Variable $x_{b,i,t}$	Lag (15-minute periods)	Description
$w_{b,0,t}$	0	Scaled current temperature.
$w_{b,12,t}$	12	Scaled temperature lagged by 3 hours.
$w_{b,24,t}$	24	Scaled temperature lagged by 6 hours.
$w_{b,48,t}$	48	Scaled temperature lagged by 12 hours.
$w_{b,96,t}$	96	Scaled temperature lagged by 24 hours.
$w_{b,192,t}$	192	Scaled temperature lagged by 2 days.
$w_{b,288,t}$	288	Scaled temperature lagged by 3 days.
$w^{+}_{b,t}$		Maximum scaled temperature over last 24 hours.
$w^{-}_{b,t}$		Minimum scaled temperature over last 24 hours.
$\bar{w}_{b,t}$		Average scaled temperature over last 3 days.
$y_{b,96,t}$	96	Scaled actual demand lagged by 1 day.
$y_{b,192,t}$	192	Scaled actual demand lagged by 2 days.
$y_{b,672,t}$	672	Scaled actual demand lagged by 1 week.

extreme temperatures (see Figure 3 for an illustration). We wish to create a parsimonious model and assuming anything beyond a linear relationship in the extremes seems contrary to that aim. Failing to enforce the linearity constraint may result in unusual relationships being predicted if extrapolating beyond the training data.

3.1.2 Pooled regression model

Our pooled regression model is fit using data from all buildings. One model is fit for each 15-minute period of the day which is then used to predict demand of each building b at time t. Note that since all buildings are included in the model, we introduce the b subscript for buildings.

Additional predictor variables are introduced in this model, such as lagged temperature variables; maximum, minimum and average temperatures; and lagged demand. A description of each predictor is presented in Table 2. We denote this set of predictor variables as $\mathcal{P}_t$, which contains the selected variables for the 15-minute period of day and month at time t. The exact combination of variables is chosen via our feature selection methodology described in Section 3.2. This model is used when selecting features as it is much faster to train than a mixed effects model.

The demand of building b at time t is given by

$$\log y_{b,t} = \sum_{x_{b,i,t} \in \mathcal{P}_t} f_{i,p}(x_{b,i,t}) + \alpha_{b,p} + \epsilon_{b,t}, \quad \epsilon_{b,t} \sim N(0, \sigma_p^2),$$

$$f_{i,p}(x) = \sum_{k=1}^{K} \beta_{i,p,k}(x - \kappa_k)_+^3,$$

where $x_{b,i,t}$ is the value of building b's i^{th} predictor variable at time t and $f_{i,p}$ is a smooth function modelling the relationship between $x_{b,i,t}$ and the logged demand for period p. A dummy variable $\alpha_{b,p}$ has been added to account for differences in each building's consumption. We call this our "Pooled Regression" (PR) model.

We do not estimate a separate smooth relationship between weather variables and demand for each building in the pooled model. Instead, we estimate the population's relationship. So, for the i^{th} predictor we construct a smooth function $f_{i,p}$ for all buildings instead of a set of smooth functions $f_{b,i,p}$ for each building.

3.1.3 Mixed models

Having specified our framework for fitting separate models to each building it is now time to explore fitting mixed models. In each mixed effects model that follows, all buildings are included by treating each as a random effect. In section 4.2 we show that mixed models improve prediction accuracy and have the added benefit of allowing us to quantify the impact of building attributes on electricity demand.

Originally, random effects were incorporated into each of the lagged weather variables, but this resulted in very poor fits presumably due to the high dimensionality. Instead, as with the PR model, we model the population relationship for all selected predictor variables and allow for subject-specific differences using random intercepts, random slopes and subject-specific curves (based on current temperature).

As with the PR model, we include subscripts b to denote each building. Unlike the individual formulations, which had separate models fit to each building, all buildings are used when training the mixed effects models and so we include an additional subscript to denote this. Again, to capture changing demand characteristics across the day, separate models are fit for each 15-minute period of the day giving 96 models for each mixed effects formulation.

Random intercept model The simplest mixed effects model is a random intercept (RI) model. We model the log of the demand by

$$\log y_{b,t} = \sum_{x_{b,i,t} \in \mathcal{P}_t} f_{i,p}(x_{b,i,t}) + u_{b,p} + \epsilon_{b,t},$$
$$f_{i,p}(x) = \sum_{k=1}^{K} \beta_{i,p,k}(x - \kappa_k)_+^3,$$
$$\epsilon_{b,t} \sim N(0, \sigma^2_{\epsilon,p}), \quad u_{b,p} \sim N(0, \sigma^2_{u,p}),$$

where $u_{b,p}$ is a random effect that controls the intercept of the model. This is similar in form to the pooled regression model, with the dummy variable $\alpha_{b,p}$ replaced by the random intercept $u_{b,p}$. We don't use this model for feature selection in Section 3.2 as it takes much longer to fit than the pooled regression model.

Random intercept and slope model Expanding on this is the random intercept and slope (RIS) model which has a random effect for both the intercept and slope of the model. We model demand by

$$\log y_{bt} = \sum_{x_{b,i,t} \in \mathcal{P}_t} f_{i,p}(x_{b,i,t}) + u_{b,p,1} + u_{b,p,2} w_{b,0,t} + \epsilon_{b,t},$$
$$f_{i,p}(x) = \sum_{k=1}^{K} \beta_{i,p,k}(x - \kappa_k)_+^3,$$
$$\epsilon_{b,t} \sim N(0, \sigma^2_{\epsilon,p}), \quad (u_{b,p,1}, u_{b,p,2})^T \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma^2_{u,1} & \sigma_{u,1,2} \\ \sigma_{u,1,2} & \sigma^2_{u,2} \end{bmatrix}.$$

Here we have included a random slope based on scaled current temperature, $w_{b,0,t}$. The random effects $u_{b,p,1}$ and $u_{b,p,2}$ control the subject-specific differences for intercept and slope, respectively. The matrix Σ is a variance-covariance matrix for the random effects. It includes terms for the variance of intercepts ($\sigma^2_{u,1}$), the variance of slopes ($\sigma^2_{u,2}$) and the covariance between intercepts and slopes ($\sigma_{u,1,2}$).

Subject-specific curves model As the relationship between demand and temperature can be quite non-linear we also explore modelling the subject-specific differences in the temperature and energy relationship using splines. We call this model the subject-specific curve (SSC) model in keeping with Durbán et al. (2005). Note that we have modified their model to work with natural cubic splines as this gives a better fit when modelling the temperature and electricity

relationship compared to penalized linear splines (Figure 3). It is given by

$$\log y_{bt} = \sum_{x_{b,i,t} \in \mathcal{P}_t} f_{i,p}(x_{b,i,t}) + g_{b,p}(w_{b,0,t}) + \epsilon_{b,t},$$
$$f_{i,p}(x) = \sum_{k=1}^{K} \beta_{i,p,k}(x - \kappa_k)_+^3, \quad g_{b,p}(x) = \sum_{k=1}^{K} u_{b,p,k}(x - \kappa_k)_+^3,$$
$$\epsilon_{b,t} \sim N(0, \sigma_{\epsilon,p}^2), \quad u_{b,p,k} \sim N(0, \sigma_{u,p}^2).$$

As with our other models we use natural cubic splines with three degrees of freedom for both $f_{i,p}$ and $g_{b,p}$. This model allows us to capture separate temperature and electricity relationships for each building while also including the population relationships between electricity demand and other selected predictors.

Subject-specific curves with attributes model As we wish to carry out scenario analysis we introduce several new variables into our model. These variables are the set of building attributes discussed in Section 2.2 which we denote by $\mathcal{A}$. We treat each of these attributes as a fixed effect. We refer to this model as the subject-specific curves with attributes (SSCATTR) model.

Our model for scenario analysis is given below

$$\log y_{bt} = \sum_{x_{b,i,t} \in \mathcal{P}_t} f_{i,p}(x_{b,i,t}) + g_{b,p}(w_{b,0,t}) + \sum_{a \in \mathcal{A}} \beta_a x_{b,a} + \epsilon_{b,t},$$
$$f_{i,p}(x) = \sum_{k=1}^{K} \beta_{i,p,k}(x - \kappa_k)_+^3, \quad g_{b,p}(x) = \sum_{k=1}^{K} u_{b,p,k}(x - \kappa_k)_+^3,$$
$$\epsilon_{b,t} \sim N(0, \sigma_{\epsilon,p}^2), \quad u_{b,p,k} \sim N(0, \sigma_{u,p}^2).$$

This is the same as our SSC model apart from the addition of the building attributes. The fixed effect $x_{b,a}$ is a Boolean variable that indicates if attribute a is present for building b.

3.2 Feature selection

Carrying out feature selection for such a wide range of models was a difficult problem to approach. We take the view that it is best to keep features consistent between each of the models in order to fairly compare each during the validation stage. Hence, each model's performance is conditional on the same set of predictor variables. As we don't expect the most important predictors to be changing rapidly throughout the year we only conduct feature selection for the first business day of each month. The selected predictors are then used for all business day forecasts in the month.

Table 2 shows a list of demand variables that were considered for our modelling. Lagged temperature variables are used to model the impact of thermal inertia in buildings. For example, high overnight temperatures in summer may result in high demand on the following day due to the increased cooling loads required to maintain suitable indoor environment quality. The maximum and minimum temperatures from the last 24 hours are also considered, as well as the mean temperature over the previous three days. Lagged demand values of 1, 2 and 7 days are included to capture any serial correlation in the observed demand time series.

Numerous studies have already shown the link between electricity demand and current temperature (Ben Taieb et al. 2016; Fan & Hyndman 2012; Roach 2019; Hong et al. 2016; Hong, Xie & Black 2019). Hence, we chose to conduct feature selection conditional on the current temperature being included. There were several reasons for this:

- Much of the literature on load forecasting already identifies the importance of current temperature in forecasting demand and we can see clear nonlinear relationships in Figure 2.
- As temperature is strongly correlated with recent values there were occasions when the current temperature would not be selected but a slightly lagged variable would be. This seemed unrealistic and was likely caused by noise in the data rather than a lagged temperature being a better predictor than actual temperature.
- Forcing current temperature to be included reduced the number of feature combinations to search through by a factor of 2.

We use the pooled linear regression model for feature selection as it is quick to fit using OLS and allows us to model buildings by using a dummy variable for each. Using a linear model also has the advantage of allowing us to efficiently compute the leave-one-out cross-validation (LOOCV) scores using (Seber & Lee 2012)

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-h_i}\right)^2,$$

where e_i are the residuals of the model and h_i are the diagonal elements of the hat-matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.

Training data are comprised of business days within a window of 120 days prior to the month we wish to select variables for. As our experimental setup for the validation phase involves one-day ahead forecasts, this variable selection prevents us from using any data from the future.

For example, when forecasting for any date in January, only data from the months preceding January would have been used to select predictors.

We use best subset selection during feature selection. Given p predictors we choose the combination of these that produce the best R^2 scores. Once the best model based on R^2 has been determined for each set of p predictors, we use the LOOCV score to determine the overall best. The LOOCV score is chosen as it gives an estimate of the out of sample performance of our models. Figure 4 shows the LOOCV scores for each predictor set of size p. During feature selection we chose to avoid greedy approaches such as forward or backward stepwise selection; or approaches that work systematically through lagged weather variables (Hyndman & Fan 2010). Naturally, greedy methods have computational benefits, but it is interesting to observe which features are chosen when *all* possible predictor combinations are assessed.

A key point to note is that feature selection was done on the weather variables and *not* the spline basis functions. Doing so would destroy the properties of a spline if only a subset of its basis functions were to be selected.

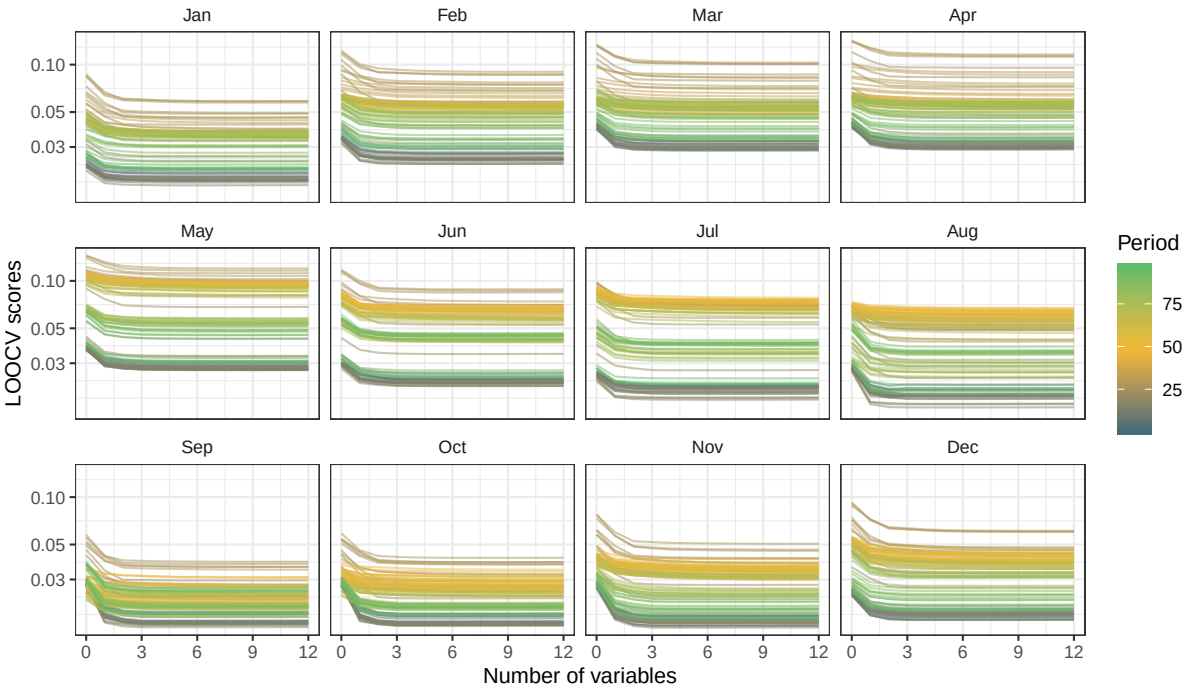

Figure 4: *LOOCV scores (log scale) for each month. In general, the LOOCV errors initially decrease as variables are added, but begin to increase slightly at a certain point for each period.*

It should be noted that feature selection could be further improved for the mixed effects models by proceeding with a step-wise selection process after the above process has completed for the pooled model. Features can be added or removed based on if an appropriate out of sample

accuracy score improves. This allows us to benefit from the relative speed of fitting via OLS before further fine-tuning with a greedy selection algorithm.

3.3 Validation

3.3.1 Rolling origin 1-day ahead forecasts

We used a historical training period comprised of recent observations for each building. Business days from a sliding window of length 120 days were selected as training data for each model. Using recent observations allows recent operational changes or trends to be captured in each model.

If, for a given 1-day ahead forecast, a building had less than 40 days of training data present then it was removed from the forecast. This was done to accommodate buildings that had recently been included in the data set or where the data had been censored. Training a building with less than 40 days of data sometimes resulted in severe overfitting.

3.3.2 Error measures

To assess the forecasting accuracy of each of our models we use four common error metrics.

1. Mean absolute error: MAE $=$ mean $(|y_t - \hat{y}_t|)$.
2. Mean absolute percentage error: MAPE $=$ mean $\left(\left|\frac{100(y_t-\hat{y}_t)}{y_t}\right|\right)$.
3. Symmetric mean absolute percentage error: sMAPE $=$ mean $\left(\frac{200|y_t-\hat{y}_t|}{y_t+\hat{y}_t}\right)$.
4. Mean absolute scaled error: MASE $=$ mean $\left(\left|\frac{y_t-\hat{y}_t}{\text{mean}(|y_t-y_{t-1}|)}\right|\right)$.

These are all well established forecasting metrics. Advantages and disadvantages of each are described in Hyndman & Koehler (2006).

When comparing these metrics in Section 4.2, we find that the SSC and SSCATTR models produce the best point forecasts. To establish that this result is statistically significant we also carry out Diebold-Mariano tests against the ILR model in Section 4.3.

4 Results

In order to build a better understanding of how a mixed model framework improves upon fitting individual models to each building we need to assess each model's performance. To do so, we create one-day ahead ex-post forecasts and calculate the MAE, MAPE, sMAPE and

MASE for each. We focus on ex-post forecasting as we wish to examine error caused by model specification and ignore errors caused by incorrect weather forecasts, as would be the case in an ex-ante forecasting scenario.

4.1 Variables chosen via feature selection

Figure 5 show the number of times each variable is selected for all months of the year. There is a lot of variation in the selected predictors, even between adjacent 15-minute periods. By overlaying all of the months we do observe some structure. Perhaps the most noticeable characteristic is that temperature variables are selected more often during business hours, which shows the influence temperature has on demand during the day and how temperature influences occupant behaviour. Outside of these hours we see fewer temperature variables selected. Another point of interest is that during business hours, lagged demand variables are selected less often than for non-business hours. It would appear as though serial correlation in the demand time series is a more useful predictor during non-business hours than temperature. Thermal inertia does not appear to influence demand as much during non-business hours.

4.2 Forecasting accuracy

Table 3 shows the MAE, MAPE, sMAPE and MASE for each model across the entire day, during business hours (7:00 am to 7:00 pm) and during non-business hours (7:00 pm to 7:00 am).

The Naive model has the worst forecasting accuracy. All of the benchmark and mixed effects models outperform it. Overall, the best performing model is the SSC model, closely followed by the SSCATTR model. These two models consistently outperform others across all metrics. Figure 6 shows an example of forecasts produced from the SSC model. The actual demand values and previous day's demand values (Naive model) are also plotted. We can see that the SSC model tends to track the general shape of each profile well and does not predict erratic spikes in demand.

Given the SSC and SSCATTR models outperform each of our benchmarks it seems reasonable to conclude that forecasting with mixed effects models is a reasonable practice that should be encouraged when data are available for similar subjects.

4.3 Diebold-Mariano test

Here we perform a one-sided Diebold-Mariano test (Diebold & Mariano 2002) to determine if our final model (SSCATTR) is more accurate than the baseline model (ILR). We perform a test

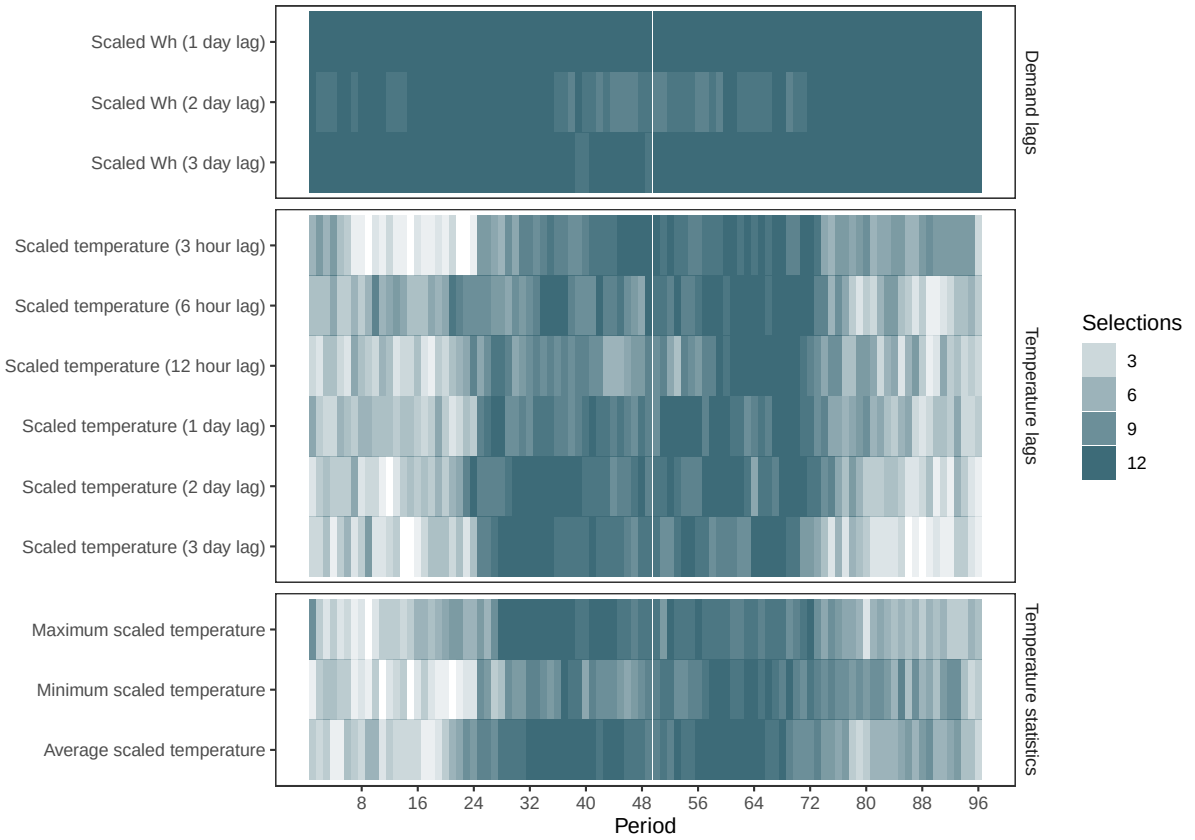

Figure 5: *Feature selection for all months. The number of times a variable has been selected across all 12 months is indicated by the transparency of each tile for a given period. In general, weather features are selected more often during business hours. During non-business hours, weather features are selected less often, but lagged demand variables are almost always selected.*

for each period of the day (Table 4). When the test is applied to each period of the day we see that our SSCATTR model produces forecasts that are significantly better than the ILR model.

4.4 Scenario analysis

Having confirmed that our mixed models produce satisfactory predictions compared to individual models provides us with justification for using them for scenario analysis. We can assess how changing certain variable values increases or decreases energy demand in buildings.

We will now show an example of producing scenario analysis for one of our buildings using this model. Building BID0010 has the attributes listed in Table 5. If any of the attribute values are modified and then the expected energy consumption is recalculated, we can take the difference between the original predictions to estimate how energy demand will change. Figure 7 shows how expected demand changes when we vary building attributes for several days in January. Figure 8 shows the cumulative change in energy consumption if the building were to switch away from using a DX system. These normalised energy savings (or increases) can then be

Table 3: *Forecasting accuracy measures for each model across the entire day, business hours (7:00 am to 7:00 pm) and non-business hours (7:00 pm to 7:00 am).*

Model	MAE	MAPE	sMAPE	MASE
All hours				
ILR	0.420	14.8	13.2	0.946
INS	0.401	14.5	12.8	0.902
Naive	0.444	15.3	13.4	1.000
PR	0.415	13.9	12.5	0.936
RI	0.413	13.9	12.4	0.930
RIS	0.384	13.3	11.8	0.865
SSC	**0.374**	**13.1**	**11.6**	**0.843**
SSCATTR	0.375	**13.1**	**11.6**	0.844
Business hours				
ILR	0.637	16.8	14.0	0.920
INS	0.601	16.2	13.4	0.869
Naive	0.692	17.9	15.3	1.000
PR	0.652	16.8	14.3	0.942
RI	0.648	16.7	14.2	0.937
RIS	0.595	15.8	13.2	0.860
SSC	**0.576**	**15.4**	**12.9**	**0.833**
SSCATTR	0.578	**15.4**	**12.9**	0.835
Non-business hours				
ILR	0.204	12.8	12.3	1.040
INS	0.201	12.7	12.2	1.020
Naive	0.196	12.7	11.5	1.000
PR	0.179	11.1	10.7	0.914
RI	0.178	11.0	10.6	0.907
RIS	0.174	10.8	**10.4**	0.885
SSC	**0.172**	**10.7**	**10.4**	0.877
SSCATTR	**0.172**	**10.7**	**10.4**	**0.876**

converted to a dollar figure by simply multiplying by the net lettable area of the building and the appropriate electricity tariffs. Such an approach allows us to quantify the expected savings, both financially and in terms of energy consumption, allowing for better decision making when it comes to building management and retrofits.

5 Conclusion

This paper explores the possibility of using mixed effects models in a forecasting role. We first specified several different models. A best subset selection approach was proposed to determine which predictor variables should be used. Feature selection was carried out for each month of the year and 15-minute period of the day, which allowed us to observe how the importance of lagged temperature and demand variables changed throughout the day.

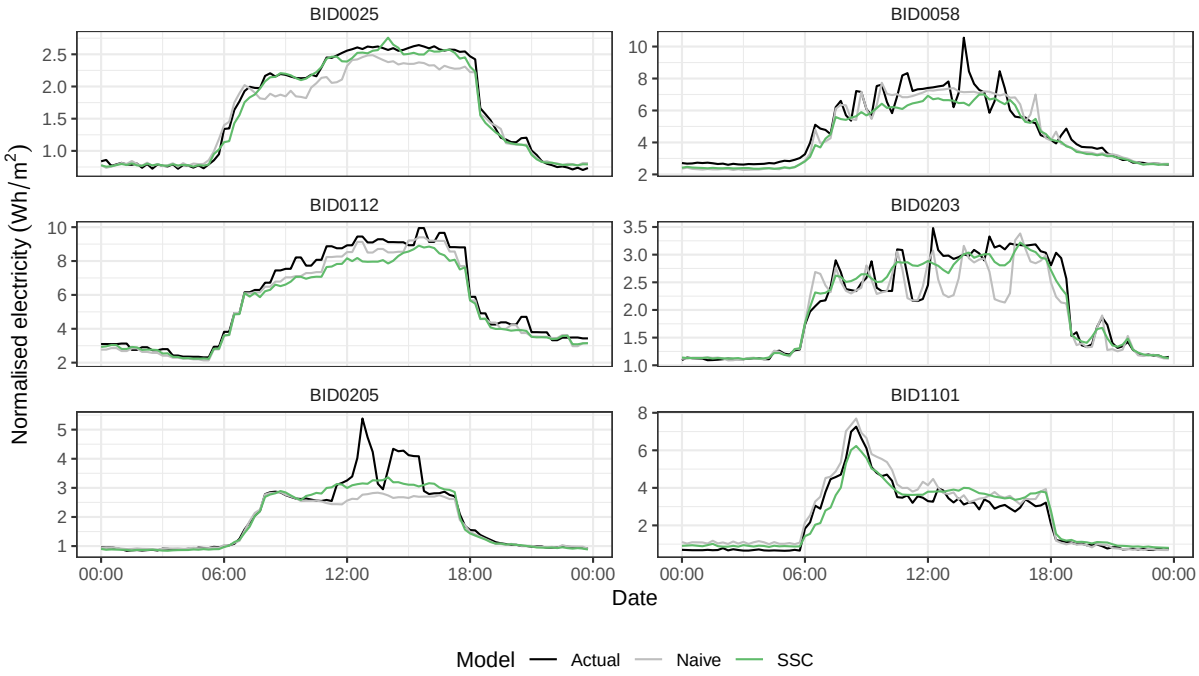

Figure 6: *One-day ahead forecasts for 23 August, 2017. The Naive (yesterday's actuals) and SSC models are shown. The naive model often includes erratic spikes whereas the SSC model tends to produce a smoother profile.*

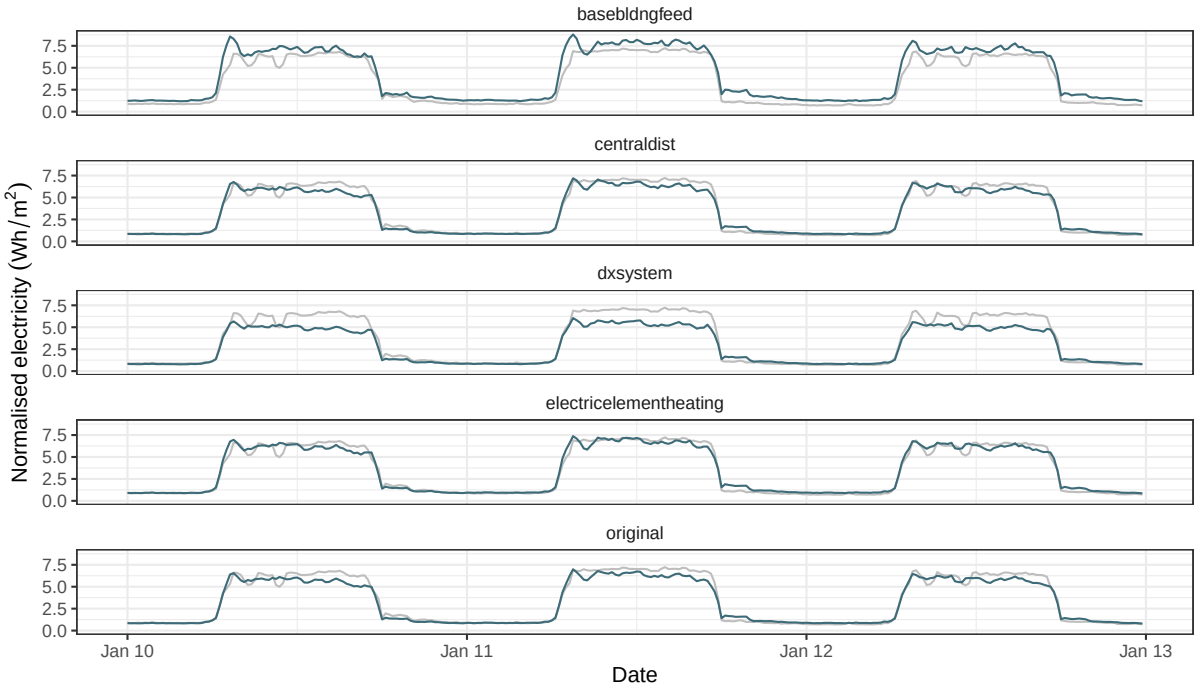

Figure 7: *Expected normalised electricity consumption for three days in January, 2017 for building BID0010. The grey line shows the actual demand that was observed on each day. The expected demand of the existing building is shown ("Original building"), as well as four scenarios each of which involves the modification of one attribute.*

Table 4: *Diebold-Mariano test to compare forecast accuracy of ILR and SSCATTR models. Alternative hypothesis is SSCATTR model is more accurate than ILR model. All periods tested separately. Only every fourth model falling on the hour is shown for legibility.*

Period	DM statistic	p-value	Significance
4	1.062	0.144	
8	-0.371	0.645	
12	-0.109	0.543	
16	-0.733	0.768	
20	-0.548	0.708	
24	2.872	0.002	**
28	3.123	< 0.001	***
32	1.513	0.065	.
36	1.040	0.149	
40	4.281	< 0.001	***
44	3.580	< 0.001	***
48	2.982	0.001	**
52	3.422	< 0.001	***
56	3.163	< 0.001	***
60	4.203	< 0.001	***
64	4.742	< 0.001	***
68	7.260	< 0.001	***
72	7.913	< 0.001	***
76	14.236	< 0.001	***
80	9.641	< 0.001	***
84	10.593	< 0.001	***
88	9.196	< 0.001	***
92	7.242	< 0.001	***
96	4.443	< 0.001	***

Table 5: *Building attributes for Building BID0010.*

Attribute	Present
basebldngfeedonly	TRUE
dxsystem	TRUE
electricelementheating	FALSE
centraldist	TRUE

We fit models to 123 buildings across Australia. Separate models for each building were fitted as a benchmark. The overall predictive power of several mixed effects models were assessed against this benchmark. One-day ahead forecasts were produced for business days over a year using all forecast methods. Based on the MAE, MAPE, sMAPE and MASE scores of each model the SSC and SSCATTR models performed best. We concluded that predicting demand using mixed effects model could improve forecast accuracy and had the additional advantage of allowing us to conduct scenario analyses.

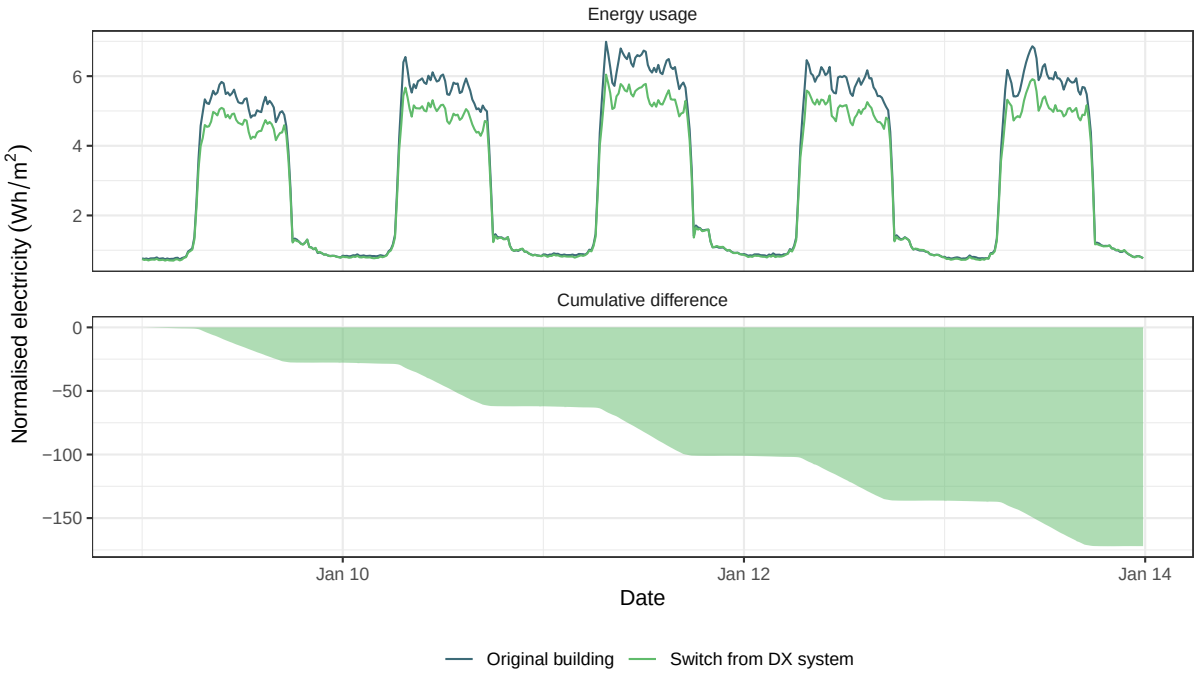

Figure 8: *Cumulative energy impact after removing DX system. We can see that the cumulative change in energy consumption steadily decreases indicating buildings without DX systems are more efficient.*

Finally, we included an example of using the SSCATTR model for scenario analysis for one of the buildings in our data set. The expected change in electricity demand was plotted when several of the attributes were varied. The expected change in consumption over the course of one workweek when a building moved away from using a DX system allowed us to quantify the potential savings in energy. Analyses such as these have applications for decision makers and facility managers that wish to understand the effectiveness of changes in building management or potential retrofits of equipment.

Acknowledgements

This research project was supported by funding from Buildings Alive. I would like to thank Buildings Alive for making data available and their guidance in understanding commercial building equipment and behaviour. I would also like to thank the reviewers for their thorough and constructive feedback.

This research was supported by use of the Nectar Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS).

Data Availability Statement

The data that support the findings of this study are available from Buildings Alive. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors with the permission of Buildings Alive.

References

Aldor-Noiman, S, PD Feigin & A Mandelbaum (Dec. 2009). Workload forecasting for a call center: Methodology and a case study. *Annals of Applied Statistics* **3**(4), 1403–1447.

Arora, S & JW Taylor (Mar. 2016). Forecasting electricity smart meter data using conditional kernel density estimation. *Omega* **59**, 47–59.

Bates, D, M Maechler, B Bolker & S Walker (2019). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. R package version 1.1-21. https://CRAN.R-project.org/package=lme4.

Ben Taieb, S, R Huser, RJ Hyndman & MG Genton (2016). Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression. *IEEE Transactions on Smart Grid* **7**(5), 2448–2455.

Ben Taieb, S, JW Taylor & RJ Hyndman (2020). Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data. *Journal of the American Statistical Association*. to appear, 1–36.

Ben Taieb, S, J Yu, M Neves Barreto & R Rajagopal (2017). Regularization in Hierarchical Time Series Forecasting With Application to Electricity Smart Meter Data. In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Brabec, M, O Konár, E Pelikán & M Malý (Oct. 2008). A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers. *International Journal of Forecasting* **24**(4), 659–678.

Diebold, FX & RS Mariano (Jan. 2002). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* **20**(1), 134–144.

Durbán, M, J Harezlak, MP Wand & RJ Carroll (Apr. 2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine* **24**(8), 1153–1167.

Fan, S & RJ Hyndman (Feb. 2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems* **27**(1), 134–141.

Frees, EW & TW Miller (Jan. 2004). Sales forecasting using longitudinal data models. *International Journal of Forecasting* **20**(1), 99–114.

Gajowniczek, K & T Ząbkowski (2014). Short Term Electricity Forecasting Using Individual Smart Meter Data. *Procedia Computer Science* **35**, 589–597.

Ghofrani, M, M Hassanzadeh, M Etezadi-Amoli & MS Fadali (Aug. 2011). Smart meter based short-term load forecasting for residential customers. In: *NAPS 2011 - 43rd North American Power Symposium*. IEEE, pp.1–5.

Grajeda, LM, A Ivanescu, M Saito, C Crainiceanu, D Jaganath, RH Gilman, JE Crabtree, D Kelleher, L Cabrera, V Cama & W Checkley (Jan. 2016). Modelling subject-specific childhood growth using linear mixed-effect models with cubic regression splines. *Emerging Themes in Epidemiology* **13**, 1.

Hong, T, P Pinson, S Fan, H Zareipour, A Troccoli & RJ Hyndman (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting* **32**(3), 896–913.

Hong, T, J Xie & J Black (Oct. 2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting* **35**(4), 1389–1399.

Hyndman, RJ & G Athanasopoulos (May 2018). *Forecasting: principles and practice*. OTexts. https://OTexts.com/fpp2.

Hyndman, RJ & S Fan (May 2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems* **25**(2), 1142–1153.

Hyndman, RJ & AB Koehler (Oct. 2006). Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**(4), 679–688.

Ibrahim, R & P L'Ecuyer (Feb. 2013). Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models. *Manufacturing & Service Operations Management* **15**(1), 72–85.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Roach, C (Oct. 2019). Reconciled boosted models for GEFCom2017 hierarchical probabilistic load forecasting. *International Journal of Forecasting* **35**(4), 1439–1450.

Roach, C (Mar. 2020). Estimating electricity impact profiles for building characteristics using smart meter data and mixed models. *Energy and Buildings* **211**, 109686.

Seber, GAF & AJ Lee (Jan. 2012). *Linear Regression Analysis*. John Wiley & Sons.

Soyer, R & MM Tarimcilar (Feb. 2008). Modeling and Analysis of Call Center Arrival Data: A Bayesian Approach. *Management Science* **54**(2), 266–278.

Ugarte, T Goicoa, AF Militino & M Durbán (Aug. 2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics & Data Analysis* **53**(10), 3616–3629.

Chapter 5

Exploring unusual sensor behaviour in buildings using BMS data and unsupervised learning techniques

The following paper was submitted to the journal *Energy and Buildings*.

All code to reproduce the paper will be made available at https://github.com/camroach87/1802-ufd after publication.

Exploring unusual sensor behaviour in buildings using BMS data and unsupervised learning techniques

Cameron Roach
Monash University
Email: cameron.roach@monash.edu
Corresponding author

17 March 2020

JEL classification: C10,C38

Exploring unusual sensor behaviour in buildings using BMS data and unsupervised learning techniques

Abstract

Building Management Systems (BMSs) are used to control HVAC equipment, lighting and other devices in commercial buildings. These systems can generate significant volumes of data with a single building typically containing tens of thousands of sensors. Due to the large volumes of data it is difficult for facility managers to quickly assess if a building is performing as expected or if faulty sensors are present. Furthermore, the data that is collected can often be difficult to deal with. Irregular time intervals, missing values, outliers and inconsistent sensor labelling all add complexity and introduce new problems that need to be addressed if accurate fault detection is to be carried out. This paper explores using unsupervised machine learning methodologies to allow end users to quickly assess if BMS points are behaving as expected. To deal with the inherent complexity of the time series data and metadata we engineer simple but useful features to improve our analysis. We test several dimensionality reduction techniques that allow us to visualise data easily. Two examples are provided to show how our approach can lend itself to quickly detecting faults or unusual behaviour. Applications of this methodology include fault detection and improving our understanding of BMS data. We also discuss further applications of our dimensionality reduction approach such as the possibility of comparing behaviour across multiple buildings.

Keywords: time series, anomaly detection, BMS data, dimensionality reduction, t-SNE

1 Introduction

Fault detection in buildings management systems (BMSs) can prevent energy wastage and help improve occupant comfort. Many methods exist for fault detection and diagnostics and can be broadly grouped into three main categories. Quantitative methods include physical models that attempt to emulate the behaviour of a correctly calibrated and functioning system. Qualitative methods rely on set rules which when violated indicate a fault. The third and more distinct

category is that which relies on historical data to diagnose faults. Fault detection methods that fall into this category include those that rely on statistical or machine learning techniques. An advantage of using historical data to detect sensor faults is that modelling is typically less complicated than for quantitative and qualitative approaches (Katipamula & Brambley 2005).

In this paper we present a data focused approach to fault detection. We extract time series and metadata features from historical sensor data before applying several dimensionality reduction approaches to identify outliers. Engineering time series features rather than using raw data is advantageous as working with raw data requires regular sampling intervals so that each period of the day can be treated as a feature. When sensors are recorded at different frequencies or times this is no longer possible and so other suitable features that describe the time series need be engineered. Metadata contains useful information that can help identify sensor types. Typical sensor names will contain abbreviations of the sensor type. Naming conventions are often inconsistent both within a building and across different buildings which makes it difficult for building engineers to process the metadata and extract useful information. To work around this issue we create n-grams for each sensor name which are then used alongside the time series features during our analysis.

We take an unsupervised learning approach to our outlier detection. Hence our algorithm does not classify sensors or days as behaving erratically, but presents the data in such a way that an end user may easily determine if a sensor is unusual compared to similar sensor types. Our focus is on presenting the data in such a way that an end user may explore the data interactively in a quick and easy manner. Figure 1 shows our intended work flow. Note that this approach may be useful for supervised learning as it allows a user to quickly tag multiple points that can then be used as training data for a suitable classification model.

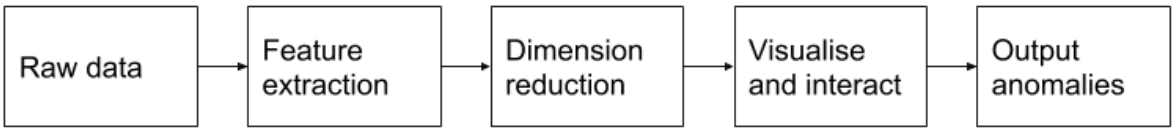

Figure 1: *Intended work flow for anomaly detection.*

We focus on producing a fault detection approach using unsupervised learning techniques as we do not require labelled data to begin diagnosing faults. Typically, in a supervised learning scenario a data set containing sensors and historical time series will be labelled as faulty or normal. However, producing a labelled data set can be both challenging and time consuming. Furthermore, each of these data sets may be domain specific. Hence, if a new building is encountered an entirely new labelled data set may be required to train a new fault detection

model. In contrast, unsupervised learning algorithms can immediately be applied to new and unlabelled datasets. While this paper focuses on exploring different unsupervised learning techniques to help improve our analysis of data our eventual goal is to move to a supervised learning approach whereby an algorithm can automatically classify anomalies.

Some papers have focused on unsupervised fault detection. Costa, Angelov & Guedes (2014) and Costa, Angelov & Guedes (2015) explore unsupervised fault detection using recursive density estimation and data clouds, which are similar to clusters but do not have a specific shape or boundary. They test their methodology on a pilot plant for industrial process control. Theirs is a two stage process that first conducts recursive density[1] estimation in the feature space before applying an evolving fuzzyrule-based classifier. A recent framework by Ardakani et al. (2016) uses a combination of clustering approaches and multivariate dynamic metamodels to identify faults in a simulation case study for a three tank system. They attempt to detect leaking and plugging within the system.

Our unsupervised approach allows for easy identification of unusual behaviour in sensors and does not require any *a priori* knowledge of a building's properties or physics underpinning its behaviour. It can easily be extended to examine different buildings or multiple buildings simultaneously. We can detect anomalies for AHUs, chillers, pumps and any other unit since we are simply using historical time series and metadata. We aim to be able to detect both *hard faults* (sensor/actuator issues) and explore *soft faults* (controls programming issues).

Many studies on fault detection utilise either simulated datasets with artificially introduced faults or small case studies of real data, both of which have drawbacks. Simulated datasets rely on assumptions by researchers on the characteristics of faults whereas real world studies tend to only have a small set of fault examples and assume that all faulty data is correctly labelled (Gunay, Shen & Yang 2017). In our paper we use a real world data set based off several buildings. We do not rely on manual tagging of faulty or normal operation and instead focus on producing a methodology that can allow us to identify system faults at a glance if particular sensors are behaving erratically. Gunay, Shen & Yang (2017) also point to the need for fault detection methods that can be adapted to different types of buildings with minimal tuning and configuration.

The paper is structured as follows. We first discuss the data and the different BMS points in Section 2. Suitable time series and metadata features are proposed in Section 3. Section 4 discusses several popular dimensionality reduction methods. These methods are then assessed

[1]Note that their definition of density is not the same as the standard statistical definition of a probability density function.

Table 1: *Number of points for each measurement type.*

Measure	BID0025	BID0126	BID1701
Cooling control valve (CCV)	20	33	6
Economy cycle dampers (ECD)	18	33	6
Enabled (ENB)	22	34	8
Return air temperature (RAT)	1	32	2
Supply air pressure (SAPR)	16	30	8
Supply air pressure setpoint (SAPRSP)	16	29	8
Supply air temperature (SAT)	20	32	8
Supply air temperature setpoint (SATSP)	16	32	6
Speed (SPD)	16	30	8
Status (STS)	20	33	8
VAV damper position max (VAVDM)	16	30	0

in Section 5 and an implementation of our preferred algorithm is presented alongside two examples. Limitations and future research directions are also considered. Concluding remarks are provided in Section 6.

2 Data

Three separate buildings are examined in this study. We focus on identifying sensor types for AHUs. Table 1 shows counts of each AHU measure type across the three buildings. Measure type is the type of measurement being recorded and for AHUs may include points such as room temperature, temperature set point, supply air pressure and so on. We only examine those sensors that have already been manually labelled which allows us to assess which dimension reduction algorithms appear to behave best. Naturally, this is only a small subset of the available sensors, but it is adequate for our testing purposes. We use two weeks of sensor reading data during January 2017. Raw data observations occur approximately every 15 minutes for most sensors, with some recorded approximately every hour. The observations do not occur on the 15-minute marks of each hour as each point in the BMS is polled sequentially, resulting in slight time offsets.

Figure 2 provides some motivation for finding a suitable way to visualise sensor readings. Time series plots can be easily interpreted and compared when working with only a few sensors. However, as more sensors are examined this approach becomes unsuitable. The plot with 100 sensors is unreadable and does not allow a user to quickly explore the data. Plotting each sensor separately is also problematic due to the large number of plots that would be created.

Instead, we require a means to adequately represent each sensor in two-dimensional space while capturing temporal and metadata features that characterise each.

3 Feature engineering

We engineer time series features based off the sensor data. We use a similar approach to that of Hyndman, Wang & Laptev (2015), but adapt the feature space to that of our domain. Manually creating time series features is somewhat subjective. Here we opt for a simple set of time series features primarily based on lower and higher order statistics, which is appropriate given the sensor data we focus on appears to be stationary conditional on time of day. Using a simple set of features allows for easy implementation in languages such as R or Python as no language specific packages are required to calculate more sophisticated time series features. Note that more advanced features may improve dimensionality reduction performance, but we show that our simple set of features appears to provide useful results in this domain. All time series features used are listed in Table 2. Each feature is normalised so that all values lie between 0 and 1. An advantage of using features based on statistical properties of the time series is that it does not matter if recordings do not fall exactly on the 15 minute marks of each hour. Hence, we are not required to carry out any interpolation which would potentially degrade the data quality and mask faults. Figure 3 shows an example of converting raw time series readings for three sensors to statistical features (without normalisation). We can observe that each time series has a dramatically different shape and is sampled at different times and intervals to the others. Despite this, the statistical features can still be calculated and capture differences between each time series.

In addition to time series features we also explore using text features derived from the metadata (Table 3). We explore using character level bigrams and trigrams. Punctuation and numeric characters are cleaned from the metadata to ensure we only focus on using alphanumeric characters. As an example, if a sensor contains the name `VAV-L2-PN2-SupAirDmpr:present-value` in its metadata, its first three character bigrams will be `va`, `av` and `vl`; and its first three trigrams will be `vav`, `avl` and `vlp`. Naturally, not all of these will be useful, but looking further along we can see that trigrams such as `sup`, `air` and `dmp` will also be produced which are more descriptive. This allows us to incorporate useful sections from the metadata that may contain clues about the sensor type.

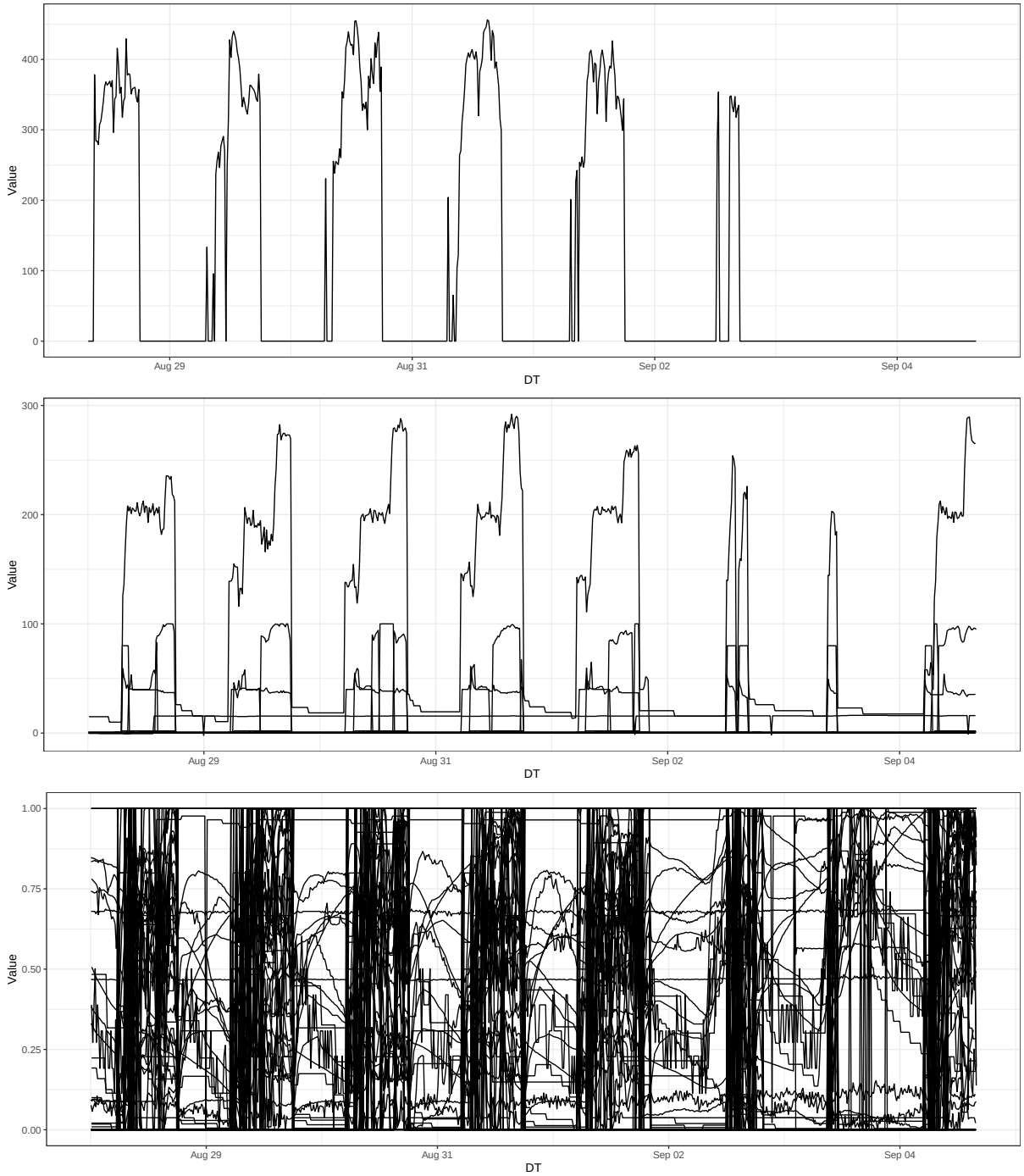

Figure 2: *Plots of time series for 1, 10 and 100 (scaled) BMS points. Attempting to identify anomalous behaviour using raw time series plots quickly becomes unwieldy even with such small sample sizes. We present a better approach to visualising these time series which can also make use of included metadata.*

Table 2: *Time series features for sensor data.*

Feature	Description
Number Unique	Number of unique values.
Mean	Mean value.
Max	Maximum value.
Min	Minimum value.
SD	Standard deviation.
Skewness	Asymmetry of distribution.
Kurtosis	Tail weight of distribution.
Max change	Maximum change in value between readings.
Min change	Minimum change in value between readings.
Mean crossings	Number of times the sensor readings cross the mean value.

Table 3: *Text features for sensor metadata.*

Feature	Description
Bigrams	Character level bigrams of sensor name excluding punctuation and numeric values.
Trigrams	Character level trigrams of sensor name excluding punctuation and numeric values.

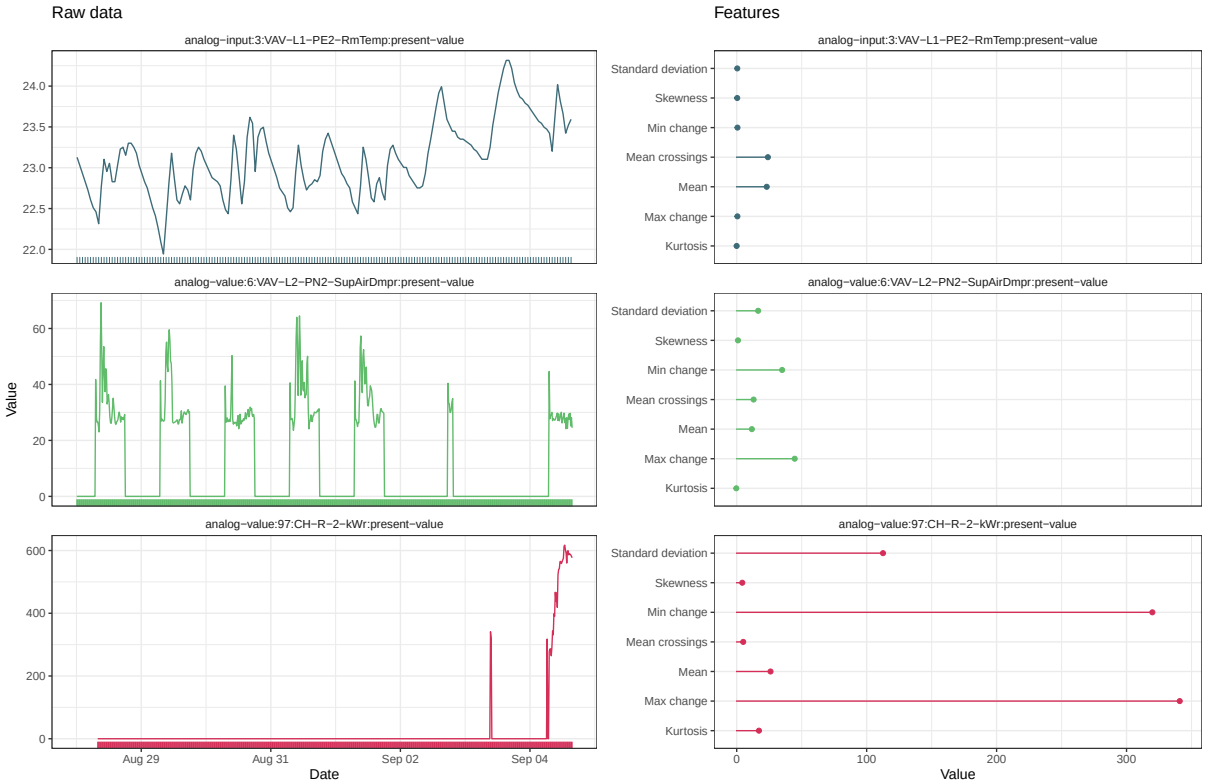

Figure 3: *Feature extraction for time series data. Note that time series are recorded irregularly as indicated by the rug plots. The extracted features are plotted on the right for each time series allowing for comparison on a common feature space.*

4 Unsupervised learning

In this section we provide brief introductions to each of the unsupervised learning techniques we explore in this paper. Each of these are tested and commented on. We do not focus on the mathematics that underpin each of these algorithms but instead give justifications for why they may be appropriate. Suitable references are provided for those wishing to explore the more technical details.

4.1 Dimensionality reduction

Dimensionality reduction allows us to project a high dimensional feature space into a low dimensional space for easy visualisation. Many approaches are available. Here we focus on principal component analysis (PCA), sparse PCA, isometric mapping (Isomap), t-distributed stochastic neighbor embedding (t-SME) and spectral embedding.

Principal component analysis. PCA is an unsupervised learning technique that has been used in various fault detection approaches (Kim & Katipamula 2018). Despite its popularity it does have some drawbacks that need to be considered. PCA focuses on producing orthogonal components that capture as much variation in the data as possible. It does not focus on preserving proximity relationships between points and neighbourhoods. Good introductions to PCA are available in Friedman, Hastie & Tibshirani (2001) and Murphy (2012).

Sparse principal component analysis. Sparse PCA (Zou, Hastie & Tibshirani 2006) is similar to traditional PCA but does not use all features to construct the principal components. Instead, it uses the lasso penalty to ensure some features receive a weight of zero and do not contribute to the principal components making it useful when working with high-dimensional feature spaces. This allows for a more easily interpretable representation with a clear distinction between features that do and do not contribute to differences in the data samples. Sparse PCA was chosen to help work with the large number of features produced when computing n-grams.

Isometric mapping. Isomap constructs a neighbourhood graph over all data points. This neighbourhood graph is then used to calculate geodesic distances between all points. Multidimensional scaling is then applied to the matrix of graph distances to create a low dimensional space. Further information can be found in Tenenbaum, Silva & Langford (2000).

t-SNE. T-SNE maps each data point to a location in a two or three dimensional space. It is well suited to visualising high dimensional spaces in two or three dimensions as it plots similar objects nearby and dissimilar objects far away with high probability. For certain tasks it has

been shown to perform better than other dimension reduction approaches such as Isomap and Locally Linear Embedding (Maaten & Hinton 2008).

Spectral embedding. Spectral embedding is designed to preserve proximity relations (Saul et al. 2006; Brand & Huang 2003). In other words, it attempts ensure that sensors that are different to others in a high dimensional feature space also appear far apart when viewed in a low dimensional space. This is different to PCA which only attempts to find components along which variance in the data is maximised. Spectral embedding constructs a weighted graph using an affinity matrix. This weight matrix is then used to construct a graph Laplacian on which eigenvalue decomposition is carried out. The eigenvectors form our low dimensional space. We test spectral embedding using both nearest neighbours and radial basis functions for our affinity matrix.

5 Discussion

5.1 Comparing dimensionality reduction methods

We compare the performance of the dimensionality reduction techniques discussed in Section 4.1. The projection of the time series and n-gram features onto two dimensions is shown in Figure 4.

Of all the approaches, t-SNE appears to show the best separation between measure types. Some measure types are grouped together (e.g. supply air pressure and supply air pressure set point). However, this is common to each of the dimensionality reduction methods and to be expected given the similarity in sensor names and time series readings.

PCA performs reasonably well and it is easy enough to see the different measure types. However, the separation between each class is not as clear as for t-SNE, especially in the center of the plot where sensors from several measure types overlap. The remaining algorithms are reasonable, but again none appear to outperform t-SNE.

5.2 Features for multiple buildings

It is important to comment on the impact that different naming conventions between buildings can have on the dimensionality reduction step. Figure 5 shows how multiple buildings can affect dimensionality reduction. One of the buildings uses a different naming convention for sensors and so points that are taken from this building are pushed far away from the other buildings.

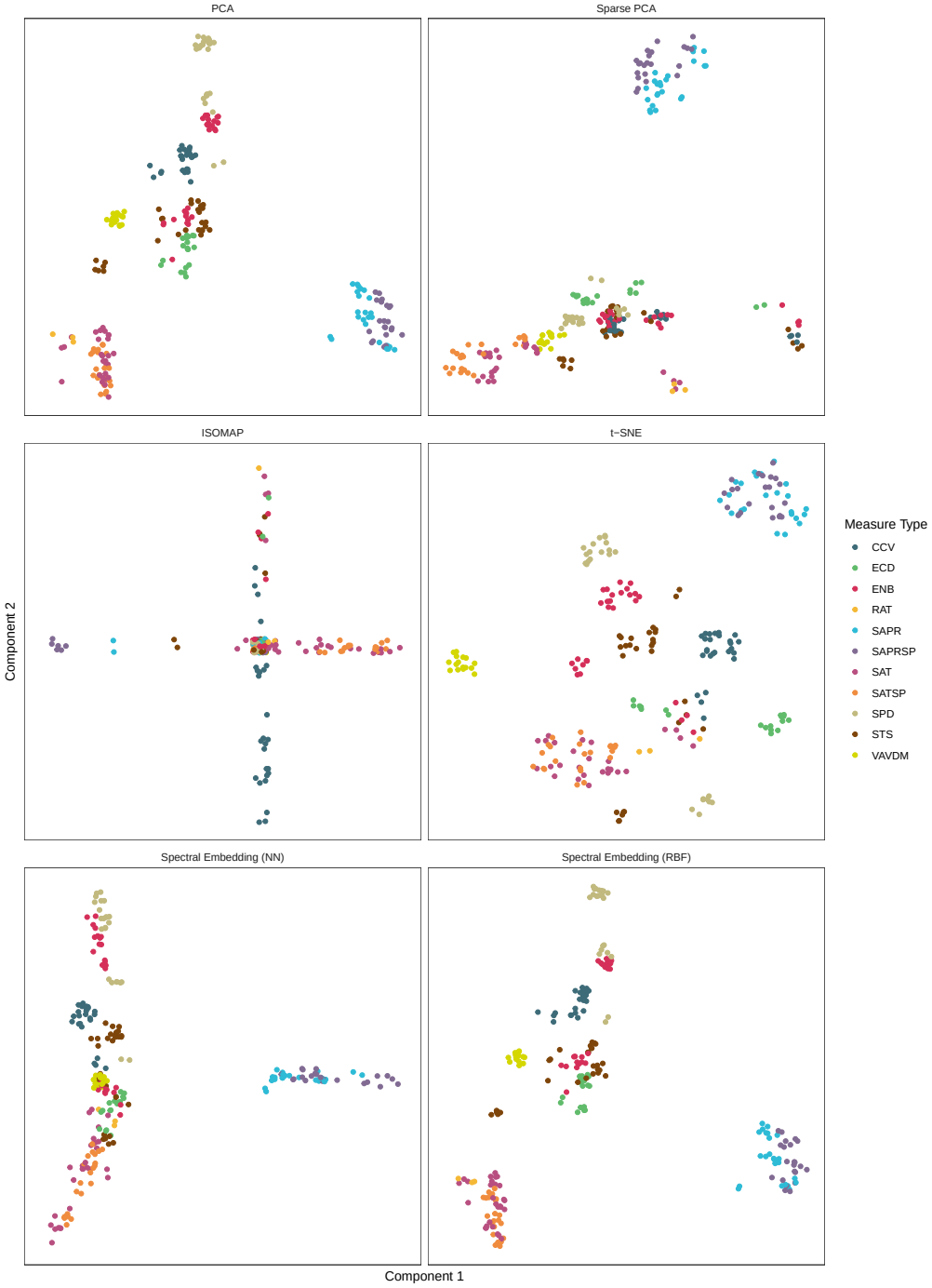

Figure 4: *Dimension reduction on AHU sensors using time series and metadata features. Each point represents a sensor. The output from each approach has been scaled to fall between 0 and 1 for easy comparison. A small amount of jitter has been added to separate points that fall on top of each other.*

Hence, the resulting dimensionality reduction doesn't so much show the different meter types present, but instead the different buildings. As such it seems sensible to only apply dimension reduction to buildings with similar naming conventions if text features are to be used[2]. It should be noted that Figure 4 only uses the two buildings with similar naming conventions to avoid this issue.

One promising point is that dimensionality reduction appears to work reasonably well for all three buildings when using only the time series features. Better separation between classes does occur when n-grams are incorporated, but as mentioned only if naming conventions between buildings are similar. If all buildings are to be included it may be best to only look at time series features.

5.3 Implementation and example

In essence, our methodology then becomes a matter of feature generation, dimensionality reduction and exploration by a domain expert. Interactive visualisations should be used to identify anomalous points. An example implementation is shown in Figures 6 and 7. Users can quickly observe all sensors of interest in a low-dimensional space and then select those that appear to be far away from others of the same type. In Figure 6 we notice that one of the CCV points is far from the others. After selecting it we can clearly see short cycling is occurring when we inspect the raw time series values. The CCV is repeatedly opening and closing resulting in wasted energy usage. Another example is shown in Figure 7. In this case we can see that the CCV point is only opening for two hours each day, which may indicate faulty behaviour. However, while unusual, this would need further investigation to decide if it is in fact a fault.

5.4 Limitations and future research

A potential shortcoming of our methodology is that due to its reliance on historical data it may project certain faulty and normal sensors to similar regions of the low dimensional space. For instance, a stuck damper and a damper that is supposed to be closed all the time will have very similar features. Any fault classification that follows will incorrectly include the intentionally closed damper with the stuck damper if they are classified as faulty. However, false positives such as these are a common shortcoming and we leave it as a topic for future research. Considering other variables and how they interact, such as if the supply air fan is running or shut off, may help to address this issue.

[2]Learning methods such as correspondence analysis may allow us to determine which text features are actually useful for distinguishing different measures. However this relies on using the sensor labels and so is not an unsupervised technique and falls beyond the scope of this paper.

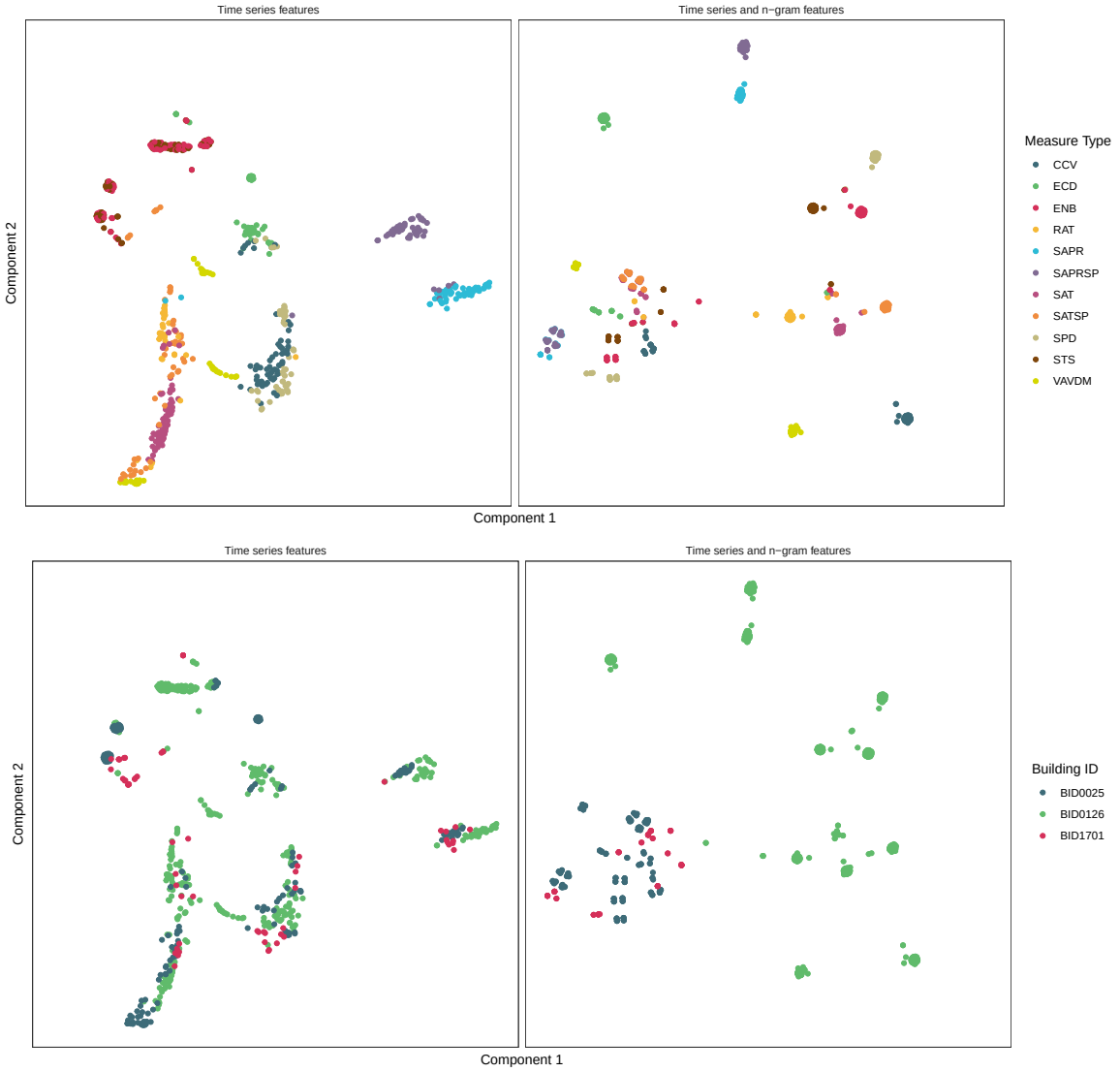

Figure 5: *Dimension reduction using t-SNE on AHU sensors for multiple buildings. The top row shows clear separation between measure types when only time series features are used (left), but poor separation when using metadata features are included (right). The bottom row makes the cause of this clear. Building BID0126 has a different naming convention to the others. Points appear to be separated by building rather than measure types.*

Another limitation of our methodology is that it focuses only on fault detection and does not attempt to diagnose the causes. However, there are already many supervised diagnosis methods available (Frank et al. 2016) and are mainly dependent on a comprehensive data set of fault-symptom relationships being available (Gunay, Shen & Yang 2017). Our paper's focus is on developing a suitable methodology to make it easier to develop such a data set which might apply to multiple buildings.

Analysing sensor behaviour between buildings is a difficult task as it does not suffice to simply compare the raw time series data from one sensor to that of an identical sensor type in another building. Other factors such as temperature set points, outside weather conditions

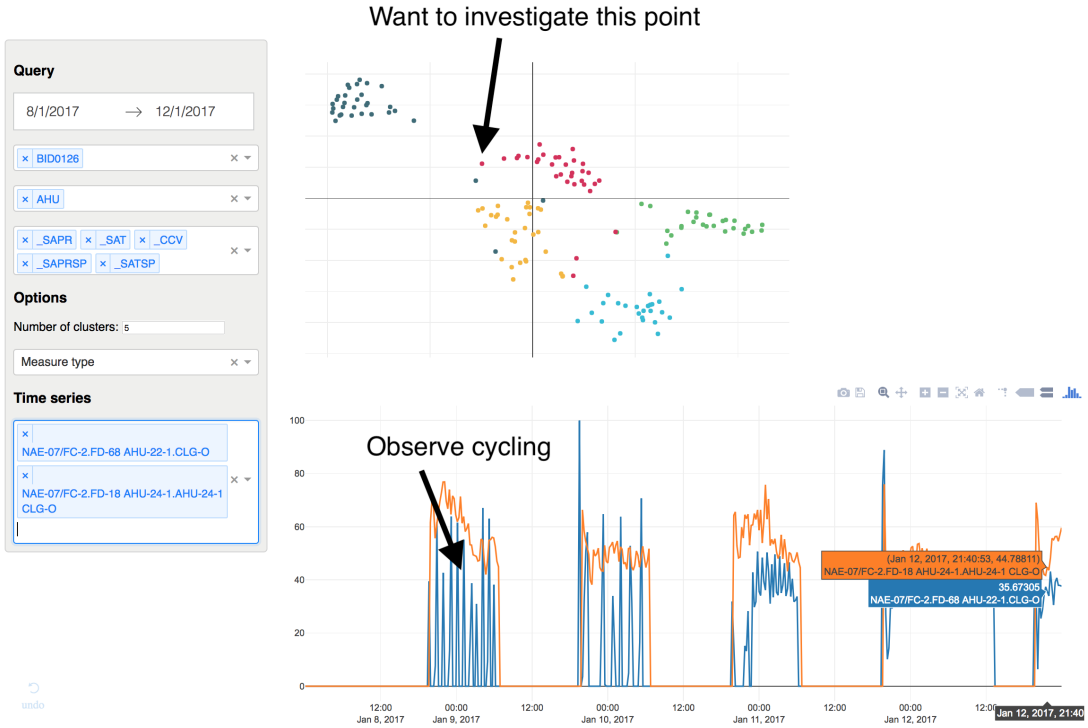

Figure 6: *Control valve short cycling. One of the CCV points (red) is far from the others. Selecting this point quickly reveals short cycling occurring in the anomalous point. Another point from the main cluster has been selected to illustrate what CCV readings from a normal point should look like.*

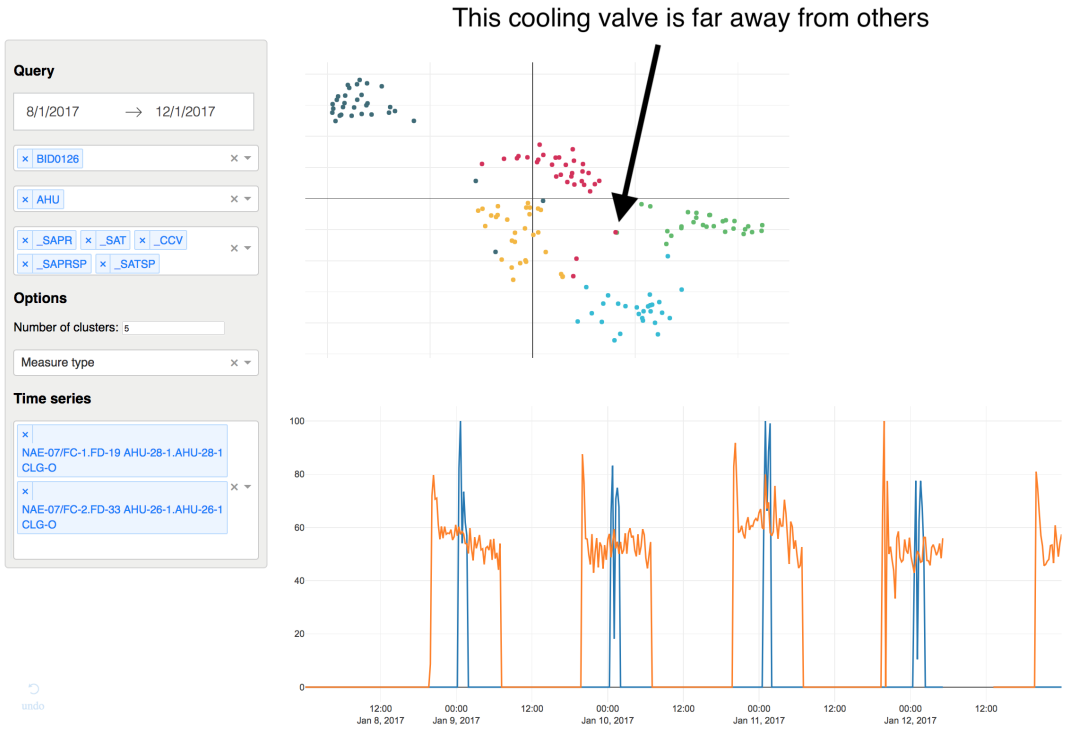

Figure 7: *Cooling control valve only open for two hours. Again, another CCV point is far from the main cluster. In this case we have quickly identified unusual behaviour that may indicate a fault, but needs further investigation.*

and occupancy should all be accounted for in order to create a fair comparison. Incorporating environmental factors each building is subject to could potentially improve separation between functioning and faulty points.

6 Conclusion

In this paper we have presented a methodology for detecting anomalous sensor behaviour within multiple buildings. We engineer various features based on time series from sensor readings and sensor metadata. Several dimensionality reduction algorithms are tested with t-SNE appearing to give the most useful two dimensional representation. Our low dimensional space allows users to easily identify anomalous points.

Future directions for this data set have been outlined with the most promising direction being improving comparisons of sensor behaviour between buildings. An implementation and example using the proposed methodology was also presented which allows a domain expert to quickly inspect many different equipment and measurement points for unusual behaviour.

Acknowledgements

This research is supported by funding from Buildings Alive. The author would also like to thank Buildings Alive for making data available and their guidance on understanding commercial office building behaviour.

References

Ardakani, MH, A Shokry, G Escudero, M Graells & A Espuña (2016). A framework for unsupervised fault detection and diagnosis based on clustering assisted kriging observer. In: *2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol)*, pp.183–188.

Brand, M & K Huang (2003). A unifying theorem for spectral embedding and clustering. In: *9th International Conference on Artificial Intelligence and Statistics*. Key West, FL.

Costa, BSJ, PP Angelov & LA Guedes (July 2014). A new unsupervised approach to fault detection and identification. In: *2014 International Joint Conference on Neural Networks (IJCNN)*, pp.1557–1564.

Costa, BSJ, PP Angelov & LA Guedes (Feb. 2015). Fully unsupervised fault detection and identification based on recursive density estimation and self-evolving cloud-based classifier. *Neurocomputing* **150**, 289–303.

Frank, S, M Heaney, X Jin, J Robertson, H Cheung, R Elmore & G Henze (Aug. 2016). *Hybrid Model-Based and Data-Driven Fault Detection and Diagnostics for Commercial Buildings: Preprint*. en. Tech. rep. NREL/CP-5500-65924. National Renewable Energy Lab. (NREL), Golden, CO (United States). https://www.osti.gov/scitech/biblio/1290794.

Friedman, J, T Hastie & R Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer.

Gunay, B, W Shen & C Yang (June 2017). Characterization of a building's operation using automation data: A review and case study. *Building and environment* **118**, 196–210.

Hyndman, RJ, E Wang & N Laptev (Nov. 2015). Large-Scale Unusual Time Series Detection. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp.1616–1619. http://dx.doi.org/10.1109/ICDMW.2015.104.

Katipamula, S & MR Brambley (Jan. 2005). Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part I. *HVAC&R Research* **11**(1), 3–25.

Kim, W & S Katipamula (2018). A review of fault detection and diagnostics methods for building systems. *Science and Technology for the Built Environment* **24**(1), 3–21. http://dx.doi.org/10.1080/23744731.2017.1318008.

Maaten, Lvd & G Hinton (2008). Visualizing Data using t-SNE. *Journal of machine learning research* **9**(Nov), 2579–2605. http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

Murphy, KP (Aug. 2012). *Machine Learning: A Probabilistic Perspective*. en. MIT Press.

Saul, LK, KQ Weinberger, JH Ham, F Sha & DD Lee (2006). "Spectral methods for dimensionality reduction". In: *Semi-Supervised Learning*. Ed. by O Chapelle, B Scholkopf & A Zien. Vol. 1. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, pp.293–308.

Tenenbaum, JB, V de Silva & JC Langford (Dec. 2000). A global geometric framework for nonlinear dimensionality reduction. en. *Science* **290**(5500), 2319–2323. http://dx.doi.org/10.1126/science.290.5500.2319.

Zou, H, T Hastie & R Tibshirani (June 2006). Sparse Principal Component Analysis. *Journal of computational and graphical statistics* **15**(2), 265–286. https://doi.org/10.1198/106186006X113430.

Chapter 6

Conclusion

Electricity analytics is rapidly moving towards more sophisticated techniques to better analyse the large amounts of data being collected. Regional grid data, smart-meter data and building sensor data all have important roles to play in reducing energy consumption and ensuring a stable grid, and so determining effective ways to analyse these large volumes of data is critical. Accurate demand forecasting and quantifying drivers of demand will help decision makers plan appropriate grid developments. There are many useful tools available to the statistician that allows them to play a role interpreting these vast data resources. Supervised machine learning techniques allow for accurate forecasting. Unsupervised dimension reduction algorithms allow for exploration of the data in a manageable fashion. Statistically significant conclusions can be reached by fitting interpretable statistical models.

The primary aim of this thesis was to improve electricity forecasting procedures and to better understand energy usage within buildings. Each of the data sources mentioned above have been analysed and used to produce new methodologies for both forecasting and inference. Techniques including gradient boosting, hierarchical reconciliation, mixed effects models, multi-model inference and dimension reduction have been used to reach important conclusions about the data or achieve improved forecasting performance.

6.1 Summary of research and contributions

Each of the main chapters of this thesis is a self-contained paper. In this section we summarise the main contributions of each piece of research.

Probabilistic forecasting within hierarchies is a common task when working with electricity demand. In Chapter 2, an approach for producing consistent probabilistic forecasts in a hierarchical setting is presented. The resulting forecasts improve when the hierarchical reconciliation stage is conducted, with the best results occurring when model residuals were used in the weight matrix. Our model showed superior forecasting performance when compared against Tao's Vanilla Model (Hong, 2010) which has been used in each global energy forecasting competition as a benchmark (Hong, Pinson, and Fan, 2014; Hong et al., 2016; Hong, Xie, and Black, 2019).

In Chapters 3 and 4 we focus on analysing smart meter data. Chapter 3 explores conducting inference on building characteristics using mixed effects models. Despite the importance of energy disaggregation, few papers have attempted to approach this problem in a statistical manner. In this chapter we presented an approach to produce demand impact profiles, that showed how each building characteristic increased or decreased demand when an attribute was either added or removed. Furthermore, the statistical significance could be evaluated for each attribute at different times of the day and year.

Chapter 4 explores using mixed effects models to improve forecast accuracy. Several papers have discussed forecasting smart-meter demand (Ben Taieb, Taylor, and Hyndman, 2020; Ben Taieb et al., 2017a, 2016; Arora and Taylor, 2016; Gajowniczek and Ząbkowski, 2014; Yildiz et al., 2017), though none have utilised mixed effects. One day ahead forecasts are tested using various model formulations. We find that our subject-specific curve formulation that incorporates an $AR(1)$ error structure improves point forecast accuracy.

Finally, Chapter 5 examines approaches to visualising and understanding building sensors in BMS systems. We discuss feature extraction for time series made up of sensor readings. We also explore including text-based features from sensor metadata and the impact it has on clustering. Having generated appropriate features, we show that t-SNE provides a sensible way of visualising each sensor in a 2-dimensional space. We provide examples

of how a facility manager can use this low-dimensional data to quickly identify unusual sensor behaviour and faults.

6.2 Limitations and future research

In Chapter 2 we proposed a method to construct hierarchically consistent probabilistic forecasts. Our approach relied on constructing simulations of demand and then hierarchically reconciling each simulation. Quantiles could then be calculated based on the reconciled simulations. This makes our approach suitable for cases when forecasts for each level of a hierarchy are all produced together. However, there may be occasions when probabilistic forecasts for different levels of a hierarchy are produced independently by different teams. In this case our methodology can not be used for reconciliation. Since the publication of our paper several others addressing this issue have been put forth (Ben Taieb, Taylor, Hyndman, et al., 2017b; Ben Taieb, Taylor, and Hyndman, 2020).

Chapter 3 highlights the potential of combining complementary data sets with smart meter demand data. To the best of my knowledge this was the first paper to attempt this using mixed effects models, and so I focused on producing a relatively simple formulation that would be easy to understand. There is no doubt that more sophisticated models could capture extra information in the data. In fact, we go on to show in Chapter 4 that including temperature variables improves forecasting accuracy at different times of the day. Future research can focus on extending the model we presented for other complementary data sets.

As discussed, probabilistic forecasting for electricity demand is important due to its inherently stochastic nature. Point forecasts alone are inadequate for decision making as they fail to capture the risk of unexpected spikes or drops in demand. In Chapter 4 we focused on improving point forecasts for smart-meter data using mixed effects models. This was adequate for conducting scenario analysis and improving point forecast accuracy. A follow up paper discussing probabilistic forecasting using mixed effects models could provide useful methodological contributions. Constructing prediction intervals and assessing their pinball-loss scores against suitable benchmark models (Ben Taieb et al., 2016) would be an interesting empirical study.

Chapter 5 examined different approaches to dimensionality reduction applied to BMS point data. Future research could include more sophisticated text and time series feature extraction which may provide better separation between normal and faulty points in low dimensions. Presenting more real-world BMS fault detection case studies using this approach could also provide some benefit.

6.3 Reproducibility

All code to reproduce published papers has been made available in dedicated GitHub repositories in an effort to aid reproducibility. A tidied data set for the GEFCom2017 competition has been released as an R package on GitHub.

Bibliography

Arora, S and JW Taylor (2016). Forecasting electricity smart meter data using conditional kernel density estimation. *Omega* **59**, 47–59.

Australian Energy Market Commission (2015). *Expanding competition in metering and related services*. Tech. rep. https://www.aemc.gov.au/sites/default/files/content/ed88c96e-da1f-42c7-9f2a-51a411e83574/Final-rule-determination-for-publication.pdf.

Australian Energy Market Operator (2019). *Electricity Demand Forecasting Methodology Information Paper*. Tech. rep. https://www.aemo.com.au/-/media/Files/Stakeholder_Consultation/Consultations/NEM-Consultations/2019/Reliability-Forecasting-Methodology/Electricity-Demand-Forecasting-Methodology-Information-Paper---draft-2019.pdf.

Ben Taieb, S, R Huser, RJ Hyndman, and MG Genton (2016). Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression. *IEEE Transactions on Smart Grid* **7**(5), 2448–2455.

Ben Taieb, S, JW Taylor, and RJ Hyndman (2020). Hierarchical Probabilistic Forecasting of Electricity Demand with Smart Meter Data. *Journal of the American Statistical Association* (just-accepted), 1–36.

Ben Taieb, S, J Yu, M Neves Barreto, and R Rajagopal (2017a). Regularization in Hierarchical Time Series Forecasting With Application to Electricity Smart Meter Data. In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Ben Taieb, S, JW Taylor, RJ Hyndman, et al. (2017b). *Coherent Probabilistic Forecasts for Hierarchical Time Series*. Tech. rep. Working Paper 03/17. Monash University, Department of Econometrics and Business Statistics. http://business.monash.edu/

econometrics-and-business-statistics/research/publications/ebs/wp03-17.pdf.

Chan, E and J Boddington (2019). *Keeping our eye on smart meters rollout*. Accessed: 2019-10-23. https://www.aemc.gov.au/news-centre/keeping-our-eye-smart-meters-rollout.

Esteves, GRT, BQ Bastos, FL Cyrino, RF Calili, and RC Souza (2015). Long Term Electricity Forecast: A Systematic Review. *Procedia Computer Science* **55**, 549–558.

Flath, C, D Nicolay, T Conte, C van Dinther, and L Filipova-Neumann (2012). Cluster Analysis of Smart Metering Data. *Business & Information Systems Engineering* **4**(1), 31–39.

Gajowniczek, K and T Ząbkowski (2014). Short Term Electricity Forecasting Using Individual Smart Meter Data. *Procedia Computer Science* **35**, 589–597.

Haben, S, C Singleton, and P Grindrod (2016). Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. *IEEE Transactions on Smart Grid* **7**(1), 136–144.

Hong, T (2010). "Short term electric load forecasting". PhD thesis. North Carolina State University.

Hong, T and S Fan (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting* **32**(3), 914–938.

Hong, T, P Pinson, and S Fan (2014). Global energy forecasting competition 2012. *International Journal of Forecasting* **30**(2), 357–363.

Hong, T, P Pinson, S Fan, H Zareipour, A Troccoli, and RJ Hyndman (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting* **32**(3), 896–913.

Hong, T, J Xie, and J Black (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting* **35**(4), 1389–1399.

Hyndman, RJ, RA Ahmed, G Athanasopoulos, and HL Shang (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* **55**(9), 2579–2589.

Hyndman, RJ, AJ Lee, and E Wang (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis* **97**, 16–32.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and É Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR* **12**(Oct), 2825–2830.

Pérez-Chacón, R, JM Luna-Romera, A Troncoso, F Martínez-Álvarez, and JC Riquelme (2018). Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities. *Energies* **11**(3), 683.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Räsänen, T and M Kolehmainen (2009). Feature-Based Clustering for Electricity Use Time Series Data. In: *Adaptive and Natural Computing Algorithms*. Springer Berlin Heidelberg, pp.401–412. http://dx.doi.org/10.1007/978-3-642-04921-7_41.

Research in China (2019). *China Smart Meter Industry Report, 2019-2025*. Tech. rep. http://www.researchinchina.com/Htmls/Report/2019/10537.html.

Singh, AK, Ibraheem, S Khatoon, M Muazzam, and DK Chaturvedi (2012). Load forecasting techniques and methodologies: A review. In: *2012 2nd International Conference on Power, Control and Embedded Systems*. IEEE, pp.1–10. http://dx.doi.org/10.1109/ICPCES.2012.6508132.

Suganthi, L and AA Samuel (2012). Energy models for demand forecasting—A review. *Renewable and Sustainable Energy Reviews* **16**(2), 1223–1240.

Taylor, JW, LM de Menezes, and PE McSharry (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting* **22**(1), 1–16.

The State of Victoria Department of Environment, Land, Water and Planning (2016). *Transition to Metering Competition in Victoria*. Tech. rep. https://www.energy.vic.gov.au/__data/assets/pdf_file/0021/43581/20161003-Final-metering-comp-consultation-paper.pdfdocx.pdf.

U.S. Energy Information Administration (2018). *How many smart meters are installed in the United States, and who has them?* Accessed: 2019-10-23. https://www.eia.gov/tools/faqs/faq.php?id=108&t=3.

Wang, Y, Q Chen, T Hong, and C Kang (2018). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid* **10**(3), 3125–3148.

Wickramasuriya, SL, G Athanasopoulos, RJ Hyndman, et al. (2015). Forecasting hierarchical and grouped time series through trace minimization. *Department of Econometrics and Business Statistics, Monash University* (Working Paper 15/15).

Yildiz, B, JI Bilbao, J Dore, and AB Sproul (2017). Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy* **208**, 402–427.

Yildiz, B, JI Bilbao, and AB Sproul (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews* **73**, 1104–1122.

Yu, N, S Shah, R Johnson, R Sherick, M Hong, and K Loparo (2015). Big data analytics in power distribution systems. In: *2015 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp.1–5. http://dx.doi.org/10.1109/ISGT.2015.7131868.