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

Contents

0.1 Raw time series plot

To get a sense of how the raw time series data looks, we plot the energy usage for 50 sampled households is plotted along the y-axis versus time from past to future in Figure ref{fig:raw-data-50}. Each of these series is associated with a single customer. Energy consumption for each customer is given at fine temporal resolution (every 30 minutes) for a period of 2-3 years.

0.2 Plot displaying missing observations

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.

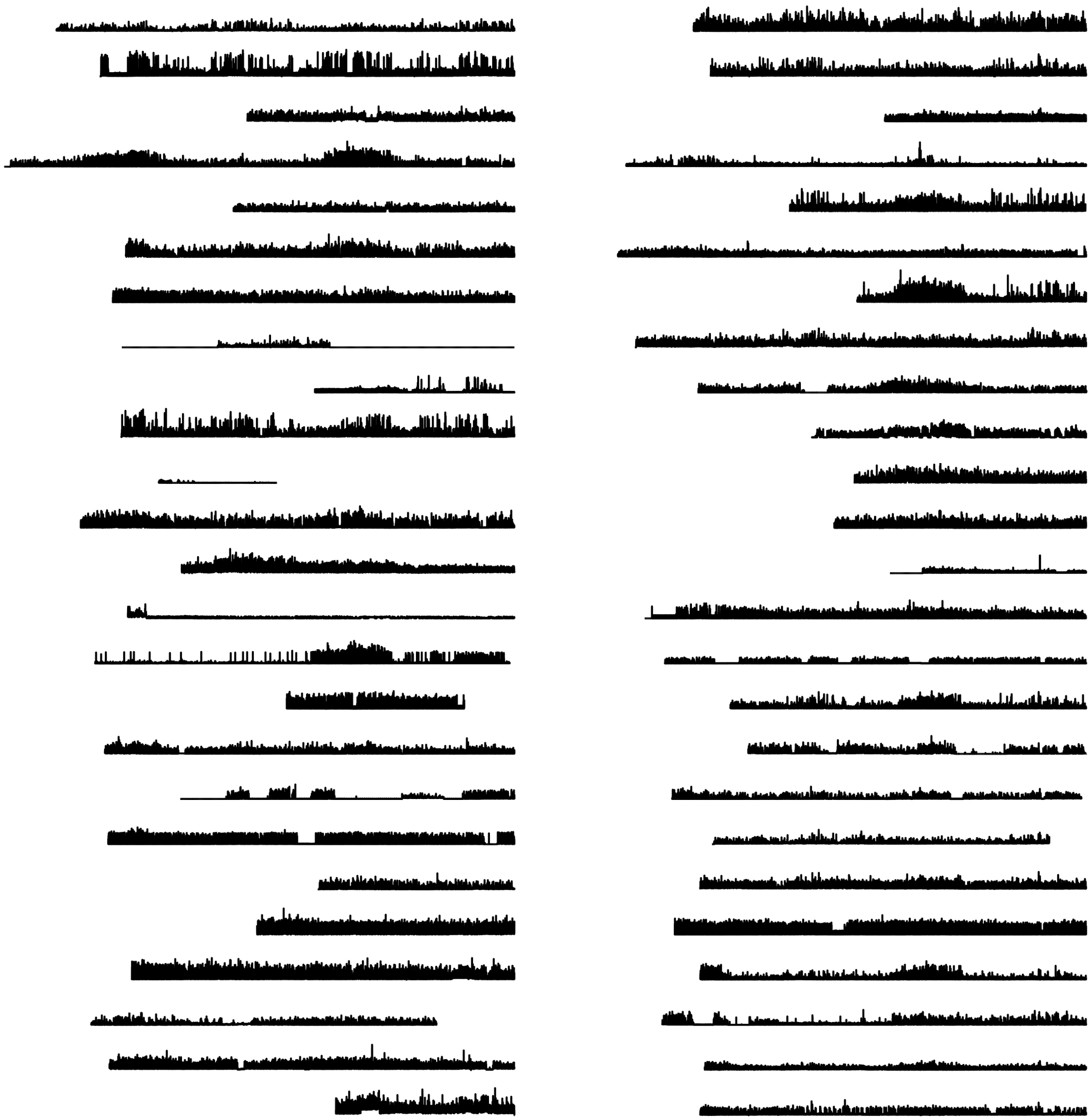


Figure 1: The raw data for 50 households are shown. It looks like there is a lot of missing values and unequal length of time series along with asynchronous periods for which data is observed. No insightful behavioral pattern could be discerned from this view other than when the customer is not at home.

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.

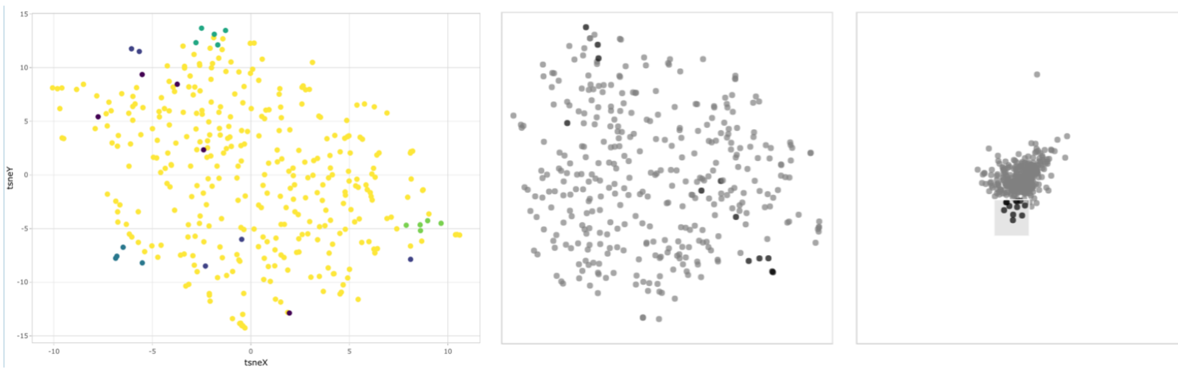
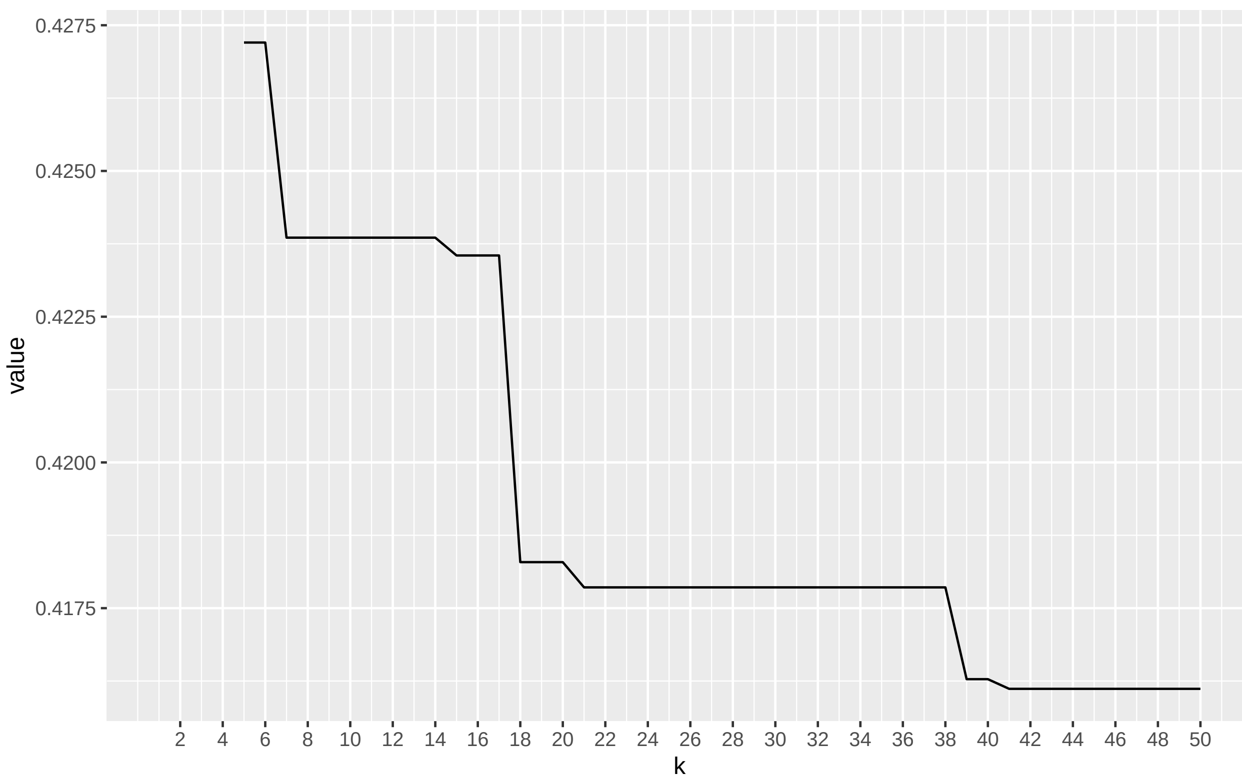
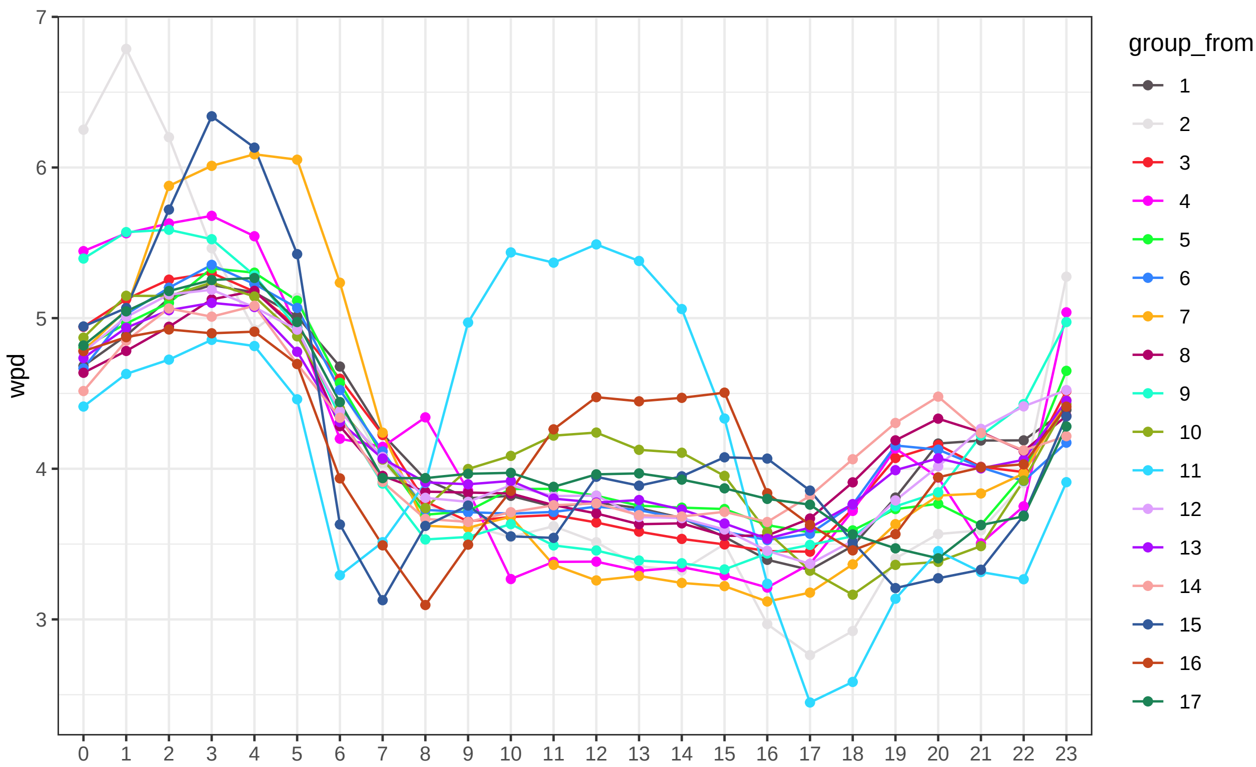


Figure 2: Instance selection using tours and projecting the points in a lower dimensional tsne cloud.

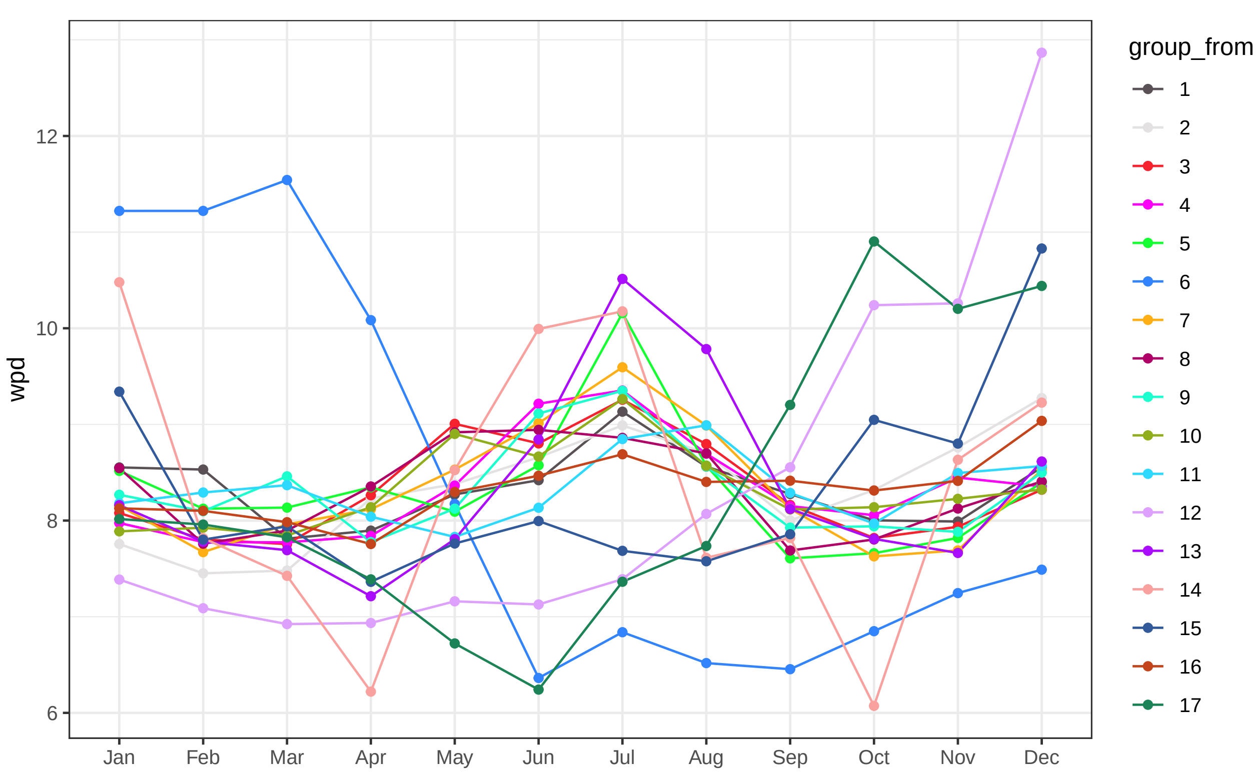
0.3 Clustering on 350 customers

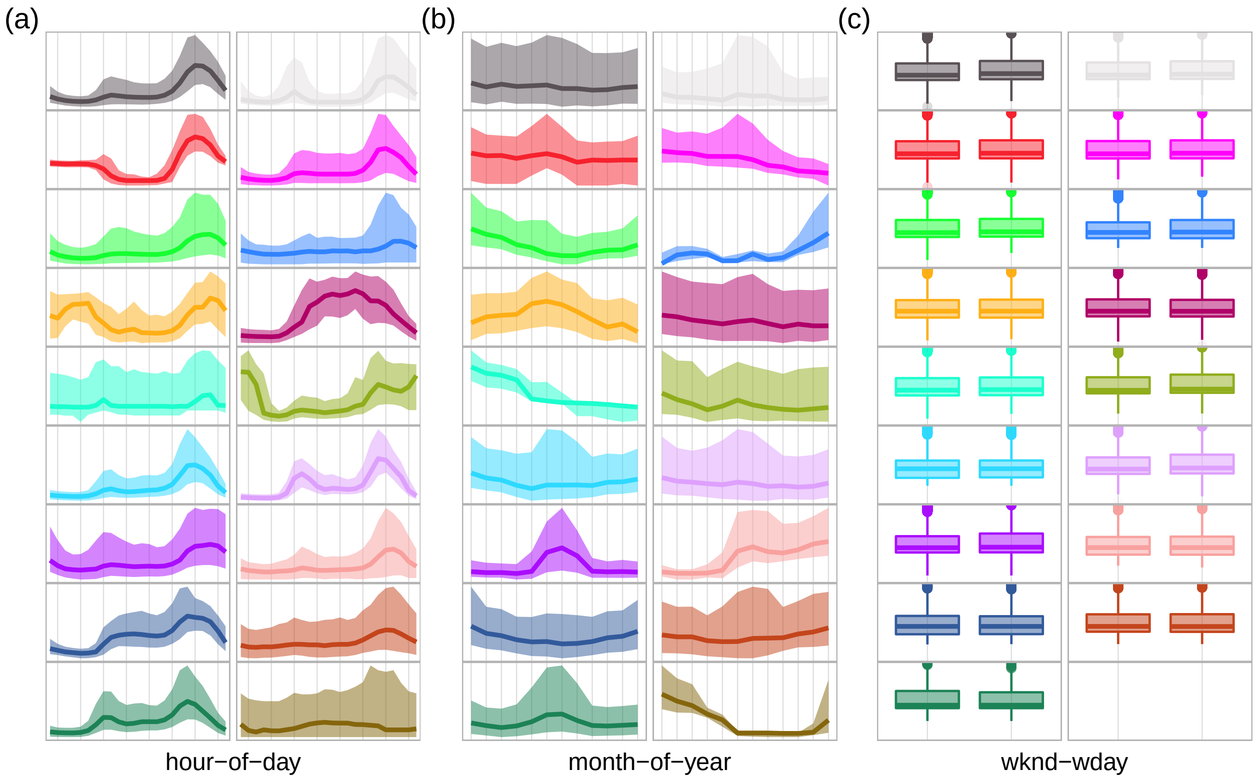


1 contribution of hod



2 contribution of moy




(a)
(b)
(c)
hour−of−day
month−of−year
wknd−wday