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

Contents

Prototype selection

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 R package liminal.

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.

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

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 4, 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.

Clustering all 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 & Graphical Statistics* 29 (3): 466–78. https://doi.org/10.1080/10618600.2019.1695624.

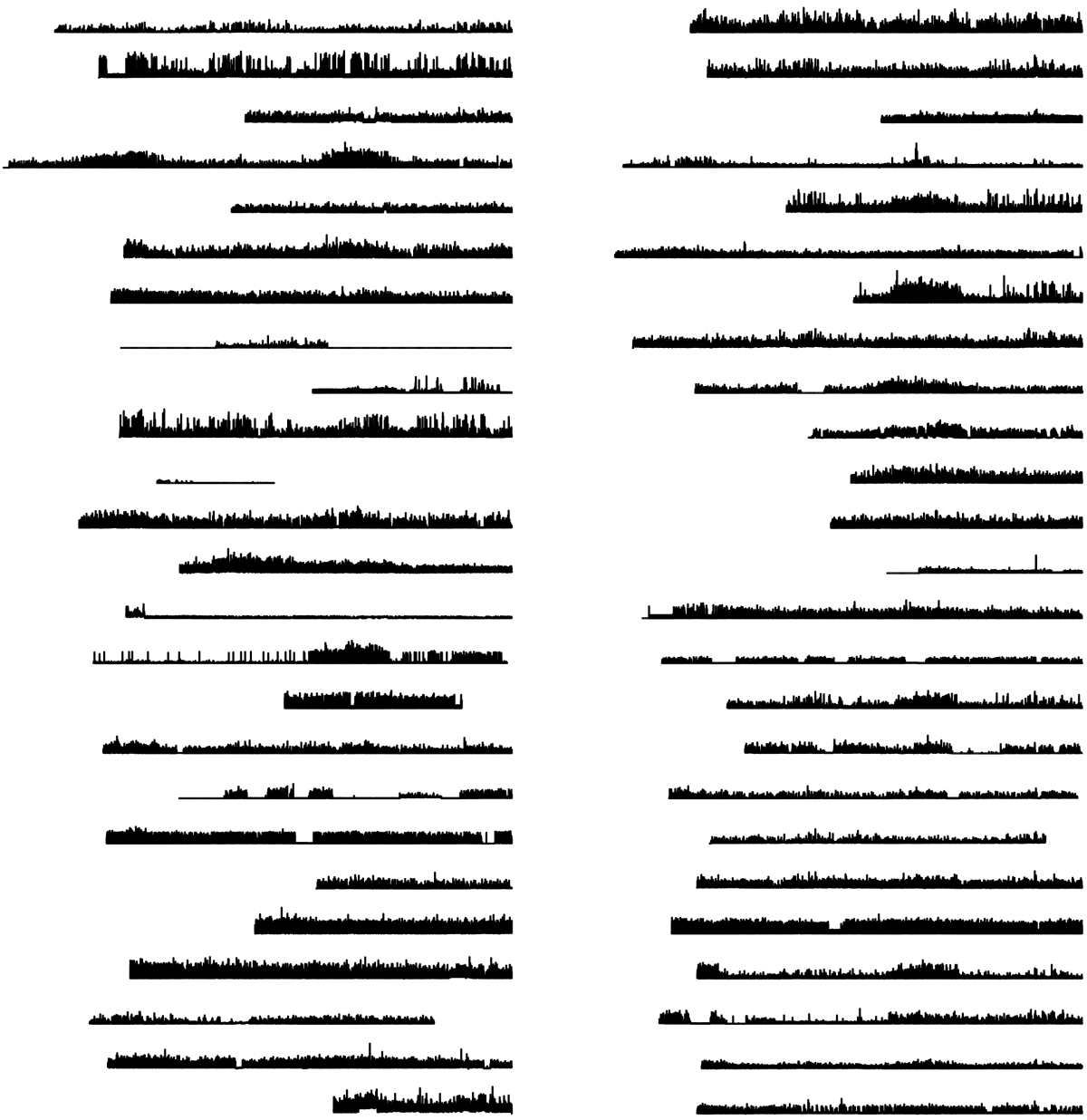

Figure 1: The raw half-hourly energy usage for 50 sampled households is plotted along the y-axis versus time in a linear scale. Each of these series is associated with a single customer. It looks like there is a lot of missing values and unequal length of time series along with asynchronous periods for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

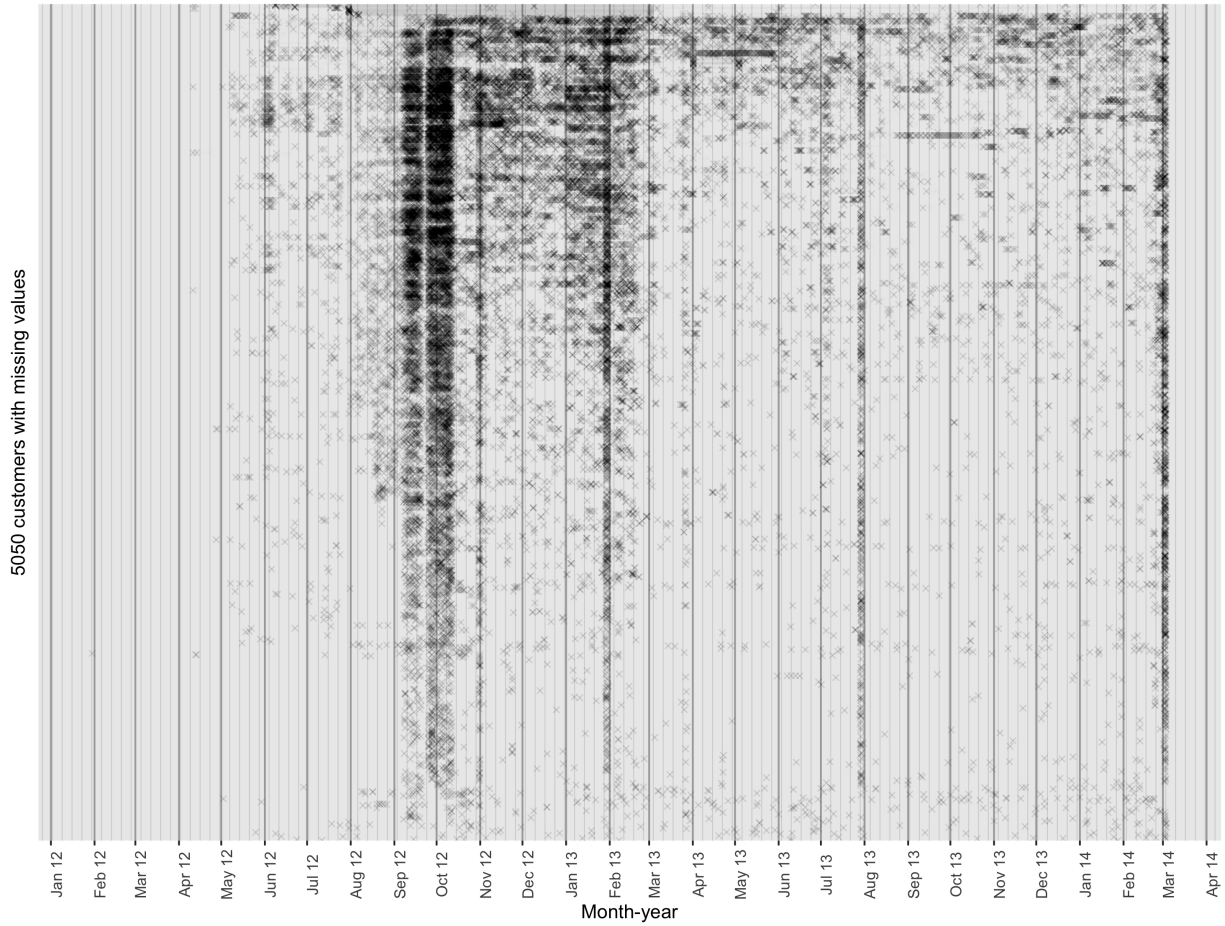

Figure 2: Investigating the temporal location of missing values for customers who have implicit missing values. There are 13,735 customers in the data set, with 8,685 having no missing values and the remaining 5,050 having at least one missing value. Each cross represents a missed observation in time, 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).

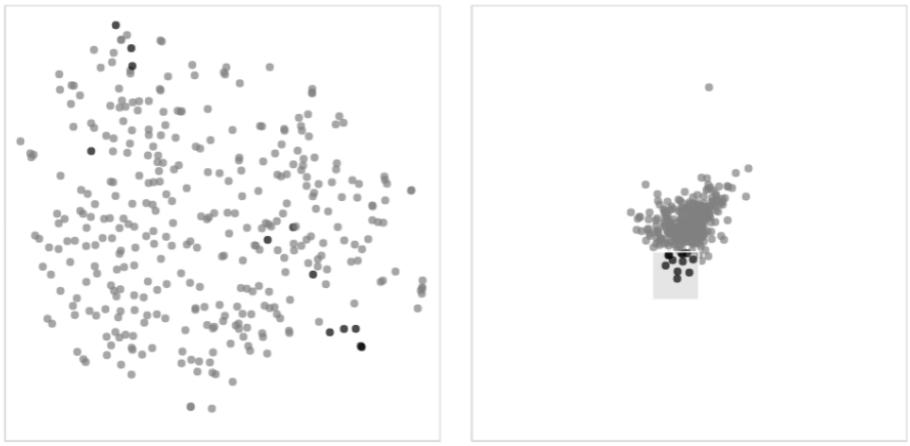

Figure 3: One instance of brushing in tours (right) and projecting the points in a lower dimensional tsne cloud (left).

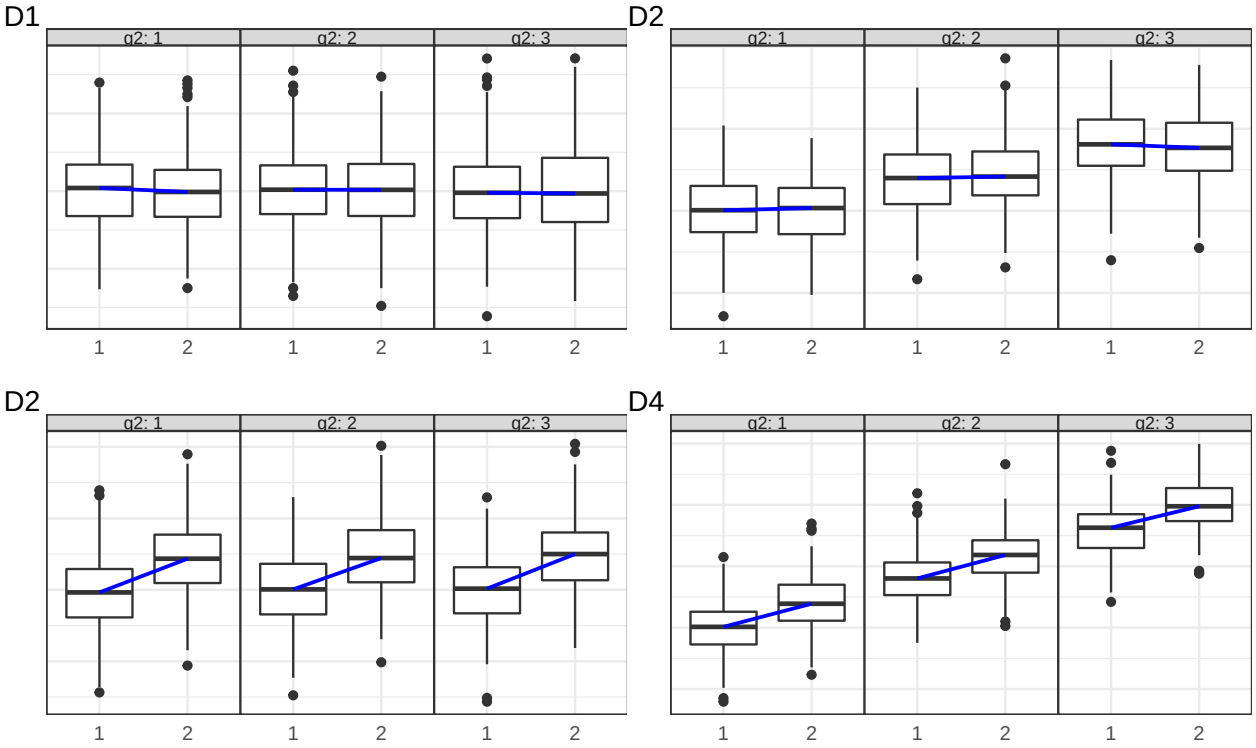

Figure 4: The distribution of simulated variable across $g1$ conditional on $g2$ is shown through boxplots for 4 designs to extend the proposed validation designs when two granularities of interest interact. D1 has no change in distributions across different categories of $g1$ or $g2$, while D2 and D3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

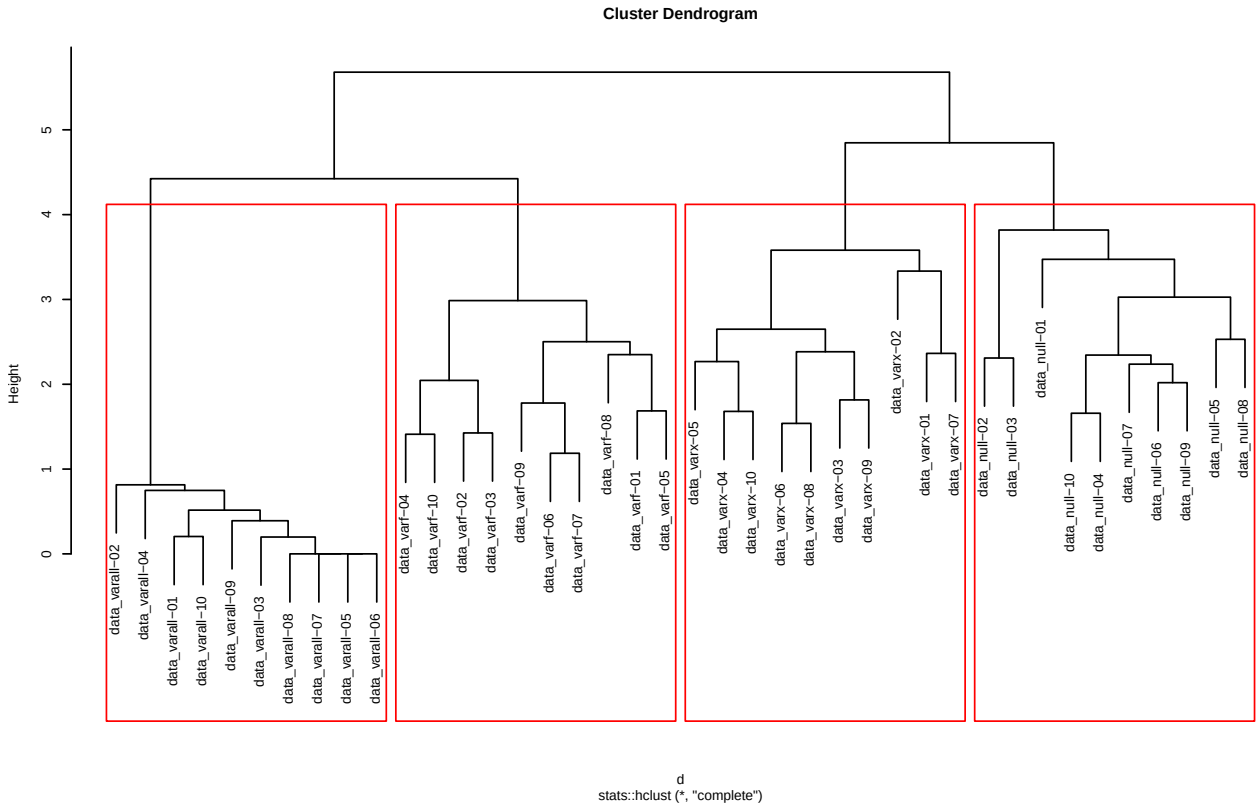

Figure 5: A dendogram showing the branching of 40 series with each of the 4 designs repeated 10 times.

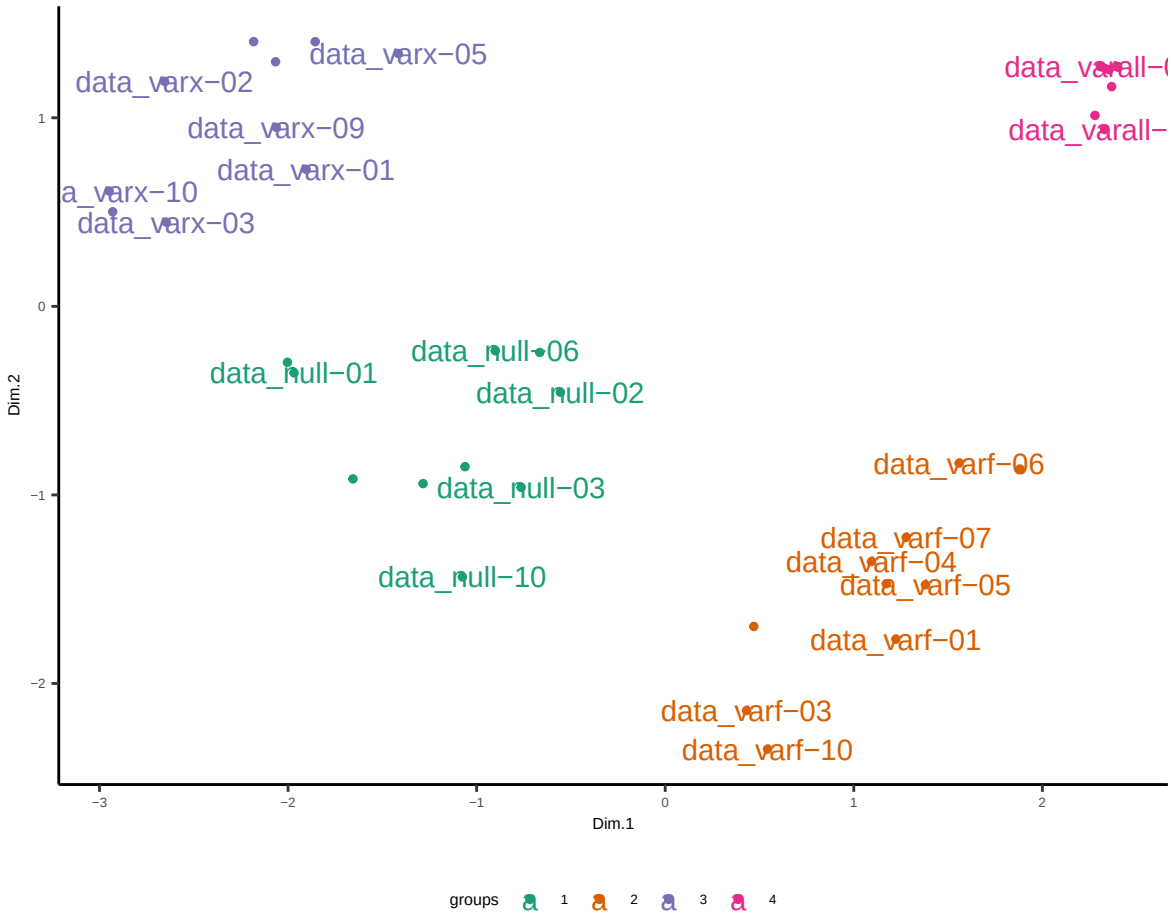

Figure 6: MDS summary plot of 40 series with each of the 4 designs repeated 10 times. It can be observed that all the series belonging to same design lie closer in this plot and all the ones in different designs are placed further away.

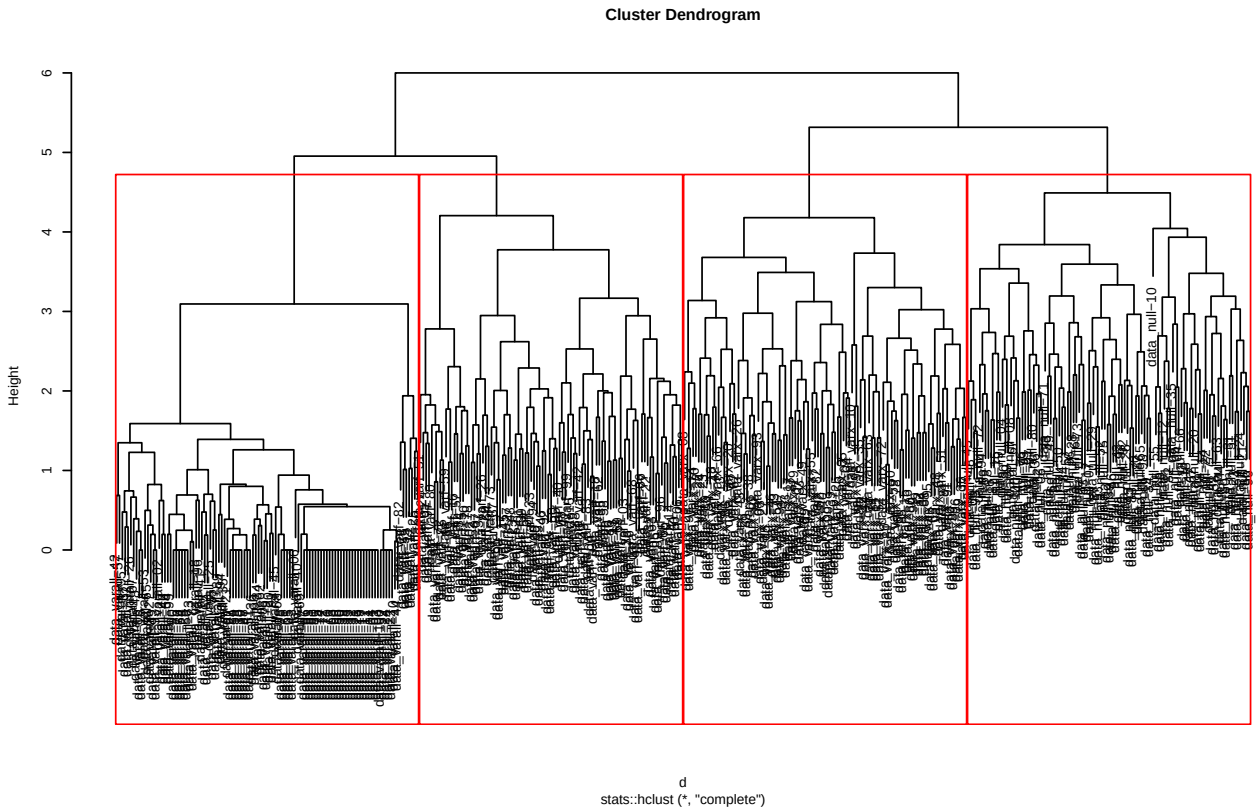

Figure 7: A dendogram showing the branching of 400 series with each of the 4 designs repeated 100 times.

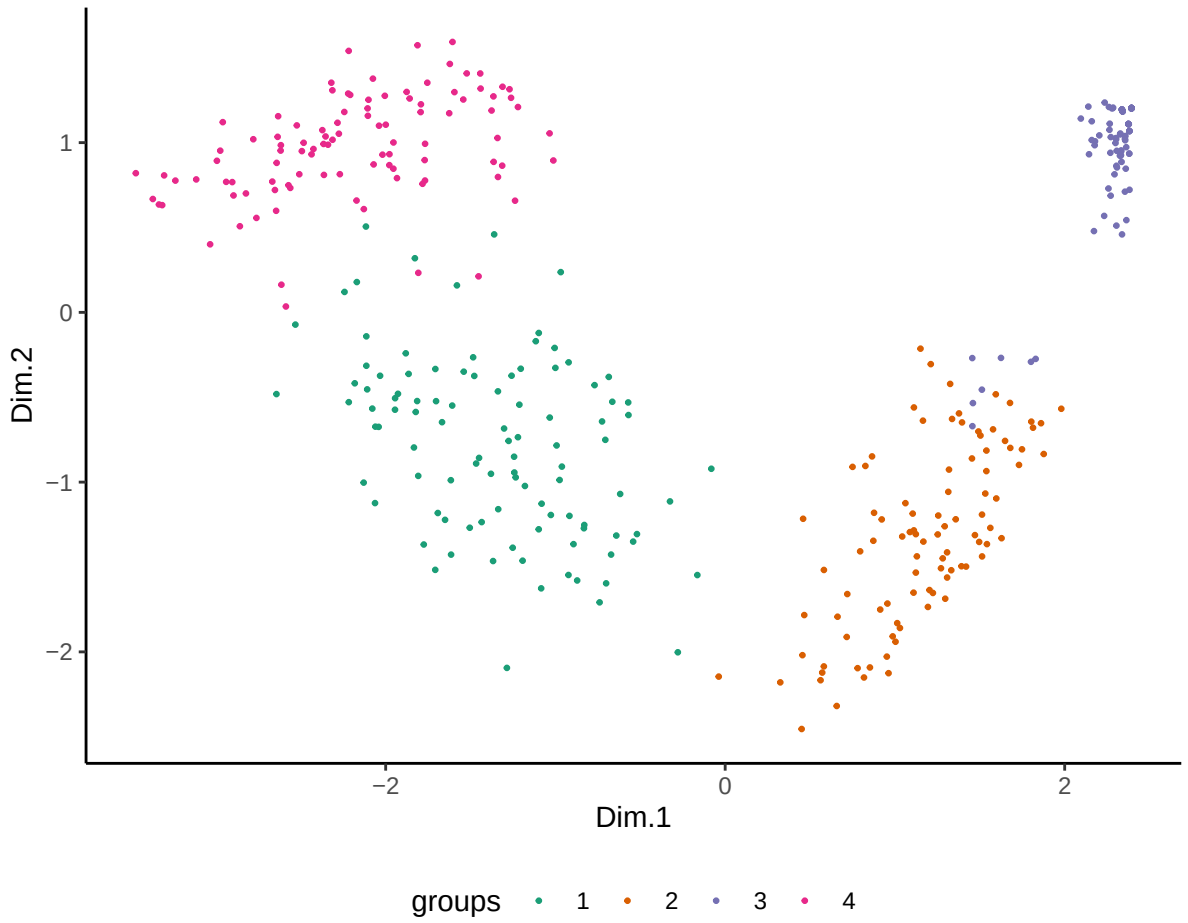

Figure 8: MDS summary plot of 400 series with each of the 4 designs repeated 100 times. It can be observed that all the series belonging to same design lie closer in this plot and all the ones in different designs are placed further away.

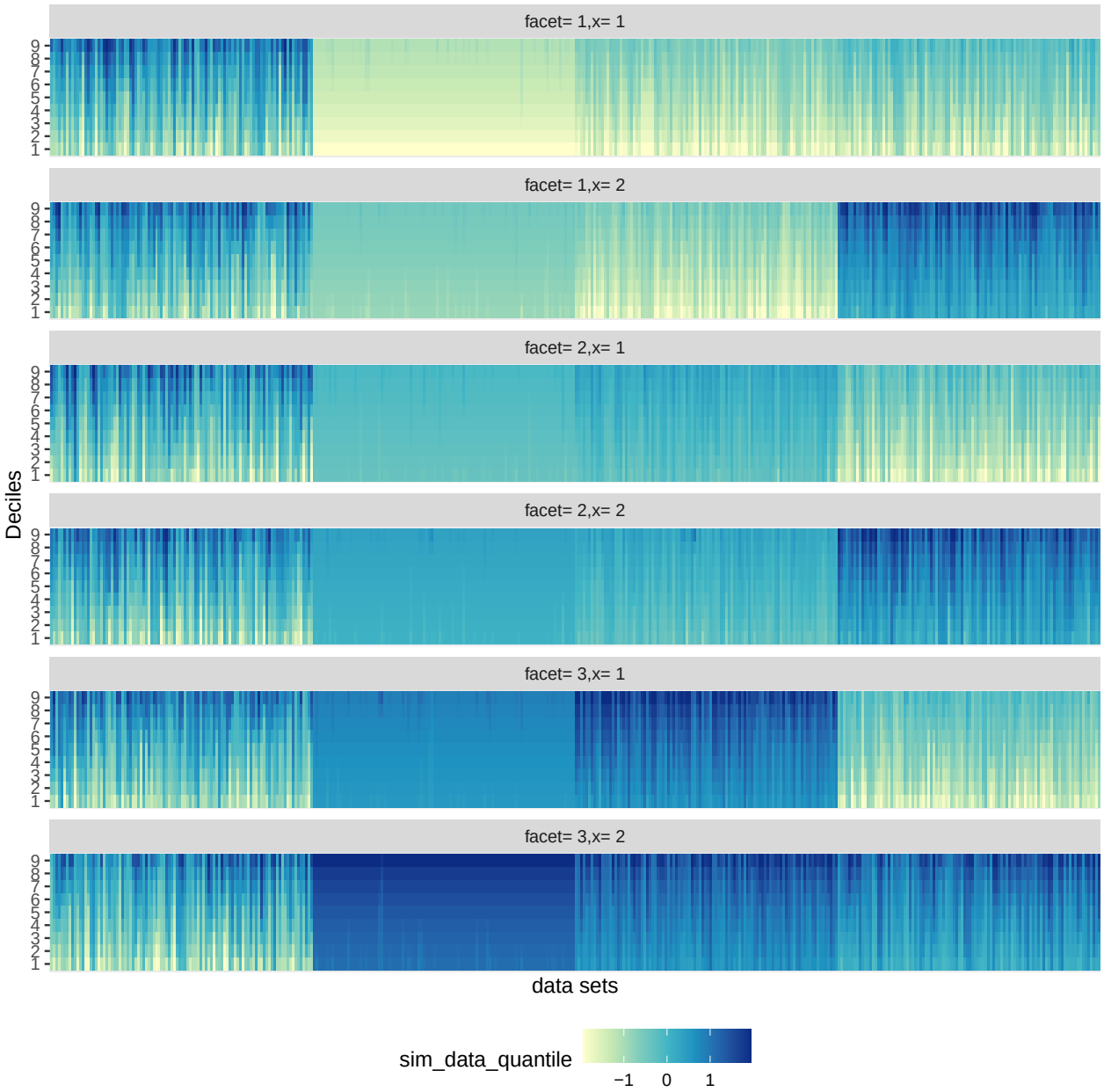

Figure 9: A heatmap showing the cluster summary of 400 series. Deciles of the clusters are plotted along the y-axis with each group on the x-axis faceted by each combination of the interacting granularity. The four distinct behavior across clusters are evident in all the combinations, with the first block corresponding to $data_null$. The deciles of the second block changes across both facet and x-axis. For similar reasons, the second, third and forth blocks correspond $data_varall$, $data_varf$ and $data_varx$ respectively.

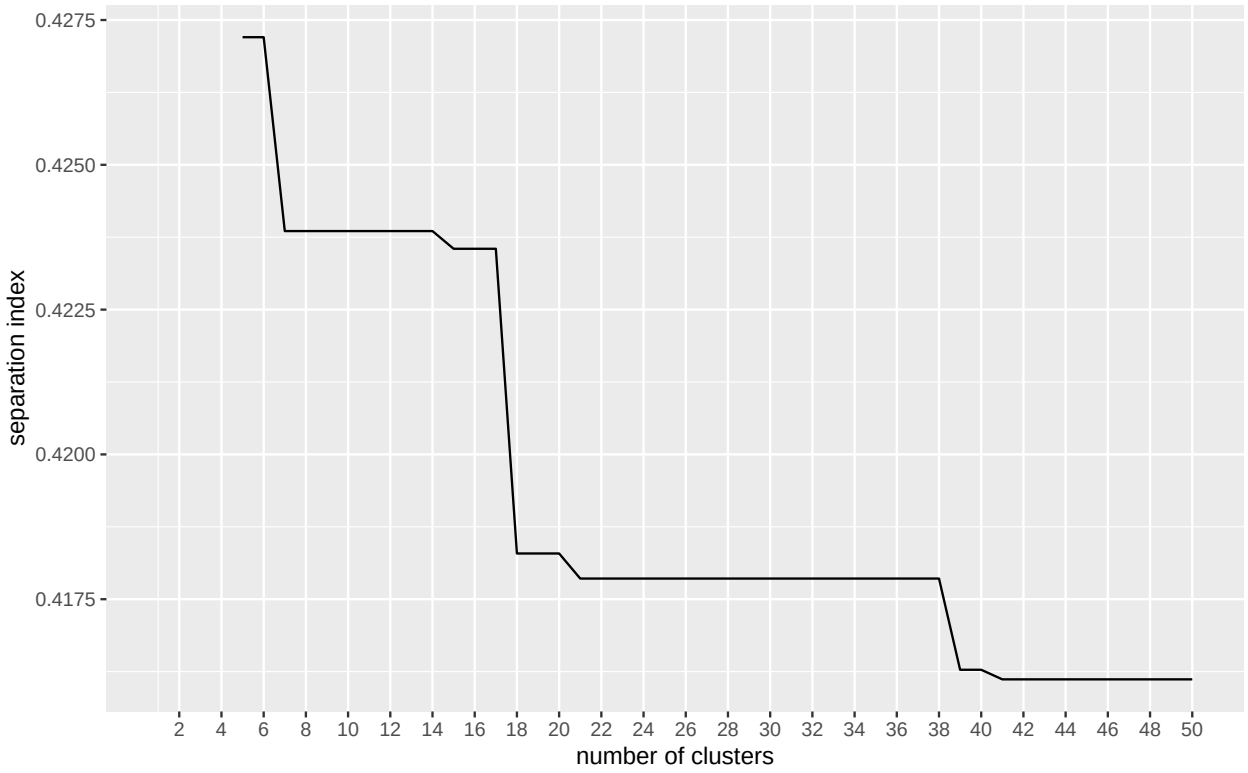

Figure 10: Cluster separation for the 353 customers across the number of clusters is shown. When the cluster size changes from 17 to 18, the separation index drops sharply and then flattens out, resulting in the appearance of the elbow. Hence, when grouping the 353 customers, the number of clusters is taken to be 17.

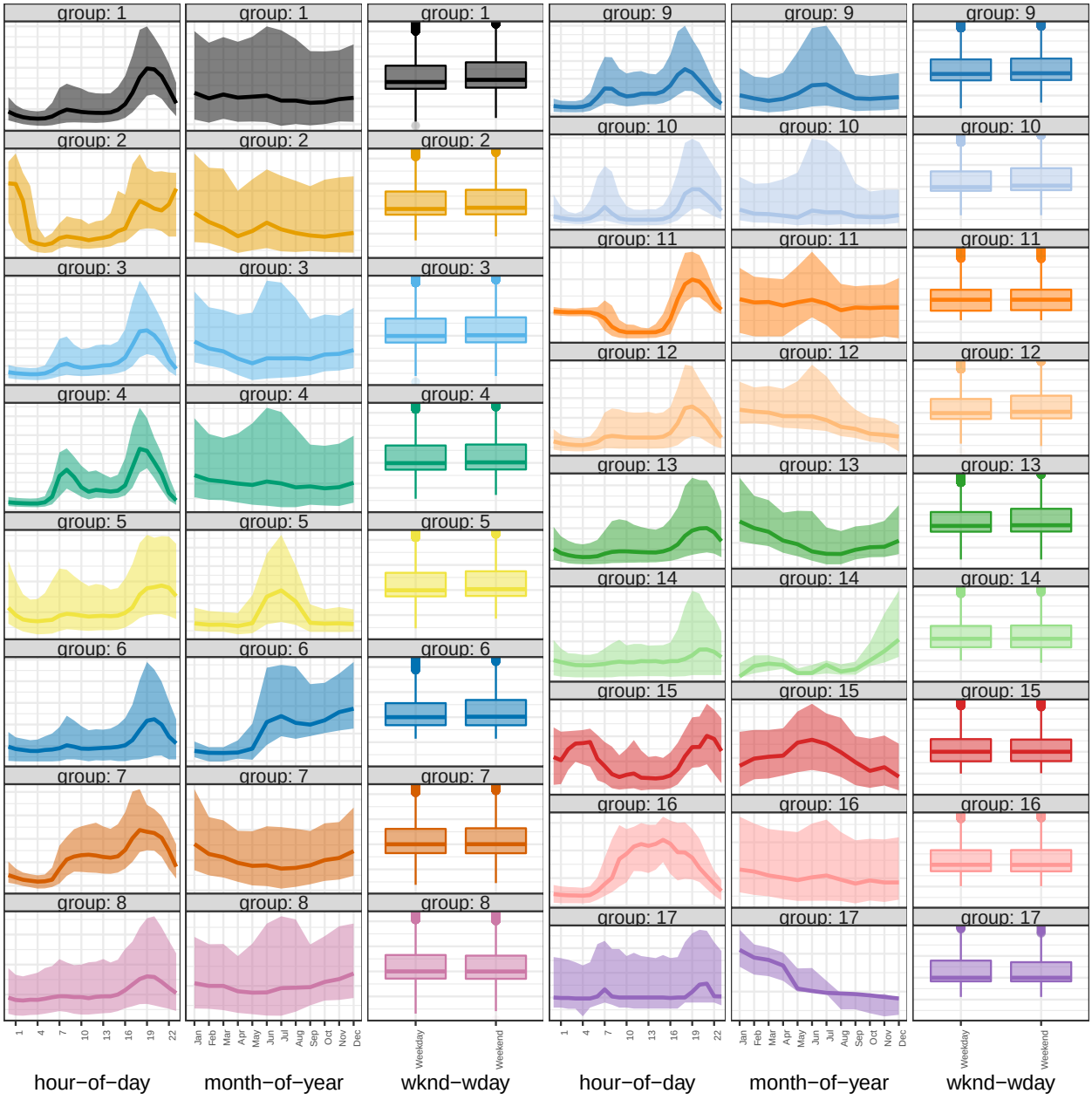

Figure 11: The distribution of electricity demand for the clusters across hod, moy and wkndwday for the 17 groups from 353 customers. Wknd-wday variations across groups are not distinguishable, but ideally each group should have an unique combination of hod and moy.

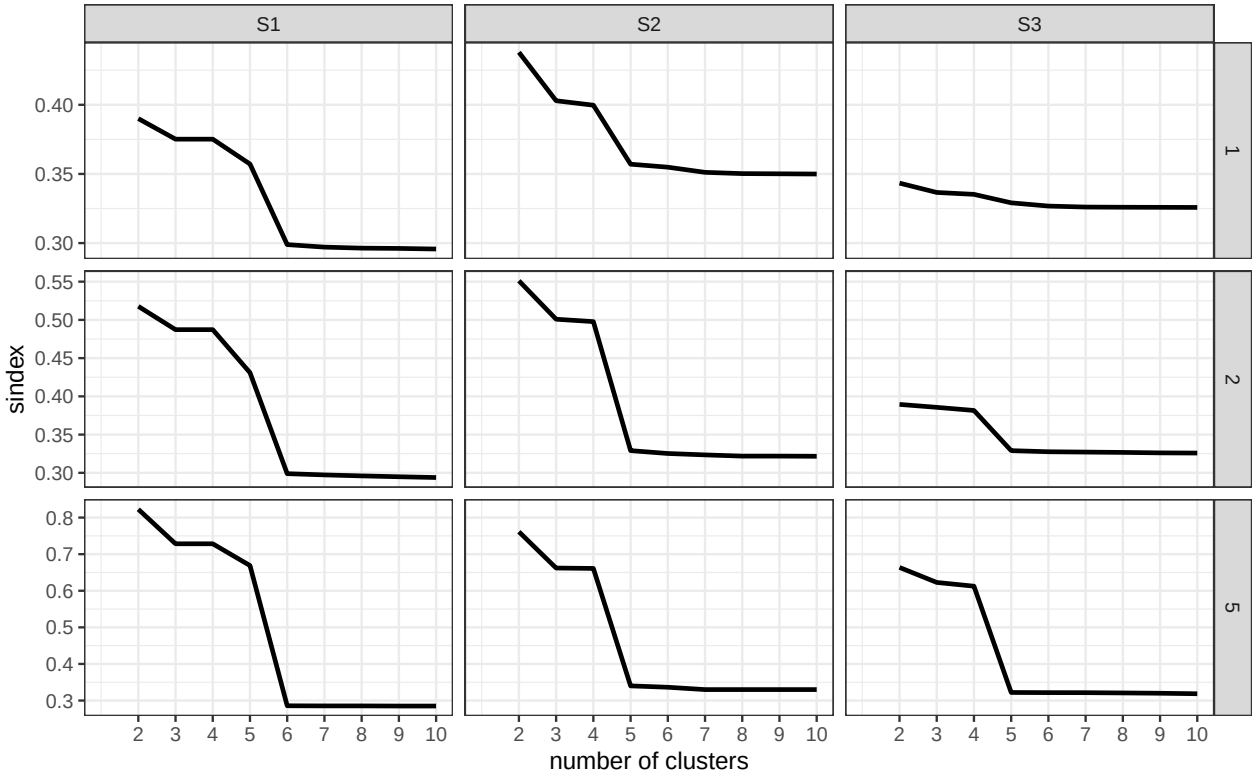

Figure 12: The cluster separation index (sindex) is plotted as a function of the number of clusters for the range of mean differences (rows) under the different scenarios (columns) for JS-RS based distance clustering. S1 has a sharp decrease in sindex from 5 to 6, whereas S2 and S3 have a decrease from 4 to 5. As a result, the number of clusters considered for S1, S2, and S3 is 5, 4, 4, respectively. This corresponds to the number of designs taken into account in each scenario