Supplementary materials for the main submission entitled - Clustering time series based on probability distributions across temporal granularities

Contents

Prototype selection method

S1. Robust scaling is applied to each customer.

S2. 50^{th} percentile for each category for each granularity is obtained for each customers. So we have a data structure with 356 rows and $(24 + 12 + 2)$ variables corresponding to 50^{th} percentile for each hour-of-day, month-of-year and weekend-weekday.

S3. Apply principal components and restrict the results down to the first six principal components (which makes up approximately 85% of the variance explained in the data) to use with the grand tour.

S4. Run t-SNE using the default arguments on the complete data (sets the perplexity to equal 30 and performs random initialisation). We then create a linked tour with t-SNE layout with liminal as shown in Figure 4.

S5. We inspect of the subspace generated by the set of low-dimensional projections in tour by looking for a simplex shape while the visualization moves from one basis to another. When we brush the corners of the simplex, we find they fall on the edge of the t-SNE point cloud. Hall, Marron, and Neeman (2005) have shown that in the extreme case of high-dimension, low-sample size data, observations are on the vertices of a simplex.

This is because in high-dimensional data analysis the curse of dimensionality reasons that points tend to be far away from the center of the distribution and on the edge of high-dimensional space. Contrary to this, is that projected data tends to clump at the center.

S6. These points should ideally correspond to different behavior with respect to all the variables considered while running PCA.

Wang, Earo, Dianne Cook, and Rob J Hyndman. 2020. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational & Graphical Statistics* 29 (3): 466–78. https://doi.org/10.1080/10618600.2019.1695624.

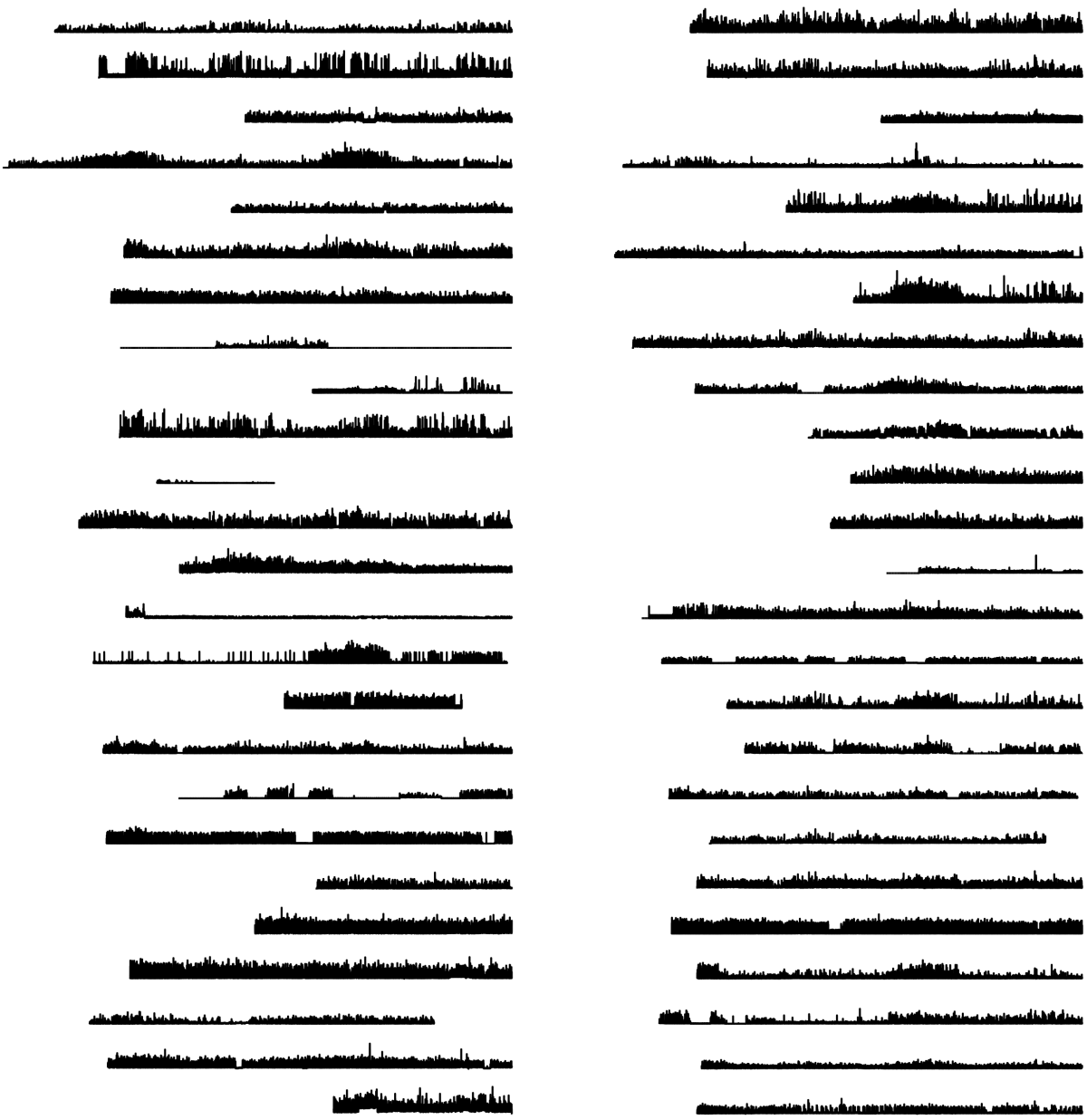

Figure 1: The raw half-hourly energy usage for 50 sampled households is plotted along the y-axis versus time in a linear scale. Each of these series is associated with a single customer. It looks like there is a lot of missing values and unequal length of time series along with asynchronous periods for which data is observed. No insightful behavioral pattern could be discerned from this view other than when the customer is not at home.

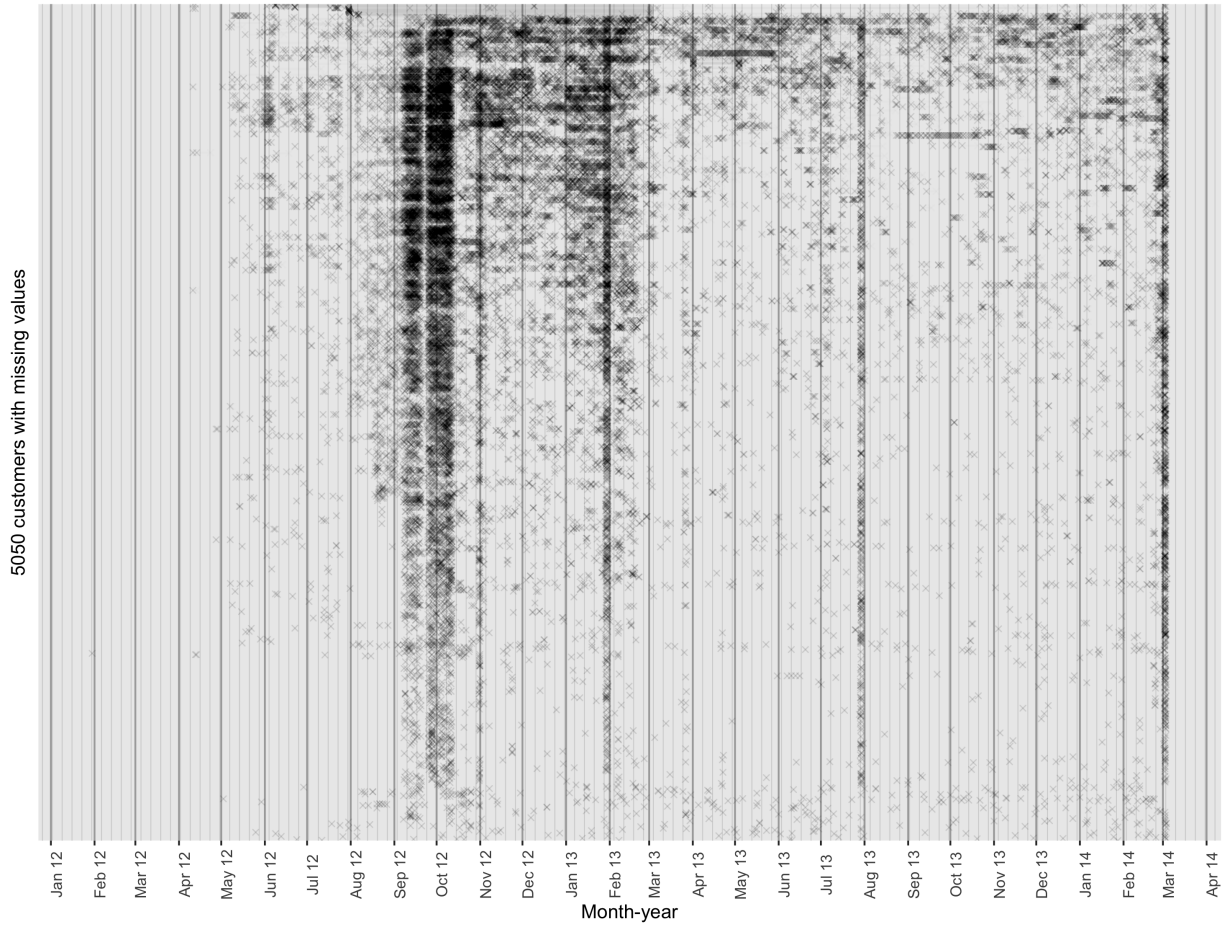

Figure 2: Investigating the temporal location of missing values for customers who have implicit missing values. There are 13,735 customers in the data set, with 8,685 having no missing values and the remaining 5,050 having at least one missing value. Each cross represents a missed observation in time, while the line connecting two dots represents continuous missingness over time. Missing values occur at random times and do not appear to follow a pattern, although there is a higher concentration of missing values in September and October 2012 for the majority of customers. This plot is inspired by Wang, Cook, and Hyndman (2020).

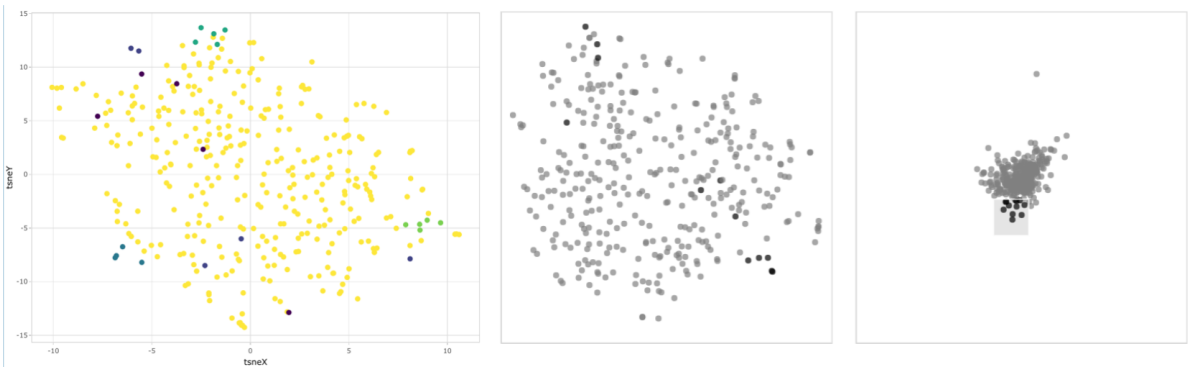

Figure 3: Instance selection using tours and projecting the points in a lower dimensional tsne cloud.

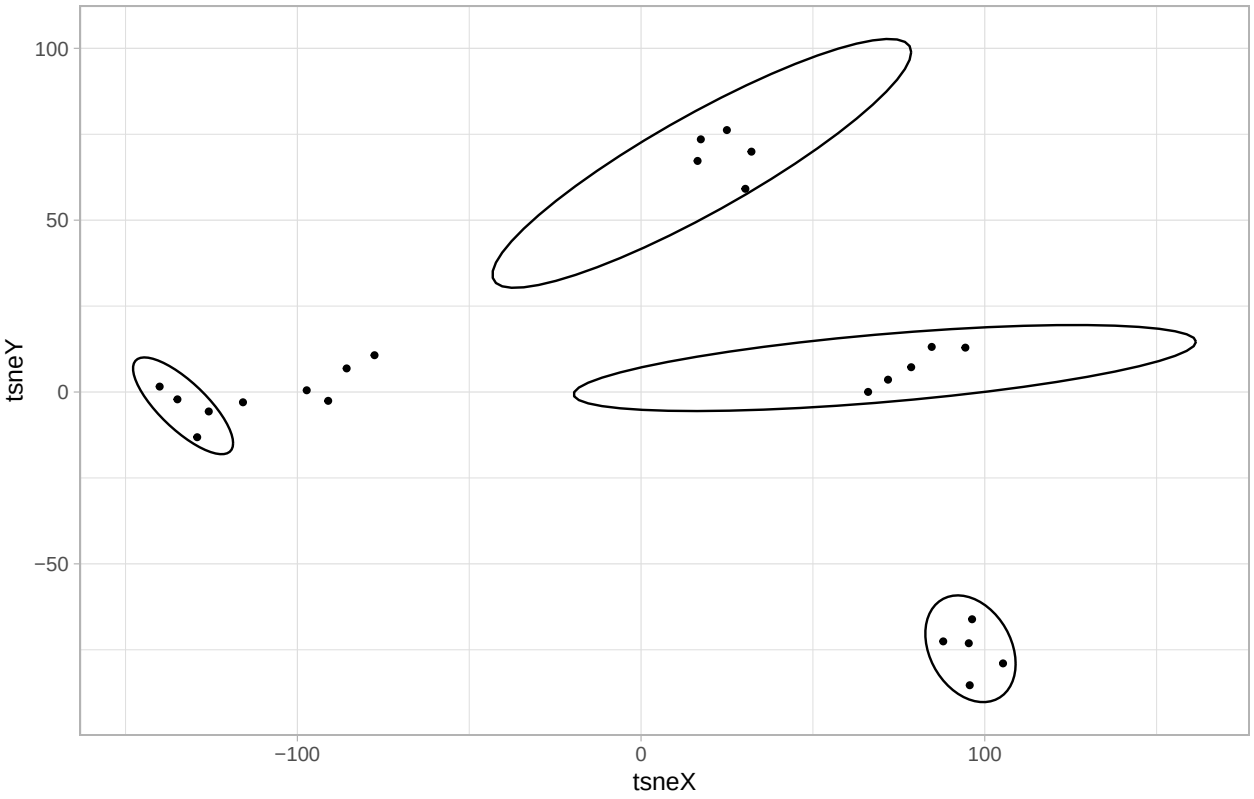

Figure 4: t-SNE summary of selected 24 prototype customers. Each ellipse corresponds to a group after clustering using our methodology. Few customers are not enclosed inside an ellipse as they are from mixed groups and there are too few points to calculate an ellipse around them. It is important to note that the selection of prototypes is based on only 50 percentiles, whereas, the clustering is based on all the deciles.

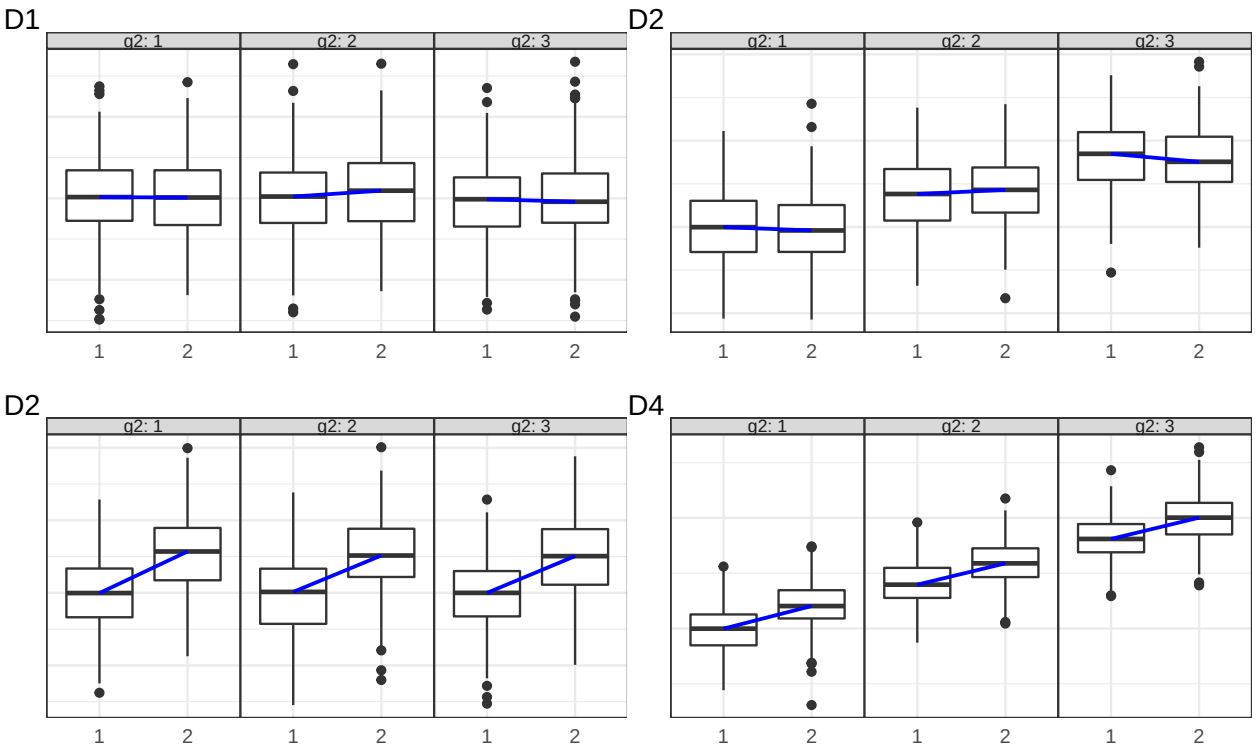

Figure 5: The distribution of simulated variable across $g1$ conditional on $g2$ is shown through boxplots for 4 designs to extend the proposed validation designs when two granularities of interest interact. D1 has no change in distributions across different categories of $g1$ or $g2$, while D2 and D3 change across only $g1$ and $g2$ respectively. D4 changes across categories of both $g1$ and $g2$.

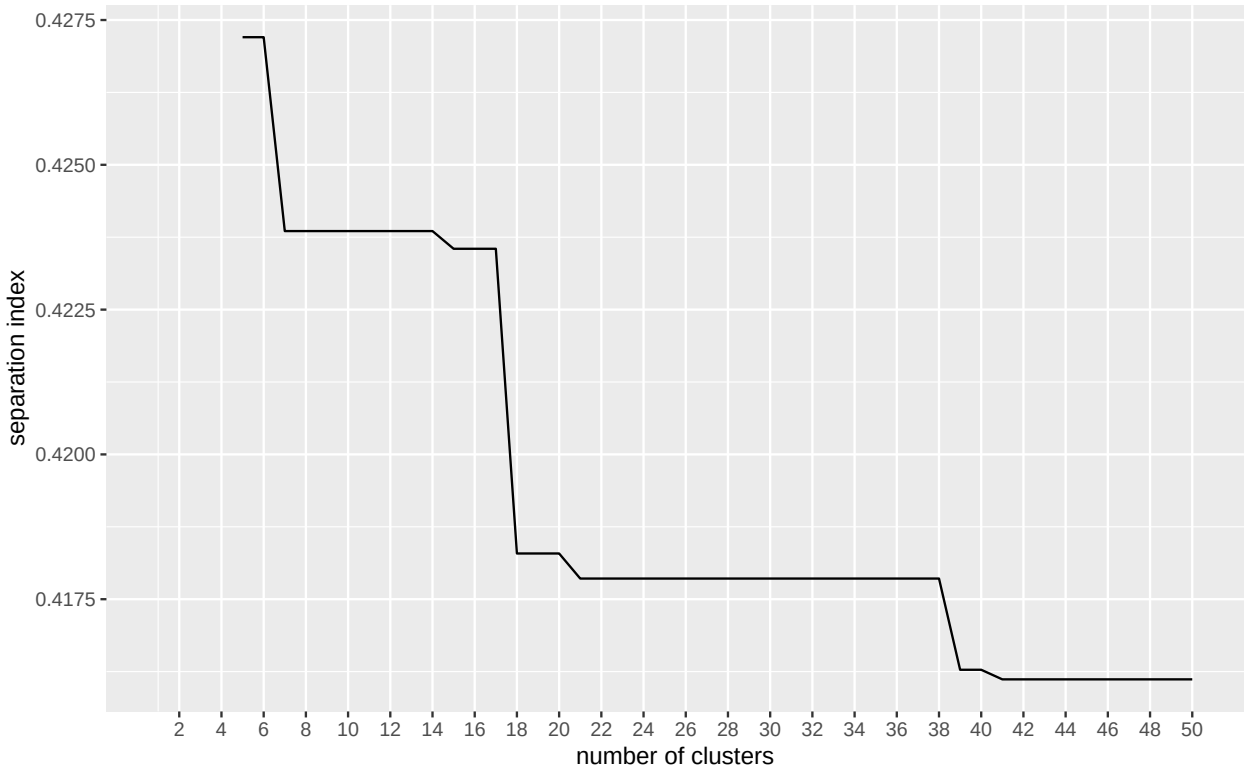

Figure 6: Cluster separation for the 353 customers across the number of clusters is shown. When the cluster size changes from 17 to 18, the separation index drops sharply and then flattens out, resulting in the appearance of the elbow. Hence, when grouping the 353 customers, the number of clusters is taken to be 17.

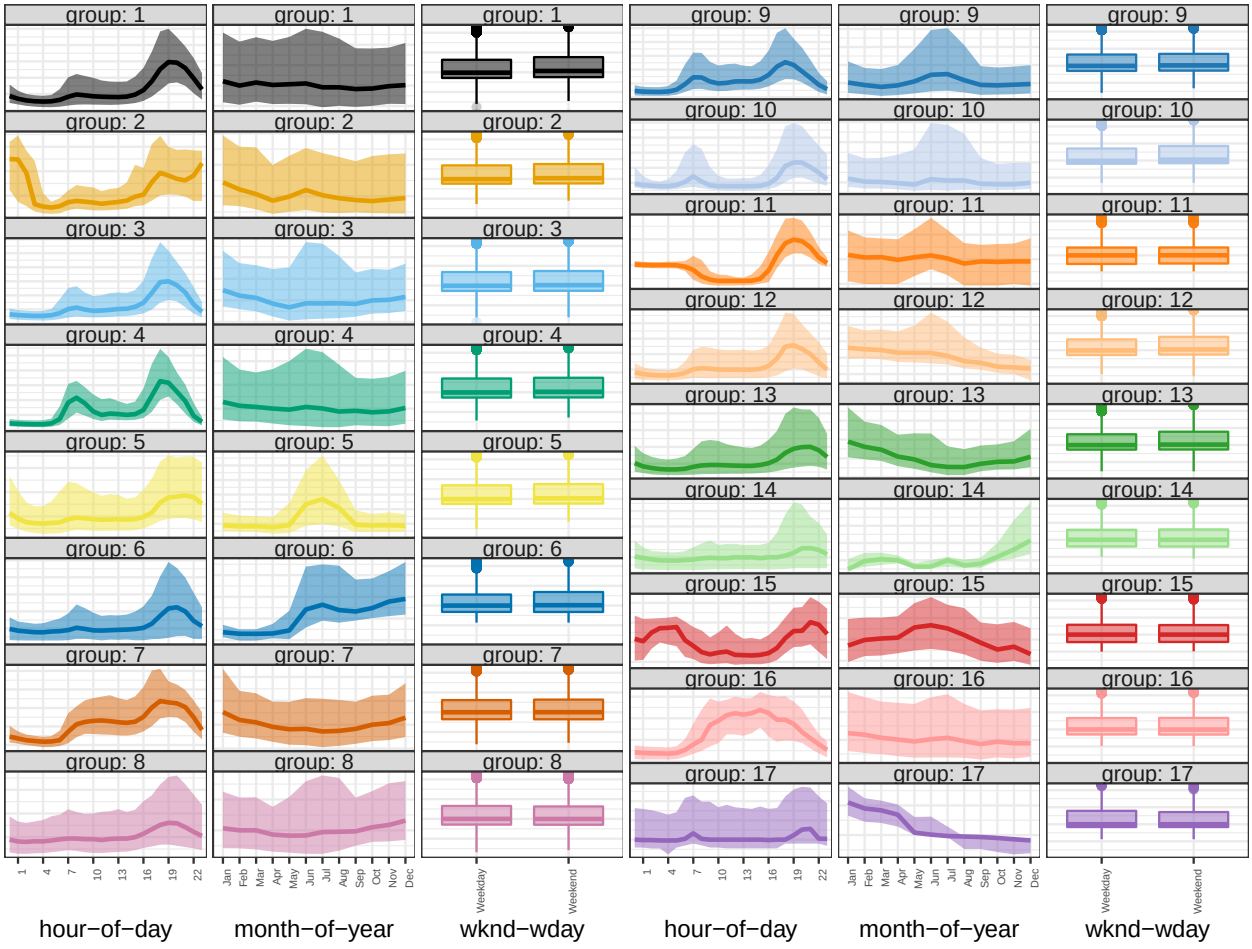

Figure 7: The distribution of electricity demand for the clusters across hod, moy and wkndwday for the 17 groups from 353 customers. Wknd-wday variations across groups are not distinguishable, but ideally each group should have an unique combination of hod and moy.