Modelling household electricity demand

Euan Enticott, Georgina Mansell, and Conor Newton

May 28, 2021

Abstract

Large changes in the nature of electricity production and consumption are expected over the coming years. Recent demand-side shocks have highlighted the importance of being able to provide accurate consumption forecasts, not only at a high level of aggregation but also for individual households. This poses challenges due to the low signal-to-noise ratio in household-level demand. We undertake an analysis of Irish smart meter data, collected at half hourly intervals. We cluster the households according to their daily demand profiles; the idea being that by fitting separate models to households who show similar demand profiles, we can borrow relevant information across households. We fit a penalised ridge regression model to each cluster of households at each time of day. Regression coefficients are sampled from a Bayesian posterior to evaluate credible intervals. We find that by modelling on individual clusters and time points we are able to both better predict aggregate demand whilst also providing accurate predictions to each individual household. We obtain an MAE of 0.142 when fitting a ridge regression to the aggregate data and an MAE of 0.0352 when aggregating the predictions from each cluster, demonstrating the advantage of this method.

1 Introduction

Over the next 50 years it is expected that the demand for electricity will increase sharply. The prominence of electrified transport is set to increase whilst more focus will be put into renewable production. This shift away from oil and gas will have marked effects on electricity forecasting. Electricity providers need to be able to accurately predict demand at least one day ahead so that they can ensure they have an adequate supply.

Renewably generated electricity is not able to react quickly in response to changes in demand therefore it is essential that they have accurate models informing how much supply they need. In particular, due to the high cost of energy storage it will be important to forecast not only the aggregate demand but individual demand as well, so they are able to correctly distribute storage to where it is needed.

Generalised Additive Models have frequently been employed to tackle this problem [1] [2]. We approach this problem by clustering the data by a households daily demand profiles. In this way we can borrow information across households with similar demand profiles which should make prediction at the household level more accurate. We also create a separate model for each time point with the idea being that the effect of features should be similar at the same time point but may vary across the day. For example, at 3am temperature is unlikely to have much of an effect on demand as most people will be asleep with no need for heating. However, in the middle of the day temperature could

have a much greater effect. Thus for each cluster we require 48 separate ridge regressions, which poses significant computational challenges.

Alongside this report, we present an R package *RcppRidge* which combats the computational issues by fitting the models with C++, parallelising across time points, and using optimised procedures for selecting hyperparameters.

2 Data

A dataset of residential electricity demand was collected as part of a household behaviour trial in 2010 in Ireland. Electricity consumption was recorded for 2672 households at half hour intervals over the year.

Figure 1 shows the total electricity demand profile for the households. There is a characteristic pattern of low demand at night, a low peak around 7am and a higher peak around 6pm. There are some seasonal effects where demand is higher in the winter months, likely due to increased heating and lighting. The profile also differs on weekends where there is no dip in demand during typical work hours.

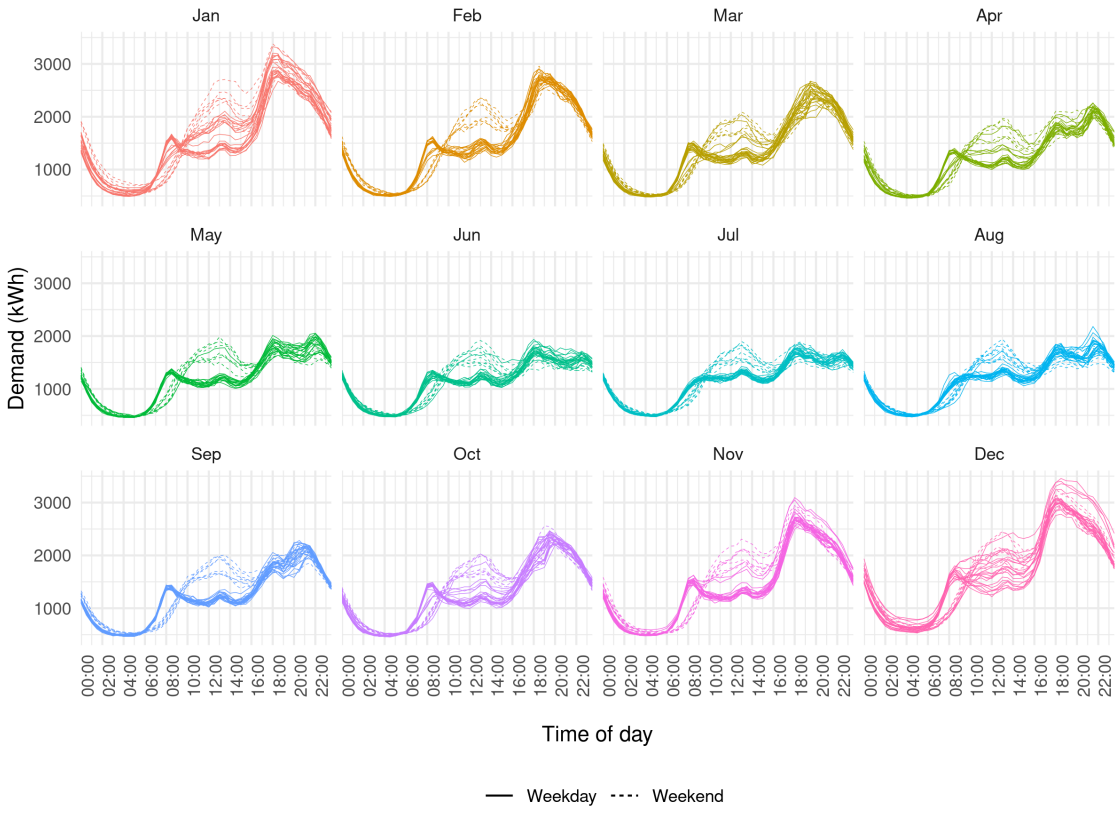

Figure 1: Total electricity demand over all recorded households; each line shows the usage over one day, split over subfigures by month.

While there is a very clear pattern to the demand at an aggregated level, individual household profiles are much more varied.

The dataset comes with additional information such as the external temperature, the year the house was built, and whether the house has electric heating.

The dataset is available at `https://github.com/mfasiolo/electBook/blob/master/data/Irish.RData`, and more information about the original study can be found at `http://www.ucd.ie/issda/data/commissionforenergyregulationcer/`.

3 Methods

All code is available on GitHub at `https://github.com/g-l-mansell/RcppRidge`

3.1 Clustering

Clustering serves us two purposes. Firstly, it can be used to identify groups of customers in the dataset that have similar demand profiles. Secondly, it allows us to break up the data into smaller and more manageable chunks.

The Irish dataset contains over 40 million rows. Fitting a model to a dataset of this size can take a significant amount time. To improve this, we first propose to break the data into clusters from which we can fit individual models in parallel.

For the clustering process we can use k-means clustering [3], [4]. In essence this algorithm splits the data into k regions that are as compact as possible. This is done by minimising the *within-cluster sum of squares* (WCSS). The objective of k-means clustering is therefore to find

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} ||\mathbf{x} - \mu_i||^2 \tag{1}$$

where $\mathbf{S} = \{S_1, \ldots, S_k\}$ is a partition of the observations into k regions and μ_i is the mean of the cluster S_i.

We can use the following iterative procedure to find an approximate solution to equation 1 efficiently, as given in [4].

1. **Initialisation:** Choose k points that will act as the initial centers of each cluster. These points are called centroids.
2. **Assignment step:** Assign each sample to the cluster with the closest centroid .
3. **Update step:** For each cluster, update the centroid to be the mean of the points assigned to the cluster.

Steps 2 - 3 are repeated until the centroids do not change between iterations. We have implemented this procedure in our R package.

A pitfall of this algorithm is that it is not guaranteed to converge to the value given by equation 1. We can refine this algorithm by running it multiple times with different initial centroids (which are chosen randomly) and selecting the result that give the lowest total within-cluster sum of squares 1. This prevents a bad choice of initial centroids generating sub-optimal clusters.

3.1.1 Choosing an optimal number of clusters

To help us choose the optimal number of clusters we can use an elbow plot [5]. This considers the total within-cluster sum of squares for different numbers of clusters. A smaller total within-cluster sum of squares suggests that our clusters are more compact and perhaps better fitting. Although we need to be careful that using more clusters

will always decrease the total WCSS. Therefore, choosing the number of clusters in the "elbow" of the graph (the region where there is a significant change in gradient) can be a reasonable choice.

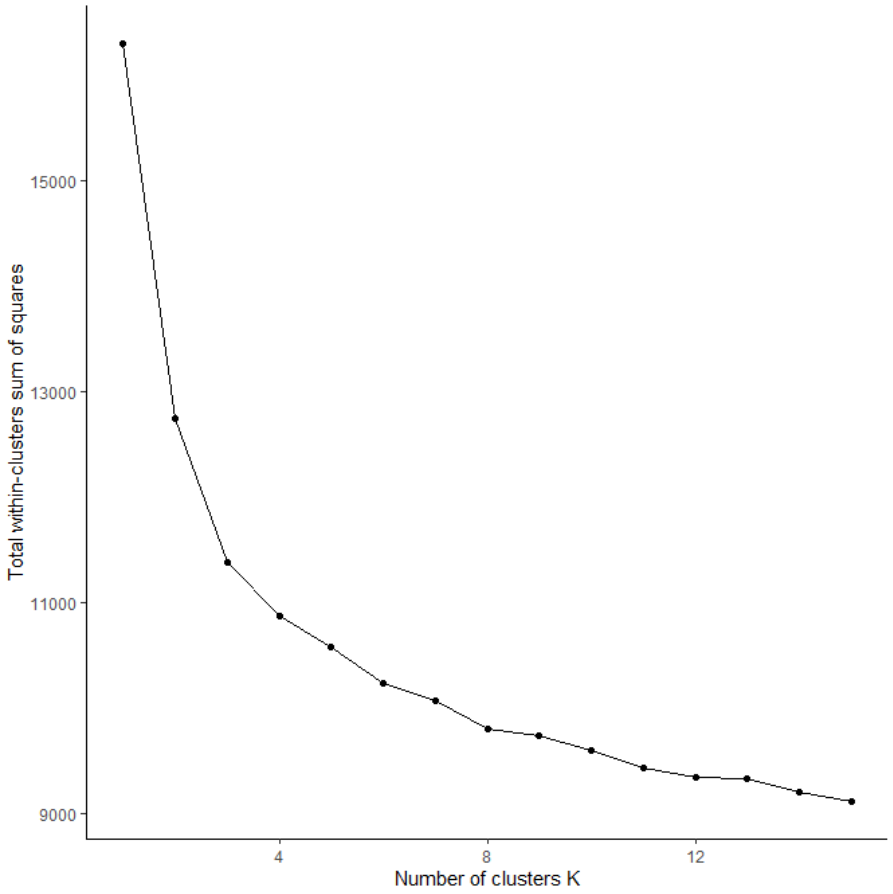

Figure 2: Elbow plot for k-means

Additionally, we can choose the number of clusters that appears to break up the data reasonably. We can visualise this by plotting the data on its two most significant principal components and colour each sample according to the cluster it belongs to.

Figure 3 displays the 5 clusters of different customers found by considering their electricity demand at different times of the day over the course of the full year. Table 1 gives a summary of the electricity demand in each cluster.

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
Mean	1.004	0.242	0.637	0.450	0.864
SD	0.218	0.072	0.071	0.063	0.109

Table 1: The mean and standard deviation of the electricity demand in each cluster.

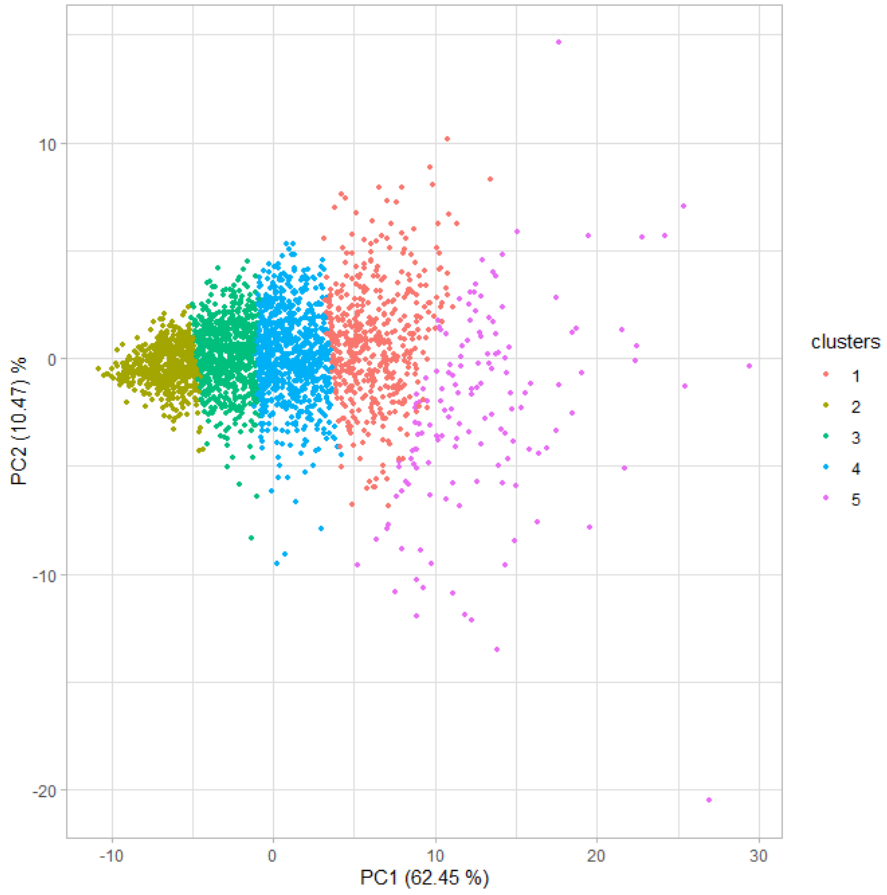

Figure 3: Customers from the Irish electricity demand data split into 5 clusters and plot against the 1st and 2nd principal components

3.2 Ridge Regression

Ridge regression [4] is a commonly used regularisation technique for least squares regression. In the standard setting we have a dataset of n samples each with a vector of p covariates $\mathbf{x}_i$ and scalar response y_i. The terms are collected in design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response vector $\mathbf{y} \in \mathbb{R}^n$. Typically the matrix $\mathbf{X}$ is standardised so that each variable has a zero mean and unit variance.

Ordinary least squares aims to find the vector of parameters $\beta \in \mathbb{R}^p$ which minimises the residual sum of squares:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{X}\beta||^2$$

However, OLS is susceptible to over-fitting; as the number of variables p increases, so does the model variance. Additionally, if $p > n$ or if there is multicollinearity between variables, the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible.

Ridge regression can overcome both of these issues by placing an L_2 penalty term on the parameter vector β:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||^2$$

Here $\lambda \geq 0$ is a hyperparameter which controls the strength of the regularisation and which should be selected through a cross validation procedure.

Solving for β gives:

$$\hat{\beta} = (\lambda \mathbf{I} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

3.3 Selecting λ

A simple way to select lambda is to calculate the mean square error on a held out portion of the dataset, then select the value of λ which minimises the MSE. An approach that is less sensitive to the specific choice of test data is leave-one-out cross validation (OCV). This is calculated as:

$$\text{OCV} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{[-i]} \right)^2$$

where $\hat{y}_i^{[-i]}$ represents the model prediction for sample i, where the model is fit using all samples except sample i.

It can be shown that OCV can be re-written as:

$$\text{OCV} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y})^2}{(1 - A_{ii})^2}$$

where

$$\mathbf{A} = \mathbf{X}(\lambda \mathbf{I} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

This calculation is less computationally expensive, as only one model is fit to all the data (per value of λ tested), rather than n separate models.

Calculation of the matrix $\mathbf{A}$ can also be optimised, using a singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U}$ is an $n \times p$ matrix, $\mathbf{D}$ is a diagonal matrix, and $\mathbf{V}$ is a $p \times p$ matrix. The SVD of $\mathbf{X}$ need only be computed once, then for each λ, $\mathbf{A}$ can be evaluated as:

$$\mathbf{A} = \mathbf{U} \, \text{diag}\{D_i^2/(D_i^2 + \lambda)\} \, \mathbf{U}^\top$$

3.4 Bayesian Perspective

Ridge regression can also be considered from a Bayesian perspective. This will allow us to quantify the uncertainty in a prediction by sampling from a posterior distribution for an unknown parameter.

Our Bayesian model for ridge regression follows [Section 3.3][3] closely and assumes that the responses y are generated by the following process:

$$y = \beta \mathbf{x} + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

This can be equivalently stated as:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \mathbf{I}\sigma^2)$$

And we assume that the parameters of the model have a Gaussian prior:

$$\beta \sim \mathcal{N}(0, \mathbf{I}\sigma^2/\lambda)$$

The posterior distribution of $\beta \mid y$ can then be shown to be:

$$\beta \mid y \sim \mathcal{N}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\sigma^2)$$

The mean of the posterior distribution is the parameter from the least squares estimator and the mode is $\hat{\beta}$, the parameter for a ridge regression.

$$\beta \mid y \sim \mathcal{N}(\hat{\beta}, (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\sigma^2)$$

To obtain a credible interval for predictions on a held-out test set $\{\mathbf{X}', \mathbf{y}'\}$, we sample M parameters from this posterior distribution $\{\beta_s\}_{s=1}^{M}$. The sampled parameters are used to create corresponding predictions $\mathbf{y}'_s = \beta_s\mathbf{X}'$, and a credible interval for each prediction can be taken.

To combine predictions (for example, to calculate a total demand estimate from household-level predictions), the total mean prediction is calculated as the sum of the components $\hat{y}_{total} = \sum_i \hat{y}_i$.

To obtain a combined credible interval, take the variance of predictions for each sample $\sigma^2_{y_i}$, then

$$\sigma^2_{y_{total}} = \frac{\sum_{i=1}^{N} \sigma^2_{y_i}}{N}$$

3.5 Analysis Pipeline

The demand data contains 48 observations per day, at half-hourly intervals. Initial testing found that fitting a ridge regression model to each time of day significantly out performs using a single model with time of day as a covariate. A number of days were removed from the dataset, such as Christmas Day, New Years Eve, and other holidays, as is common in energy prediction models.

A simple feature transform of that dataset was performed, to account for non-linear relationships and categorical variables such as the day of the week were one-hot-encoded. A validation set was held out from all analyses, containing the last day of each month.

For each cluster of samples, and for a dataset of total demand, 48 ridge regression models were fit in parallel. This resulted in a penalty parameter λ and a vector of coefficients β for each model. For each model $M = 1000$ values of β were sampled from the posterior $\beta|y$, and a mean estimate for each held-out sample was predicted, along with a 95% credible interval.

4 Results

4.1 Total Demand

The total demand at each of the 16704 time points was calculated. For each of the 48 times of day, a model was fit using 336 samples, and tested on 12 (the held-out last day of each month). Since we are predicting aggregated demand, we are unable to include household level covariates. On this small dataset, our R package is able to fit the models in parallel very fast.

The resulting predictions are shown in Figure 4.

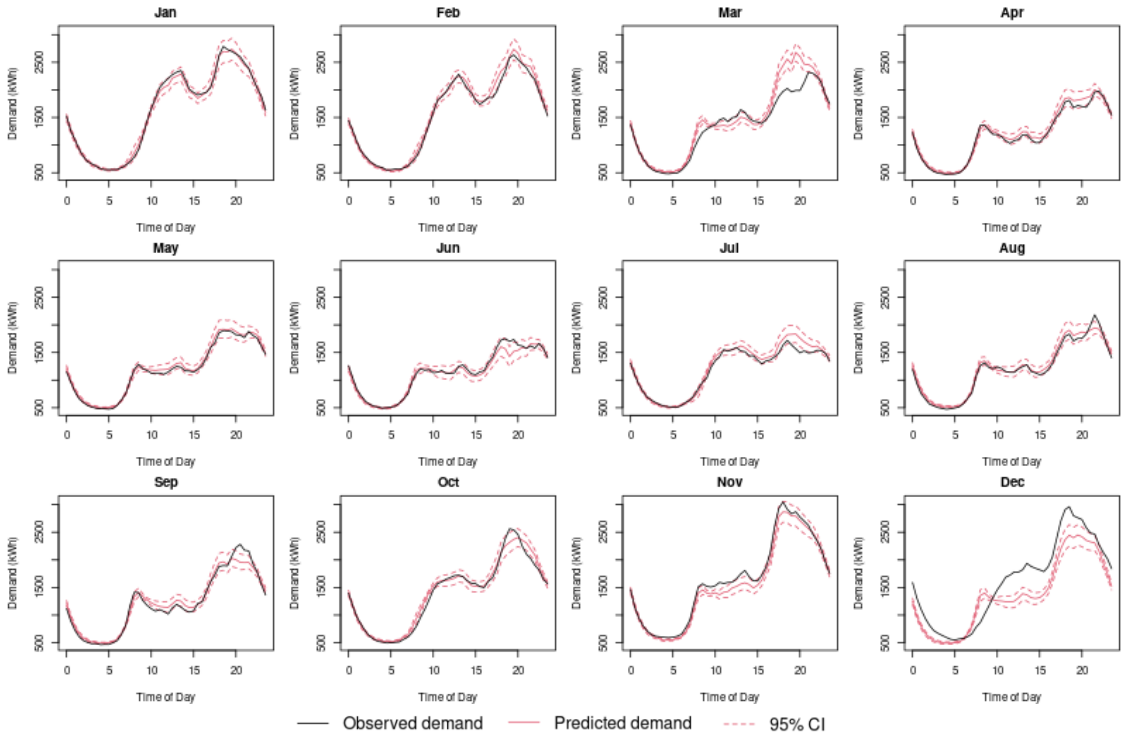

Figure 4: Total demand predictions for the last day of each month.

Since a regression model is fit to each time of day, we have 48 sets of parameters $\{\beta_t, \lambda_t\}$. The mean regression coefficients $\hat{\beta}_t$ are shown in Figure 5. We can see that the effect of the day of the week varies significantly over the time of day, and that temperature is a key predictor.

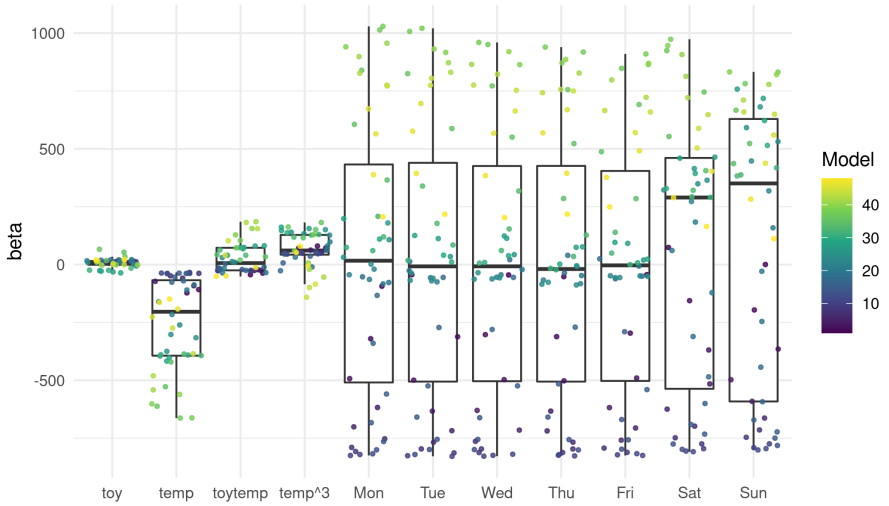

Figure 5: $\hat{\beta}_t$ for each of the 48 total demand models.

We can also visualise the credible intervals for β_t; Figure 6 shows an example of this for model 48 (23:30-00:00).

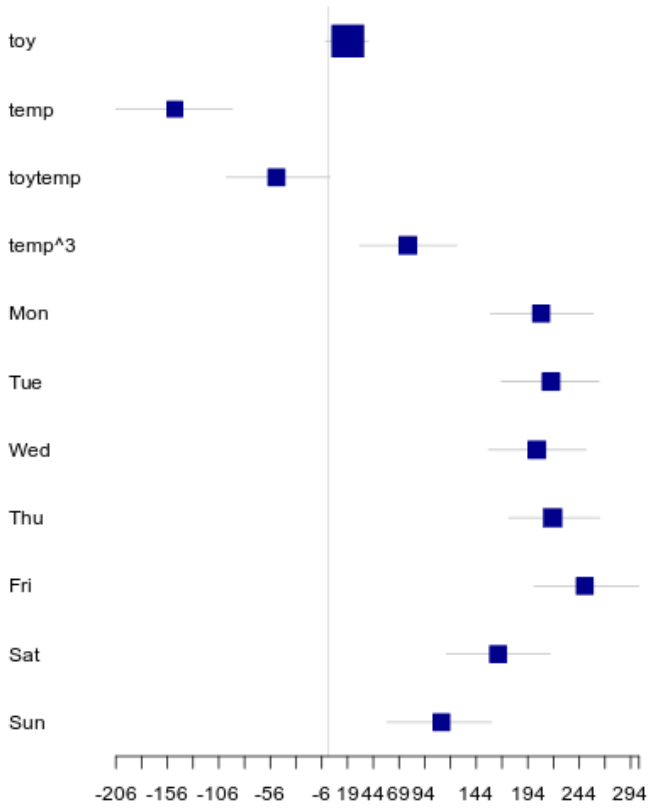

Figure 6: $\hat{\beta}_t$ for each of the 48 total demand models.

The mean absolute error of the predictions is shown in Table 2. The MAE normalised per household is also shown, to give a value comparable to the household-level models.

Model	**MAE**	**MAE/n**
Total Demand	81.7	0.142

Table 2: Mean absolute error of the total demand models on the validation dataset.

4.2 Clusters

We now look at the results of fitting to the 5 clusters. The number of households within each cluster is shown in Table 3.

Model	**Size**
Cluster 1	481
Cluster 2	546
Cluster 3	720
Cluster 4	780
Cluster 5	145

Table 3: The number of households within each cluster.

We see that the clusters are of relatively equal sizes with the exception of cluster 5 which is quite small. Within each cluster we fit 48 separate ridge regression models, one for each of the half hourly intervals. Thus in total we are fitting 5 lots of 48 models or 240 models in total. Each of these includes 36 parameters. We are able to do fit this many parameters as there is an abundance of data available with over 40,000,000 rows. In order to fit these models we wrote C++ code that uses OpenMP to parallelise across the times of day. The predictions on the test set and associated credible intervals are presented in the appendix.

We see in these plots that the model performance is worst in the month of December. This is likely due to our test data being from the end of the month and the last days in December are taken off work for many people, leading to deviations from typical demand. We have removed New Years Eve but it appears that deviations still occur on the 30th December. Overall the plots show that the model is fairly accurate at predicting the demand profiles within each cluster.

4.2.1 Credible intervals

We take a parametric bootstrap of the Bayesian posterior to find confidence intervals for both β and y. We do this by finding the mean and variance for the multivariate normal Bayesian posterior. We then sample estimates of $\hat{\beta}$ from this 1000 times and take the relevant quantiles to attain the credible interval. With these randomly sampled $\hat{\beta}$ parameters we predict the demand and then take quantiles of this demand to find intervals for y for each time of day.

Model	**MAE**
Cluster 1	0.0561
Cluster 2	0.0165
Cluster 3	0.0242
Cluster 4	0.0361
Cluster 5	0.0858
Aggregate	0.0352

Table 4: Mean absolute error of the cluster models on the validation dataset.

We can take a weighted average of the MAEs from each cluster to find the MAE predicting the total demand. We find this as 0.0352. This is around four times better than simply fitting straight to the total demand. Thus there are clear advantages to clustering the data before fitting.

Figure 7 shows the predictions for each of the 5 clusters, with 95% credible intervals, for the January validation data. The average demand is taken per cluster, representing an average household, so that clusters of different sizes are comparable. We can clearly see the different demand profiles of the clusters, with cluster 2 on the left-hand-side of Figure 3 having the lowest demand, and cluster 5 on the right-hand-side of Figure 3 having the highest mean demand. We see that the least accurate fit is within cluster 5, this could be as it is the smallest data set as well as the fact that the data within cluster 5 were quite varied as seen in Figure 3. In all the other clusters estimated demand profile is almost identical to the true demand profile.

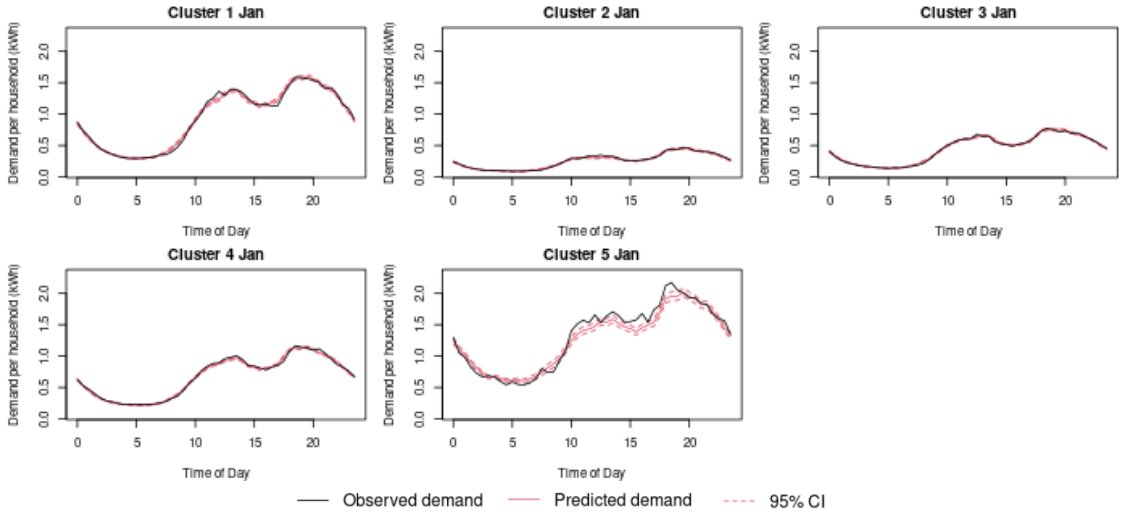

Figure 7: Average demand predictions for January for each of the 5 clusters.

5 Conclusion

We aimed to pioneer a new approach to electricity forecasting, namely clustering by household demand profiles in order to improve model performance. We see that there are clear advantages to fitting with clustered sets of data. Grouping the households by their average daily demand profile is a simple but effective way of clustering the data set that has clear benefits for both fitting speed and accuracy. As there is more consistency within each data set the parameter estimates remain more relevant even under the relatively simple model of a penalised ridge regression.

We find that the performance of the ridge regressions fitted to the clusters significantly outperforms the same model but fitted to aggregate demand. The aggregate demand model achieved an MAE of 0.142 whilst the clustered models achieve an MAE of 0.0352 when the predictions are summed for total demand. Our analysis was somewhat limited by the simplicity of the ridge regression model, yet we have achieved a good gain in accuracy. If this gain could be repeated under more complex models then it could prove to be quite an effective approach.

Though ridge regression is a fairly simple model, it appears that we have captured the complexity of the data by fitting to suitable subsets (customer clusters per time of day). We included a number of predictor variables, including a non-linear feature transform, however these regression coefficients were often close to 0.

In the future it would be interesting to fit more complex models such as GAM's to the clustered data. These could further improve upon the performance we attained here. However fitting multiple GAM models could prove prohibitively computationally expensive. One advantage of a GAM is that rather than fitting at multiple times of day you can include a smoothed time of day covariate.

Overall we conclude that the approach of clustering and splitting models by time of day is a promising technique for demand forecasting yet more work would need to be done using a variety of different model types to properly analyse the effectiveness.

References

[1] C. Capezza, B. Palumbo, Y. Goude, S. N. Wood, and M. Fasiolo, "Additive stacking for disaggregate electricity demand forecasting," 2020.

[2] J.-H. Meier, S. Schneider, C. Le, and I. Schmidt, "Short-term electricity price forecasting: Deep ann vs gam," in *Information and Communication Technologies in Education, Research, and Industrial Applications* (V. Ermolayev, F. Mallet, V. Yakovyna, H. C. Mayr, and A. Spivakovsky, eds.), (Cham), pp. 257–276, Springer International Publishing, 2020.

[3] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[4] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001.

[5] A. Kassambara, *Practical guide to cluster analysis in R: Unsupervised machine learning*, vol. 1. Sthda, 2017.

A Results per cluster

Cluster 1:

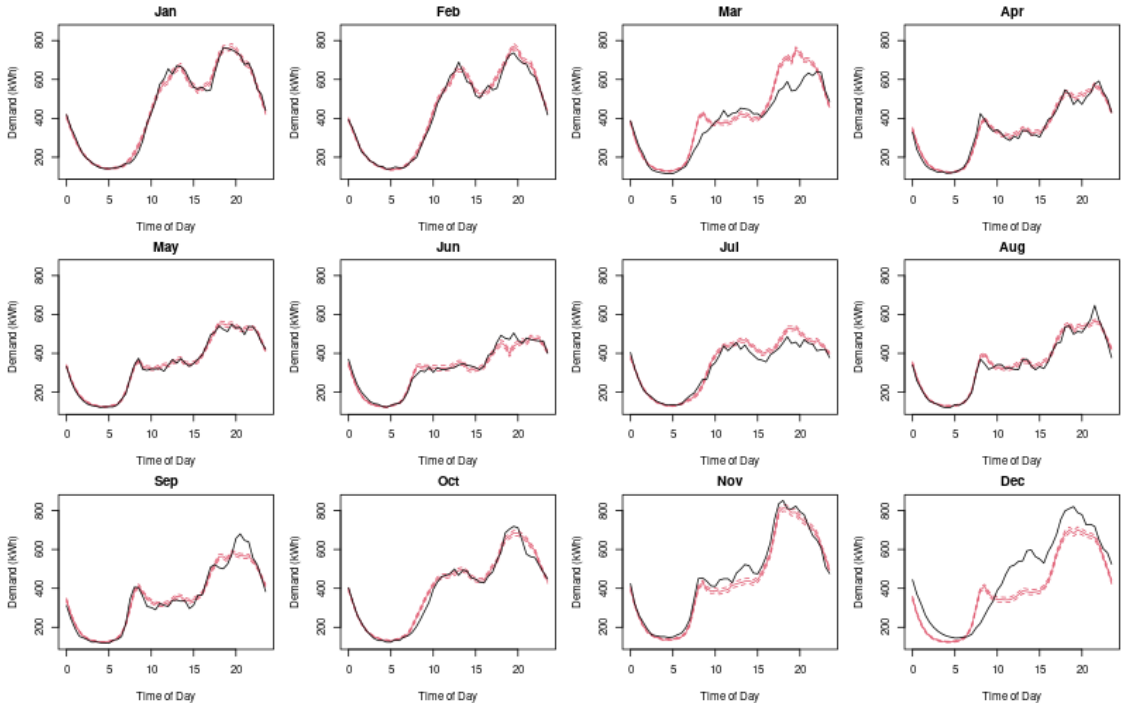

Figure 8: Aggregated demand predictions for cluster 1.

Cluster 2:

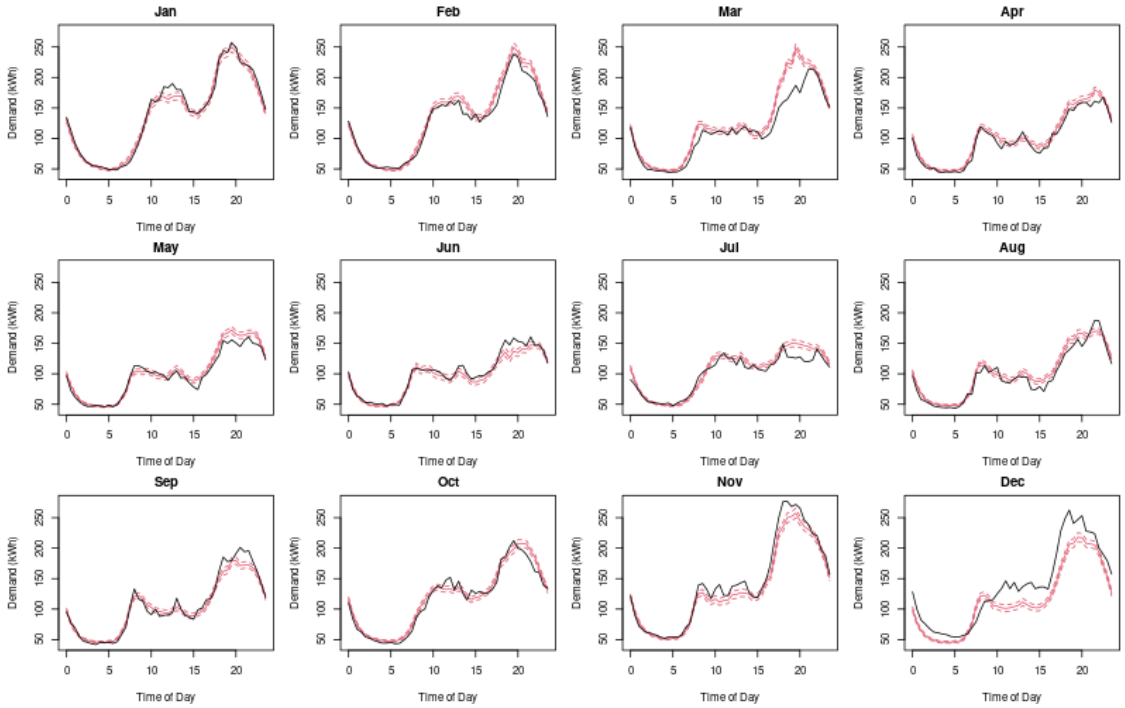

Figure 9: Aggregated demand predictions for cluster 2.

Cluster 3:

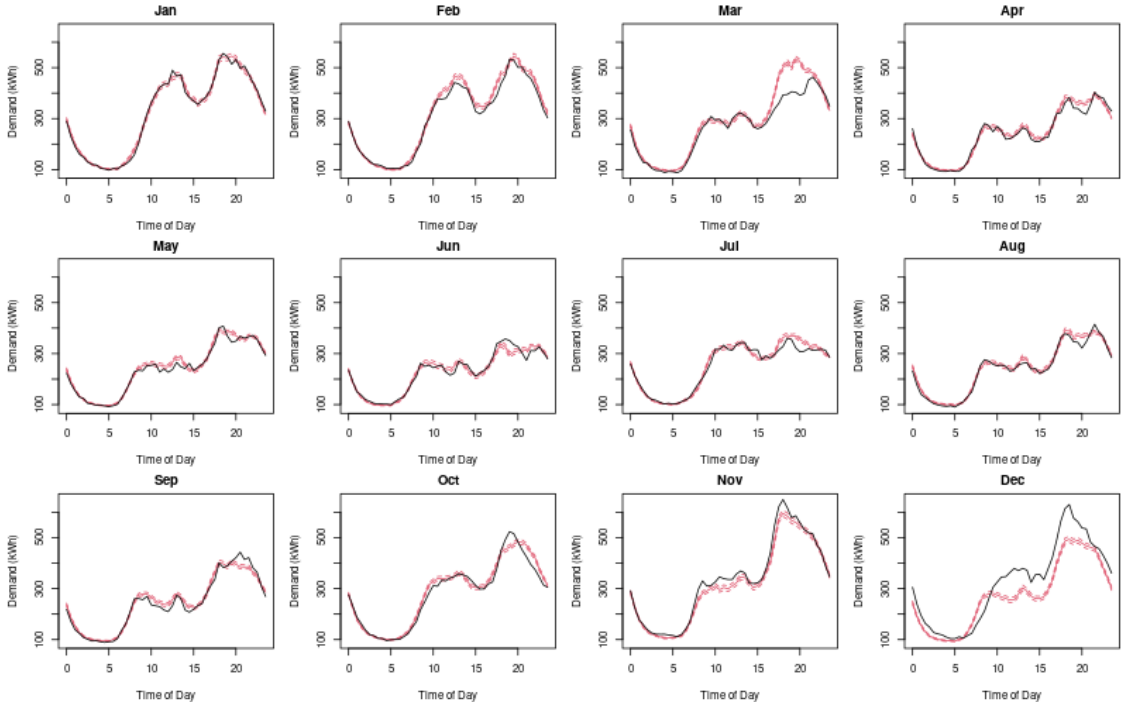

Figure 10: Aggregated demand predictions for cluster 3.

Cluster 4:

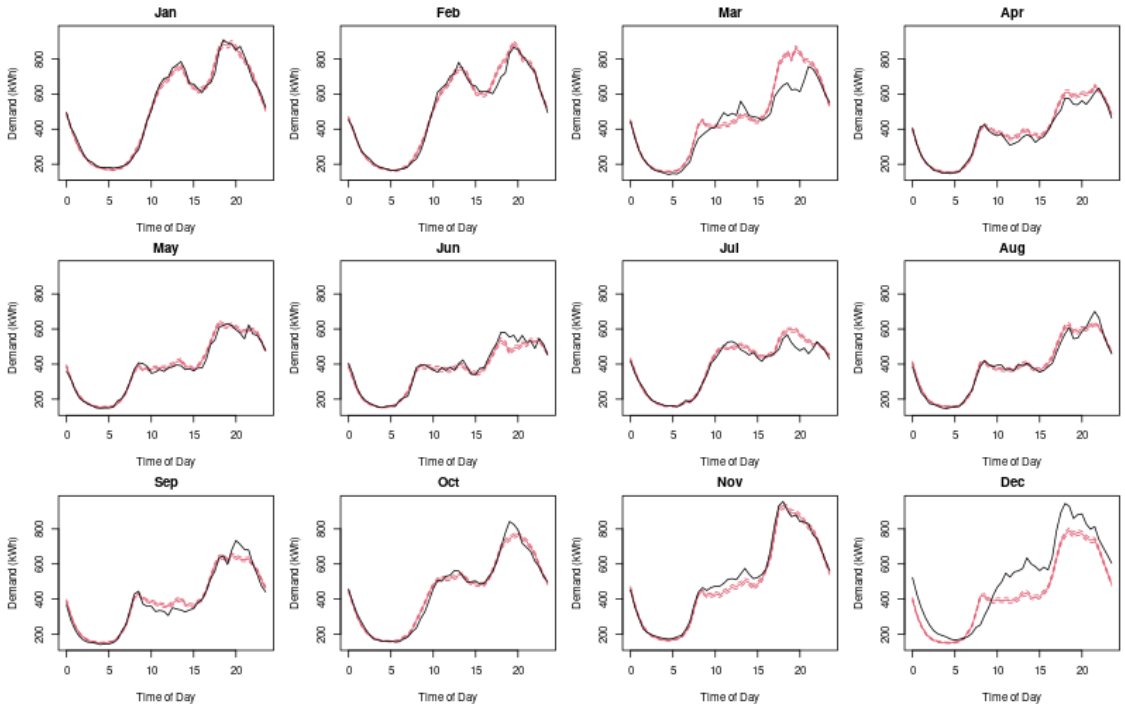

Figure 11: Aggregated demand predictions for cluster 4.

Cluster 5:

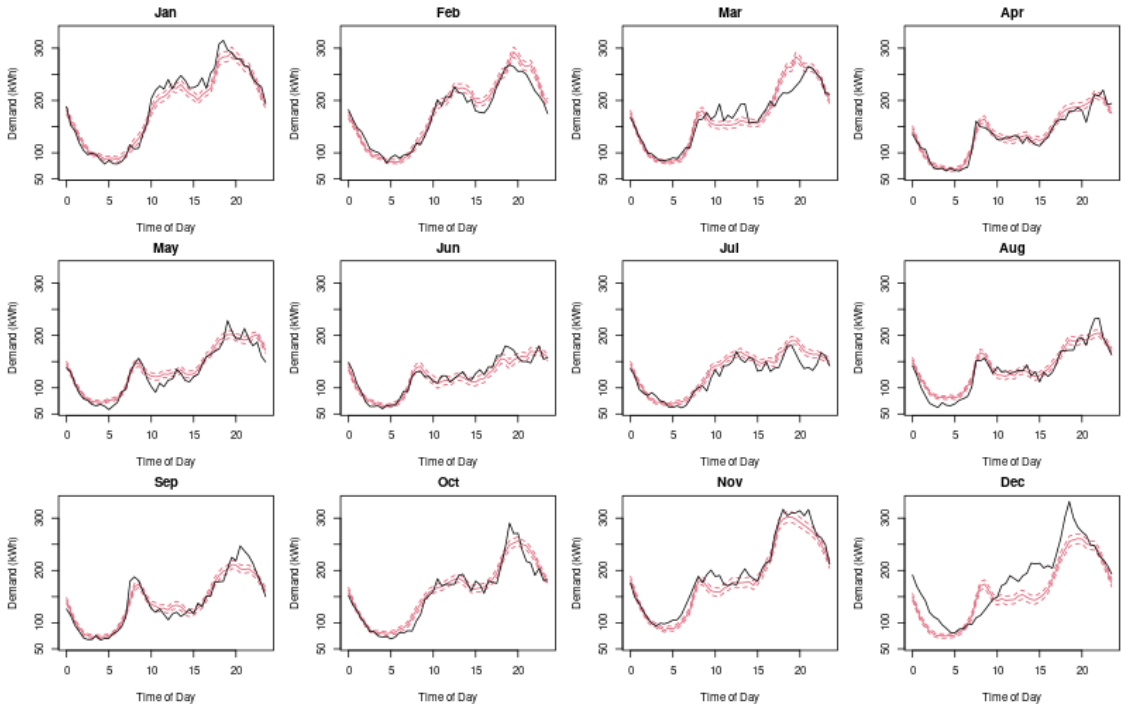

Figure 12: Aggregated demand predictions for cluster 5.

B Example cluster model betas

As in Figure 6, we can visualise the credible intervals for the regression parameter β, for any of our regression models. For the cluster models we have many more predictor variables than the total demand model. A forest plot showing these intervals for an example model (Cluster 5, at 8pm) is shown below:

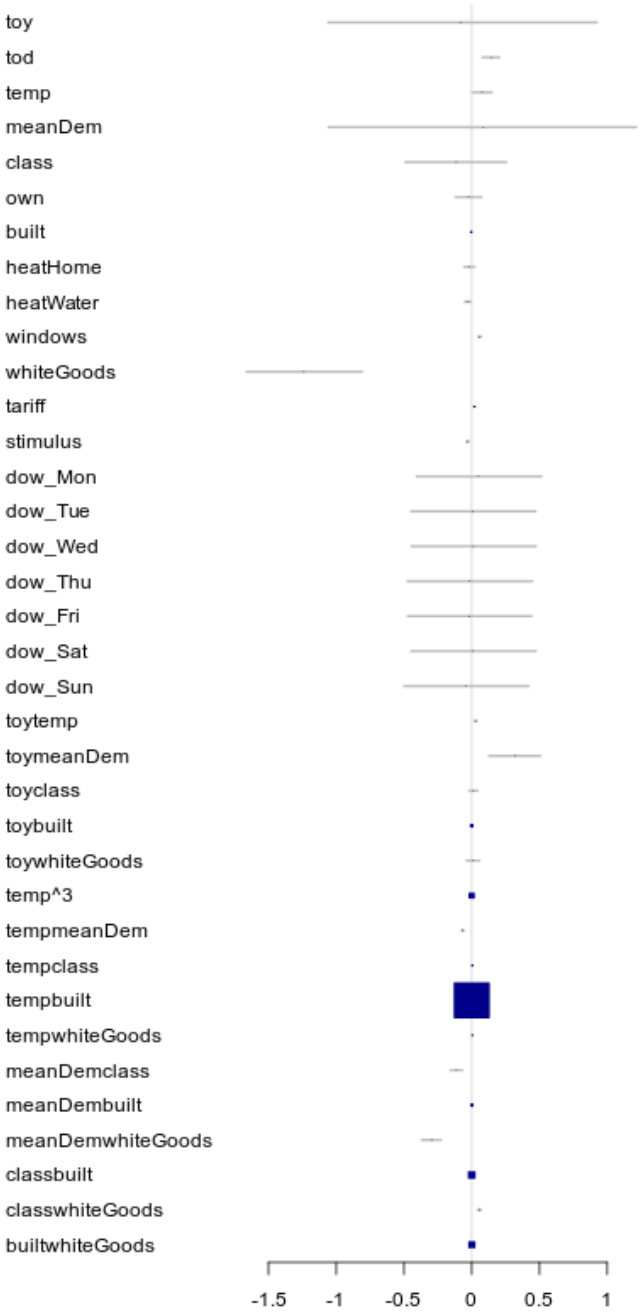

Figure 13: Example fitted regression parameter β for a household-level cluster model.

Many of the intervals cross 0, indicating they are not significantly predictive, with the notable exception of white goods. In many of the other models, the household mean demand had the largest regression coefficient.