Nonlinear mixed effects models 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

11 December 2020

JEL classification: C10,C14,C52

Nonlinear mixed effects models 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 nonlinear mixed models. The predictive performance of the mixed effects models are tested against naive, linear and nonlinear benchmark models fitted to each building separately. 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. Mixed effects models are useful when non-independence is observed in the data, which often occurs when multiple observations are recorded from the same subject. 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. Electricity demand readings at 15-minute intervals are taken from smart meters located at each of the buildings. These readings exhibit within-subject correlation which are well suited to modelling using mixed effects models, which we explore here. 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.

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 for buildings with different characteristics.

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. These subject specific curves are nonlinear functions that are allowed to vary by each subject. In this paper we first justify the use of subject-specific curves for modelling metered electricity demand before building on the existing methodology using 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, our methodology is tested against several other benchmark models to quantify how forecasting accuracy is improved. 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 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

Electricity demand data from 123 commercial office buildings in Australia are used to train and validate our models. The electricity demand data is recorded at 15-minute intervals by smart meters located at each building. 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. The different relationships for each building necessitates the use of a nonlinear demand response function that can adequately capture the unique response characteristics of each building. To accomplish this, we model our data with mixed effects models and capture the nonlinear building specific demand response using subject-specific curves. Note that when splines are fitted independently for each building they can be very sensitive to outliers or data at domain extremes. This also motivates our use of mixed effects models where data from multiple buildings can be used to fit these relationships giving models that are more robust to outliers or sparse data at covariate 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.

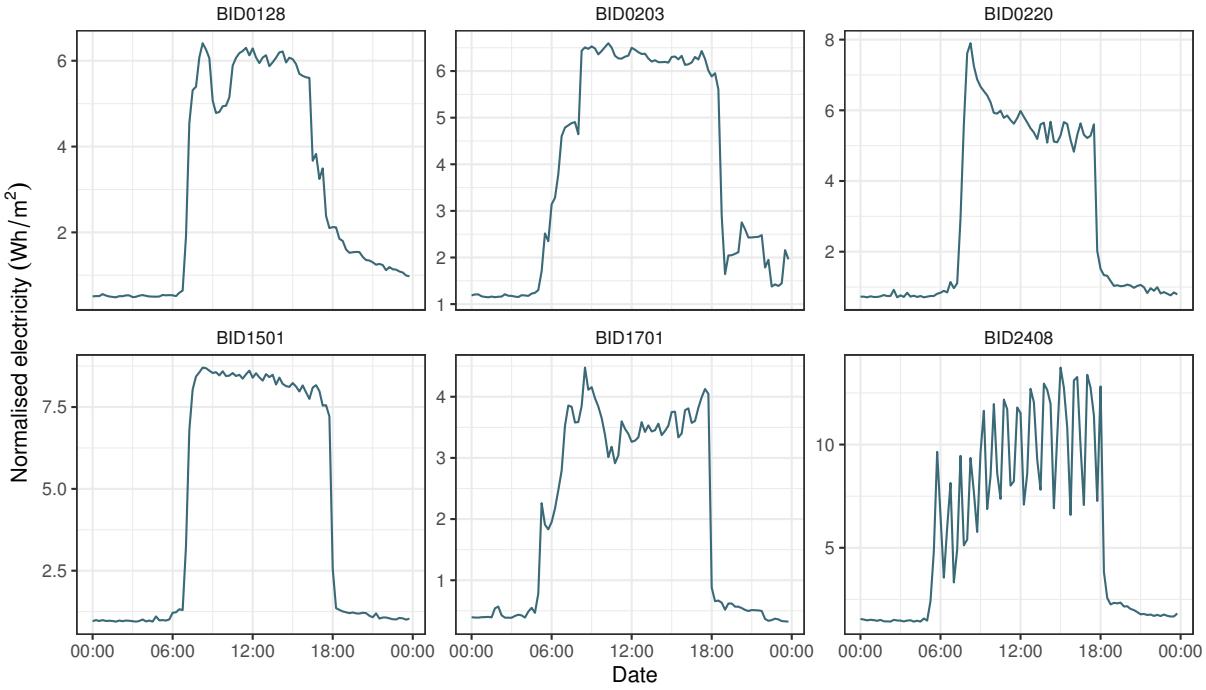

Figure 1: *Normalised electricity demand of six commercial office buildings in Australia on 9 January, 2017. Only one day of data is shown although it is enough to see clear differences in the demand profiles. Both the magnitude and volatility of demand varies greatly between buildings.*

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.

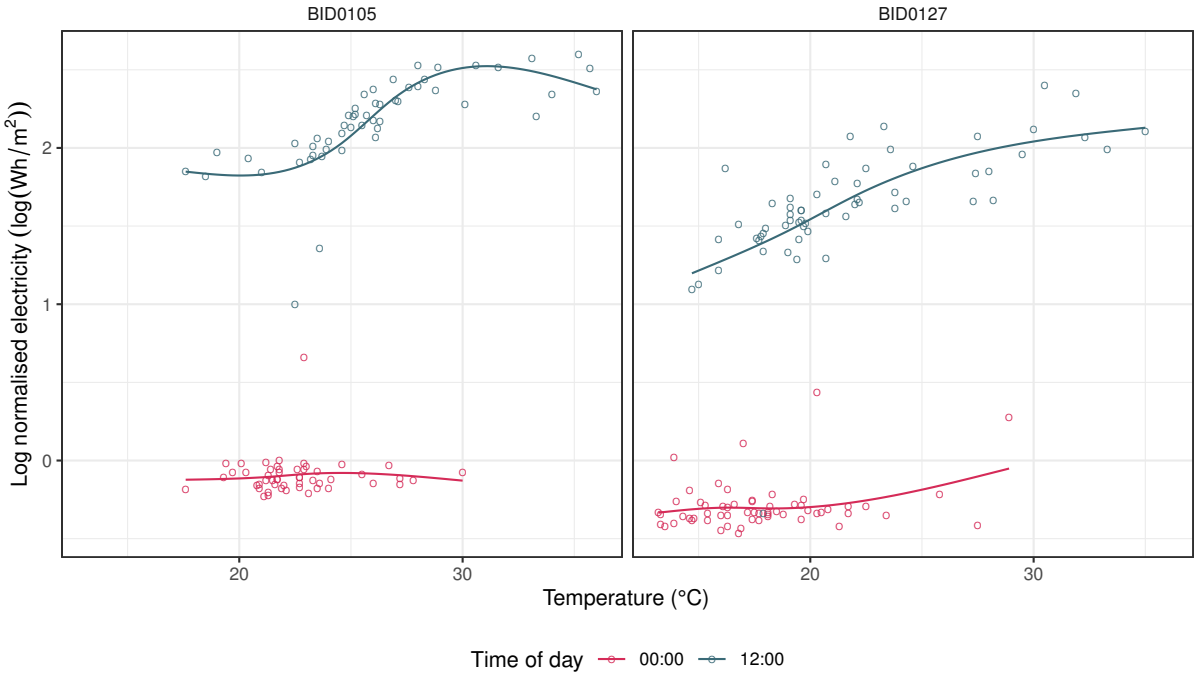

Figure 2: *Relationship between temperature and normalised electricity demand at midday and midnight for two Australian office buildings during Summer. Logged values are shown as we use these as our response variable when fitting models to enforce a positivity constraint. The relationship between temperature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

3 Methodology

We focus on the problem of one day ahead electricity demand forecasting for commercial buildings. This has important applications such as allowing facility managers to adequately prepare a building for demand response (by adjusting set points and operational schedules) and reducing peak demand to avoid high capacity[2] or time of use charges.

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 and mixed effects models and their formulations. All analysis was produced using the R statistical programming language (R Core Team 2020). Mixed effects models were fit using the `lme4` package for mixed effects models (Bates et al. 2019). All code used to run this analysis has been made available in the https://github.com/camroach87/1901-nlmets GitHub repository.

[2]Peak demand events over a certain time frame are often factored into a commercial building's electricity tariffs. These are known as capacity charges.

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

3.1 Model formulation

To justify our final model 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.

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 fit 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). An example for temperature and electricity demand is shown in Figure 3. We see that a natural spline gives a more reasonable fit at the sparsely populated extremes compared to other spline functions.

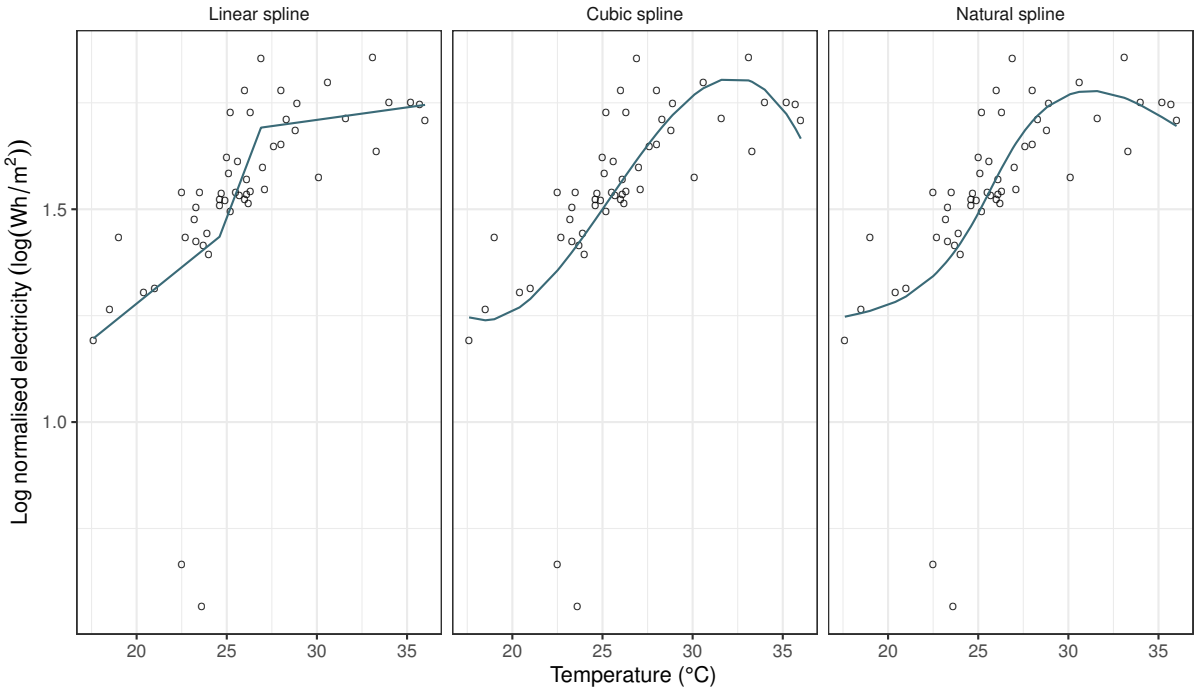

Figure 3: *Linear and cubic splines with three degrees of freedom fit to one building's demand data during Summer at 11:45 am. We observe that the linear spline results in severe kinks in the relationship which seems unrealistic. The cubic spline gives a much smoother fit and appears to be the more reasonable 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.*

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[3] 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

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

where f_p is a smooth function modelling the relationship between $w_{0,t}$ and the logged demand for period p. The function f_p is modelled using natural cubic splines with three degrees of freedom. Knots are set at the 33.3^{rd} and 66.6^{th} quantiles of $w_{0,t}$. 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. 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 extreme temperatures (see Figure 3 for an illustration). We wish to create a parsimonious model and assuming anything beyond a linear relationship in the

[3]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.

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. When doing one day ahead forecasts we do not have demand observations within the last 24 hours to use as lags and so we restrict our lagged demand variables to 1 day, 2 days and 3 days. The temperature variables can have 1-24 hour lags as temperature forecasts can be used to supplement the already observed data. Note that in this paper we restrict ourselves to using actual temperature data and not forecast temperature data to ensure our results are dependent on model formulations and not on any errors in temperature forecasts. In practice, forecast temperature data can be used instead. 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

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

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. Again, the function $f_{i,p}$ is modelled using natural cubic splines with three degrees of freedom and knots set at the 33.3$^{\text{rd}}$ and 66.6$^{\text{th}}$ quantiles of each predictor $x_{b,i,t}$. 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 predictor 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 temperature and lagged demand 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

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

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

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

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 nonlinear 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

$$\log y_{bt} = \sum_{x_{b,i,t} \in \mathcal{P}_t} f_{i,p}(x_{b,i,t}) + g_{b,p}(w_{b,0,t}) + \epsilon_{b,t},$$
$$\epsilon_{b,t} \sim N(0, \sigma^2_{\epsilon,p}).$$

As with our other models we use natural cubic splines with three degrees of freedom for our smooth function $f_{i,p}$. The function $g_{b,p}$ is a subject specific curve and is defined as

$$g_{b,p}(x) = u_{b,p,0} + \sum_{k=1}^{3} u_{b,p,k} B_{p,k}(x),$$
$$\left(u_{b,p,0}, \ldots, u_{b,p,3}\right) \sim N(0, \Sigma),$$

where $B_{p,k}$ are cubic B-spline basis functions and $u_{b,p,k}$ are correlated random effects. 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 Here 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 is given below

$$\log y_{bt} = \sum_{x_{b,i,t} \in \mathcal{P}_t} f_{i,p}(x_{b,i,t}) + g_{b,p}(w_{b,0,t}) + \sum_{a \in \mathcal{A}} \beta_a x_{b,a} + \epsilon_{b,t},$$
$$\epsilon_{b,t} \sim N(0, \sigma^2_{\epsilon,p}),$$

where all terms and functions are defined as in the SSC model. The new fixed effect $x_{b,a}$ is a Boolean variable that indicates if attribute a is present for building b. Including building attributes as fixed effects allows one to conduct scenario analyses by predicting the expected demand with and without certain attributes present.

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 m 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 m 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 m. During feature selection we chose to avoid greedy approaches such as forward or backward stepwise selection; or approaches that work systematically through lagged predictor variables (Hyndman & Fan 2010). Naturally, greedy methods have computational benefits, but it is interesting to observe which features are chosen when *all* possible combinations are assessed.

A key point to note is that feature selection was done on the predictor 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.

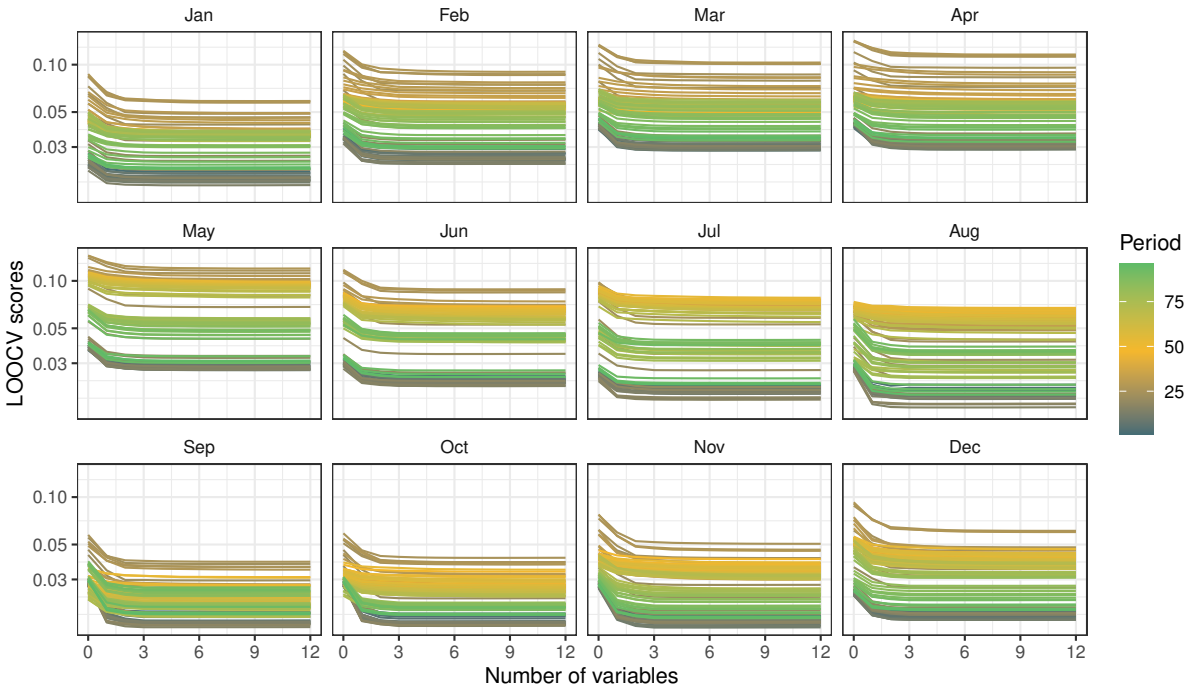

Figure 4: *LOOCV scores (log scale) for each month. In general, the LOOCV errors initially decrease as variables are added, but begin to increase slightly at a certain point for each period.*

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.

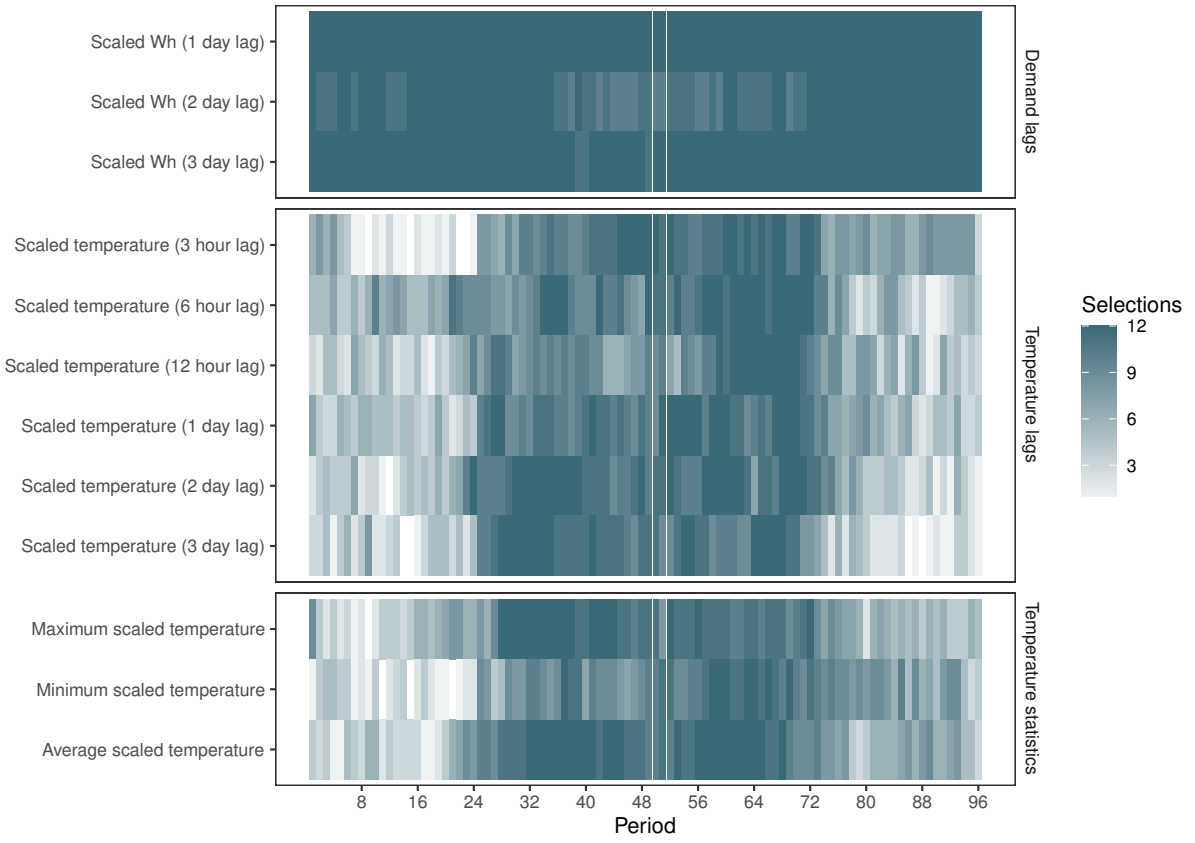

Figure 5: *Feature selection for all months. The number of times a variable has been selected across all 12 months is indicated by the transparency of each tile for a given period. In general, temperature features are selected more often during business hours. During non-business hours, temperature features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

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.

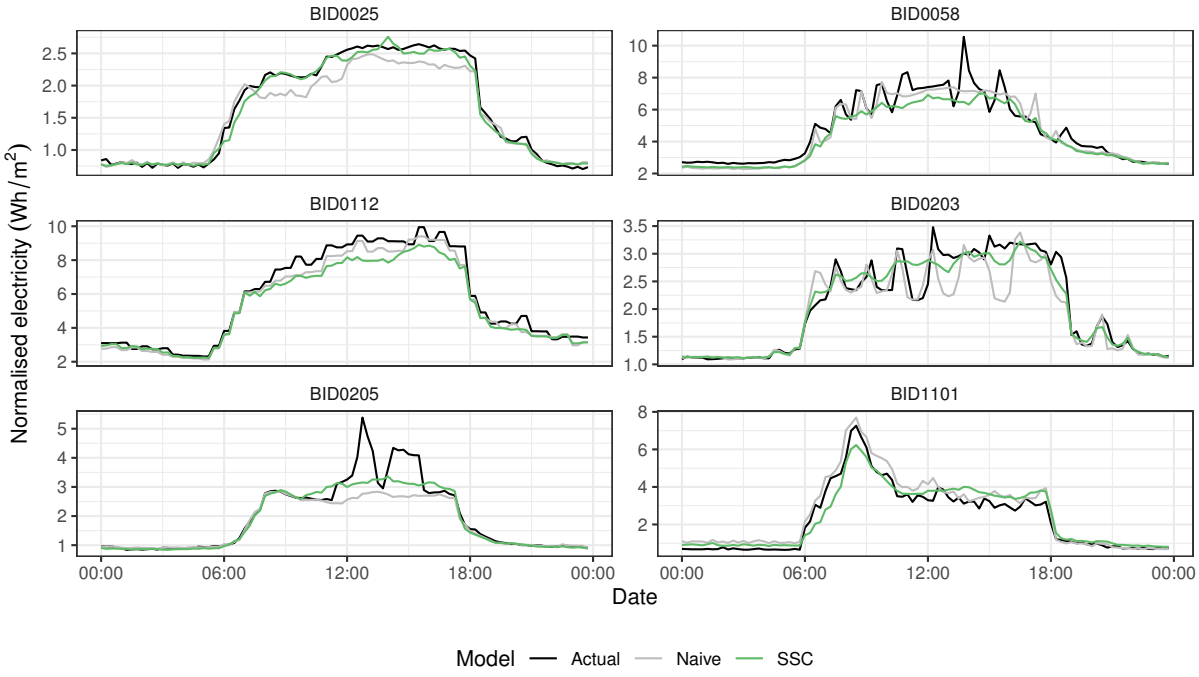

Figure 6: *One day ahead forecasts for 23 August, 2017. The Naive (yesterday's actuals) and SSC models are shown. The naive model often includes erratic spikes whereas the SSC model tends to produce a smoother profile.*

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

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**

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.

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.

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	***

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 electricity demand using nonlinear mixed effects models can improve forecast accuracy.

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 (2009). Workload forecasting for a call center: Methodology and a case study. *Annals of Applied Statistics* **3**(4), 1403–1447.

Arora, S & JW Taylor (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 (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 (2002). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* **20**(1), 134–144.

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

Fan, S & RJ Hyndman (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 (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 (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 (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 (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting* **35**(4), 1389–1399.

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

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

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

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

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

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

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

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

Soyer, R & MM Tarimcilar (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 (2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics & Data Analysis* **53**(10), 3616–3629.