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

Contents

0.1 Plot displaying missing observations 1
0.2 Clustering on 350 customers 6

0.1 Plot displaying missing observations

The data set contains 13,735 customers, of which 8,685 have no missing values and the remaining $5,050 customers have at least one missing value. We wanted to see if there was any structure to these missing observations, so we plotted them over time in Figure **??**. Each row represents a customer, with the blank indicating non-missing values and the shapes indicating missing values. Figure 9 of Wang, Cook, and Hyndman (2020) inspired this plot. For the most part, it appears that data for September and October 2012 are missing. Otherwise, missingness can occur at any time and there is no pattern.

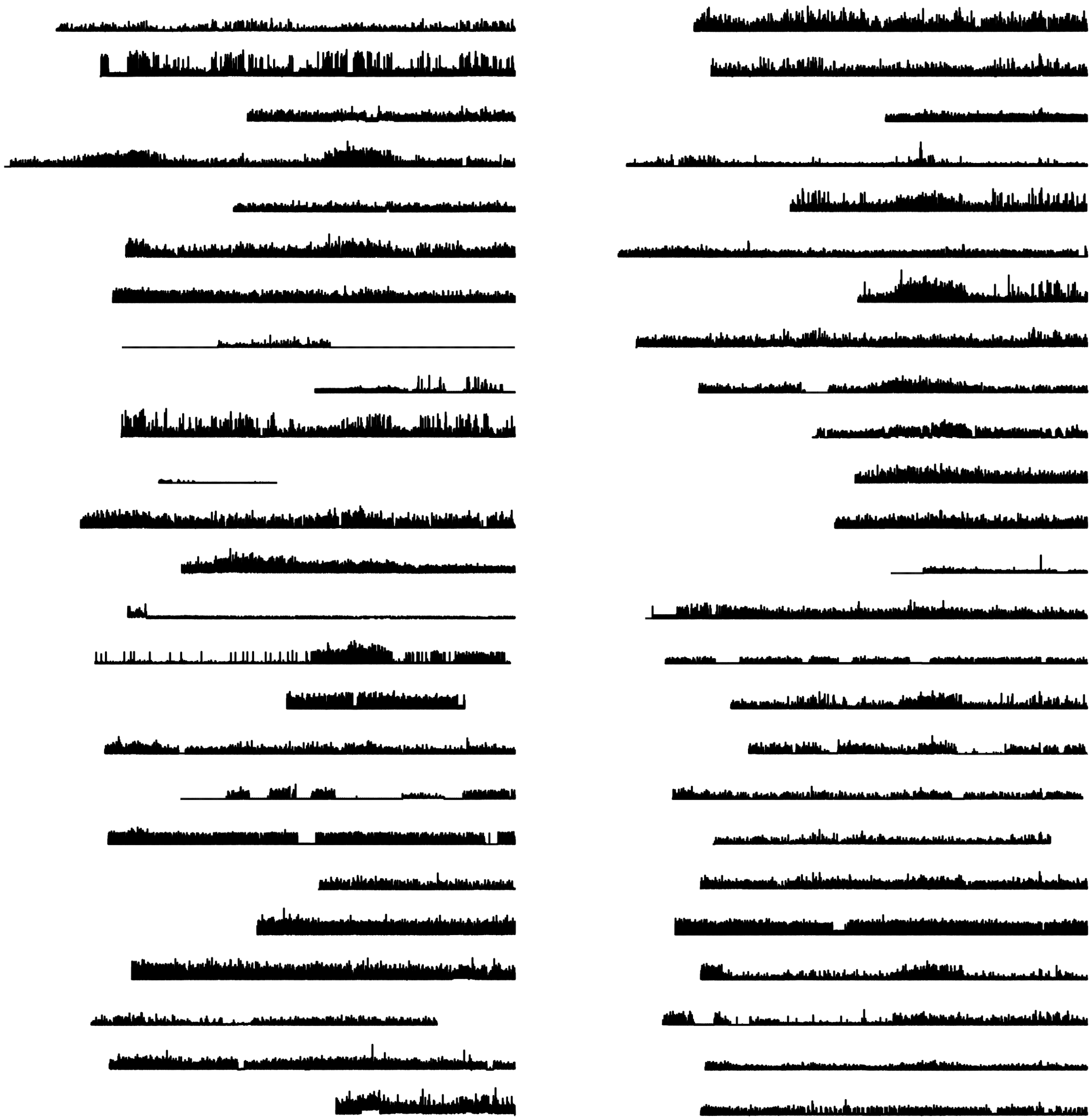
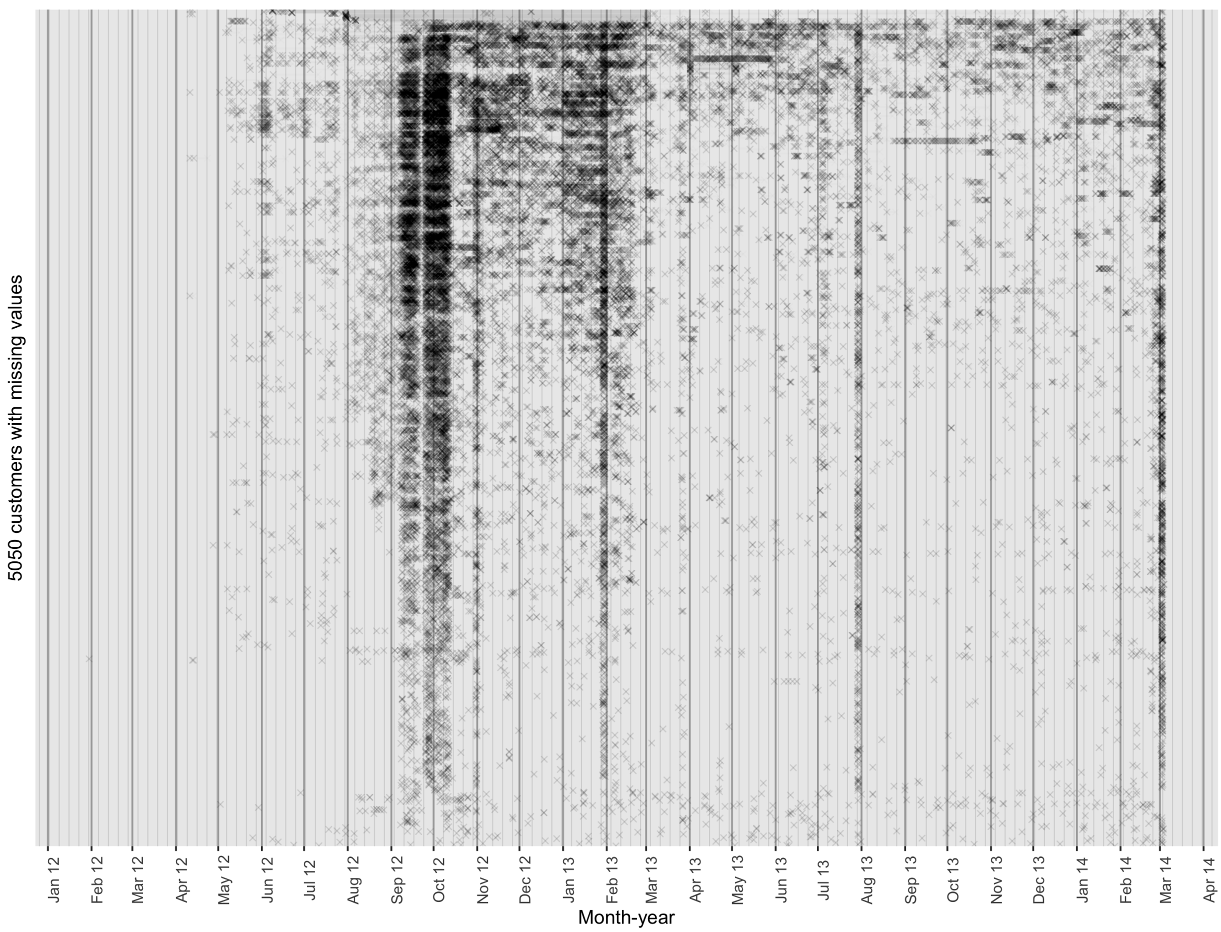


Figure 1: The raw half-hourly energy usage for 50 sampled households is plotted along the y-axis versus time in a linear scale. Each of these series is associated with a single customer. It looks like there is a lot of missing values and unequal length of time series along with asynchronous periods for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



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.

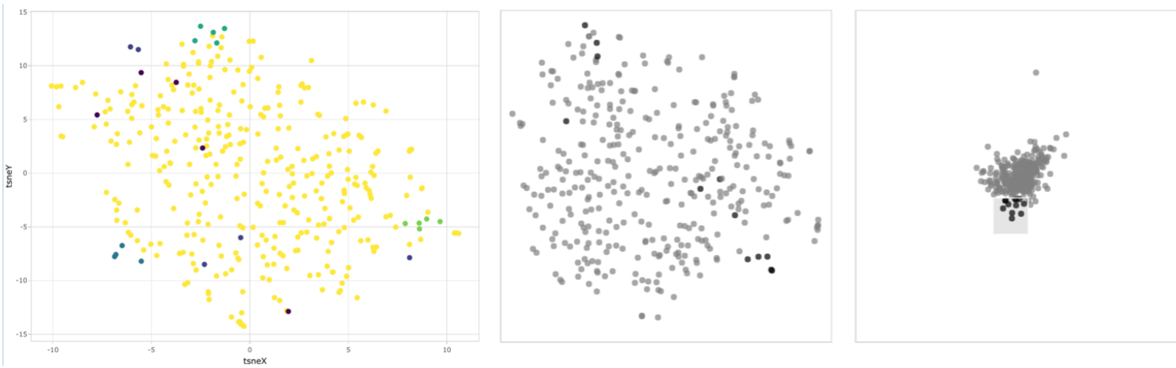


Figure 2: Instance selection using tours and projecting the points in a lower dimensional tsne cloud.

0.1.1 Interaction of granularities

Consider a case in which there are only two interacting granularities of interest, $g1$ and $g2$. In contrast to the previous situation, when we could study distributions across $n_{g_1} + n_{g_2} = 5$ separate categories, with interaction, we must evaluate the distribution of the $n_{g_1} * n_{g_2} = 6$ combination of categories. Consider the 4 designs in Figure 3, where various distributions are assumed for different combinations of categories, resulting in different designs. Design $D1$ exhibits no change in distributions across $g1$ or $g2$, whereas Designs $D2$ and $D3$ alter across only $g1$ and $g2$, respectively. D4 varies across both $g1$ and $g2$ categories. D3 and D4 appear similar based on their relative differences across consecutive categories, but $D4$ also changes across facets, unlike $D3$, which has all facets look the same.

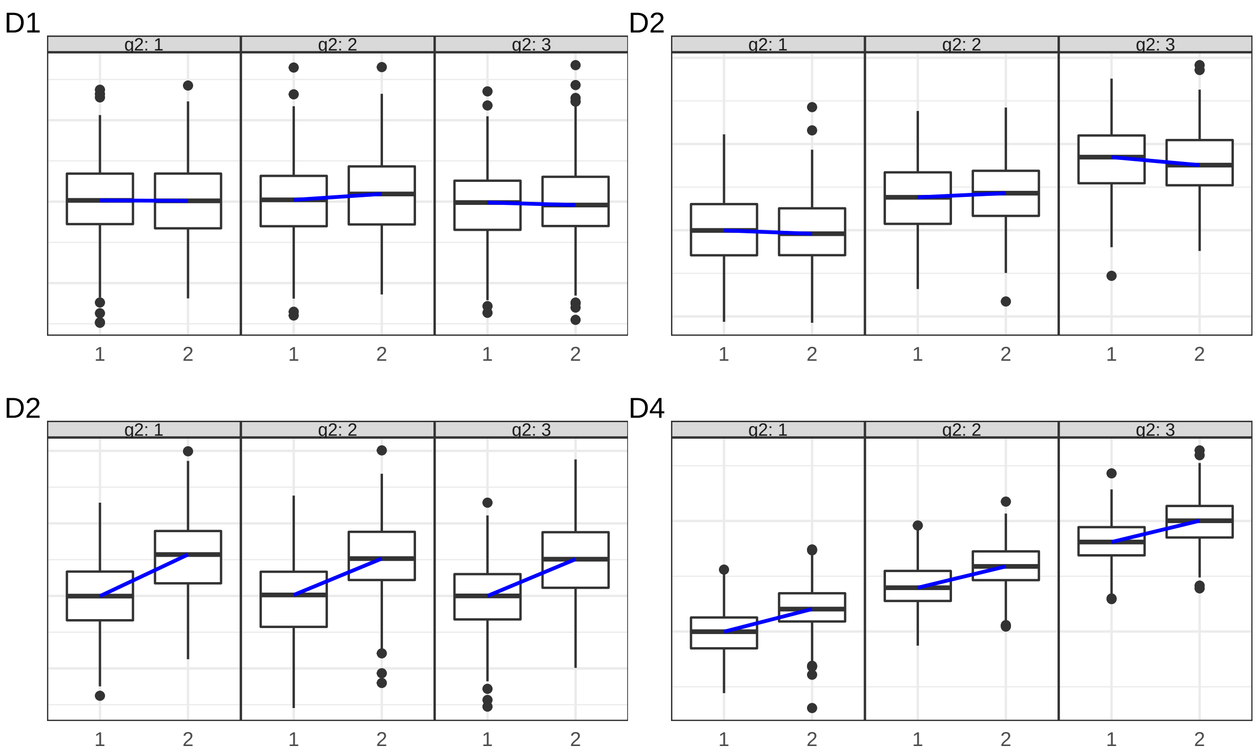
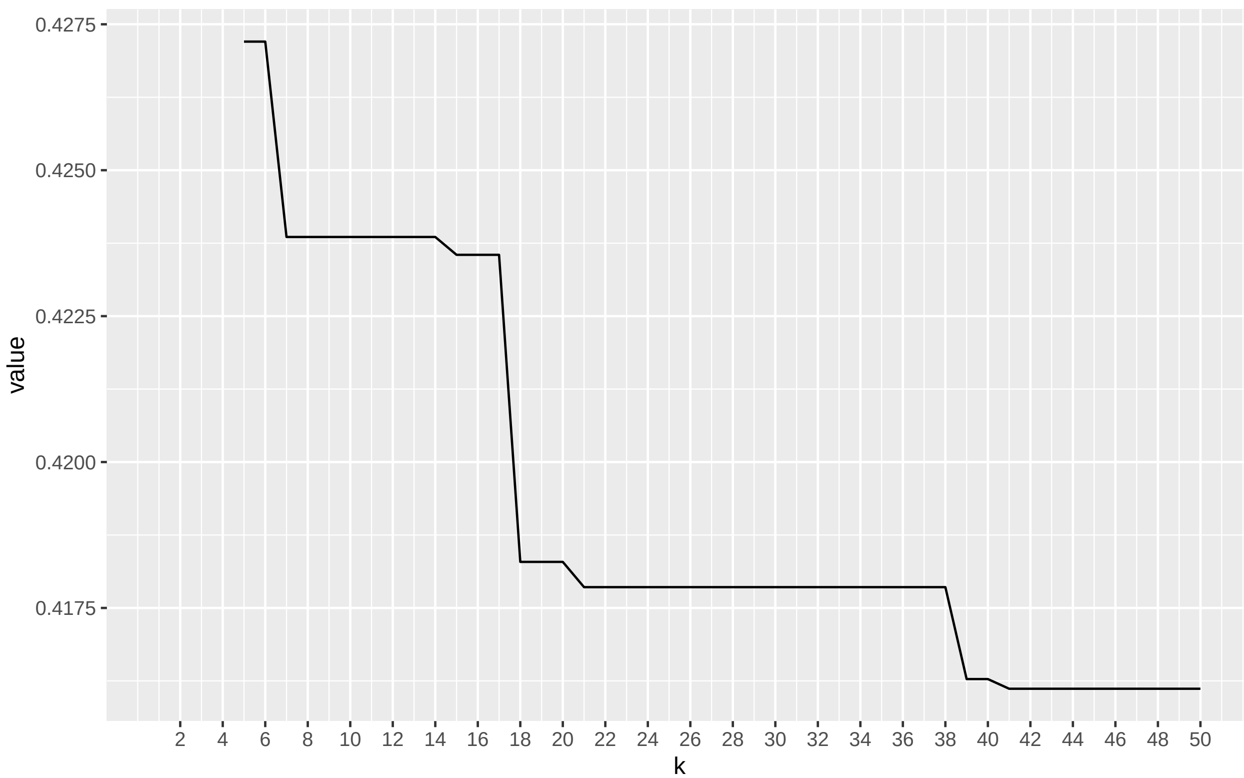


Figure 3: Distribution of the simulated variable across $g1$ conditional on $g2$ is shown through boxplots for 4 designs. D1 has no change in distributions across different categories of $g1$ or $g2$, while D2 and D3 change across only $g1$ and $g2$ respectively. D4 changes across categories of both $g1$ and $g2$.

0.2 Clustering on 350 customers



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 and Graphical Statistics*. https://doi.org/10.1080/10618600.2019.1695624.

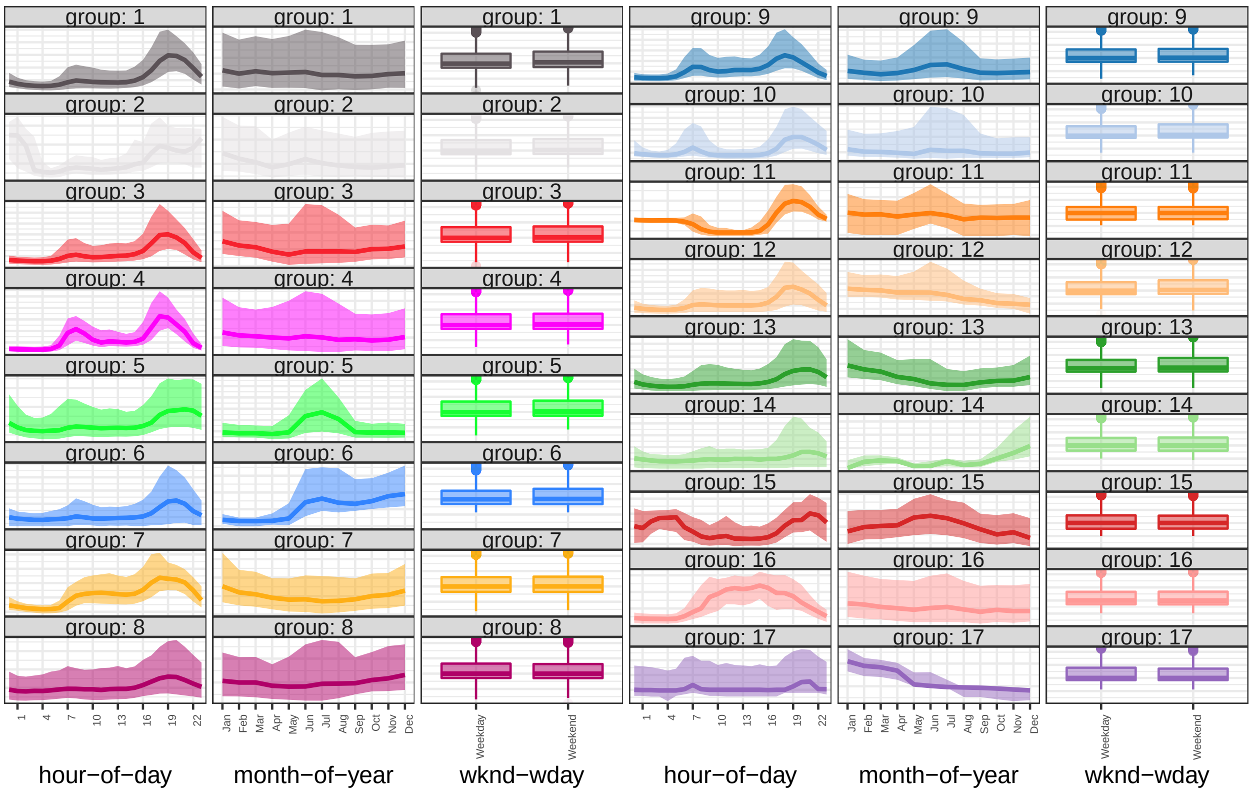


Figure 4: The distribution of electricity demand for the clusters across hod (a), moy (b) and wkndwday (c). It seems like group 2 and 5 have a hod pattern across its members, while group 1, 3, 5 have a moy pattern. Wknd-wday variations across groups are not distinguishable, indicating that it is not a critical variable for clustering. It is helpful to compare the summarised distributions of groups to that of individuals to confirm that the most of individuals in the group have the same characterisation.