Application

28/10/2021

What data you have, what are their features. What you doing in this section.

The use of our methodology is illustrated on smart meter energy usage for a sample of customers from SGSC consumer trial data which was available through Department of the Environment and Energy and Data61 CSIRO. It contains half-hourly general supply in KwH for 13,735 customers, resulting in 344,518,791 observations in total. No additional information about the customers is used for the application.

The linear view of the time series data generally has too many measurements all squeezed in that representation, it hinders us to discern any repetitive behavioral pattern for even one customers (let alone many customers together). In most cases, electricity data will have multiple seasonal patterns like daily, weekly or annual. We do not learn about these repetitive behaviors from the linear view. Hence we transition into looking at cyclic granularities, that can potentially provide more insight on their repetitive behavior. The raw data for these consumers is of unequal length, with varying start and finish dates. Because our proposed methods evaluate probability distributions rather than raw data, neither of these data features would pose any threat to our methodology unless they contained any structure or systematic patterns. Additionally, there were missing values in the database but further investigation revealed that there is no structure in the missingness (see Supplementary paper for raw data features and missingness).

Prototype selection

Supervised learning uses a training set of known information to categorize new events. Instance selection (@olvera2010review) is a method of rejecting instances that are not helpful for classification. This is equivalent to subsetting the population along all dimensions of interest such that the sampled data reflects the underlying distribution's primary features. Instance selection in unsupervised learning has received little attention in the literature, yet it could be a useful tool for evaluating model or method performance. @Fan2021-bq proposes one such process that picks related examples (neighbours) for each instance (anchor) and considers them as the same class. In this part, consumers with prototype behaviors are chosen to serve as study population for our suggested methodology.

Data filtering and variable selection

- Choose a smaller subset of randomly selected 600 customers with no implicit missing values for 2013.
- Obtain *wpd* for all cyclic granularities considered for these customers. It was found that `hod` (hour-of-day), `moy` (month-of-year) and `wkndwday` (weeknd/weekday) are coming out to be significant for most customers. We use these three granularities while clustering.
- Remove customers whose data for an entire category of a significant granularity is empty. For example, a customer who does not have data for an entire month is excluded because their monthly behaviour cannot be analyzed.
- Remove customers whose energy consumption is 0 in all deciles. These are the clients whose consumption is likely to remain essentially flat and with no intriguing repeated patterns that we are interested in studying.

There are several ways to approach the prototype selection. Use any dimensionality reduction techniques like MDS or PCA to project the data into a 2-dimensional space. Then pick a few "anchor" customers who are far apart in 2D space and pick a few neighbors for each. Unfortunately, this does not assure that consumers with significant patterns across all variables are chosen. We perform a linked tour with t-SNE layout and (liminal) to identify customers who are more likely to have distinct patterns across the variables studied. Tours can help us see separation between variables that was not obvious in the single variable display. Please see the Supplementary article for further details on how the prototypes are chosen. Figure **??** shows the raw time plot, distribution across `hod`, `moy` and `wkndwday` for the set of chosen 24 customers. Few of these customers have similar distribution across `moy` and some are similar in their `hod` distribution. These 24 prototypes are clustered using the methodology described in **??** to see if the grouping is useful.

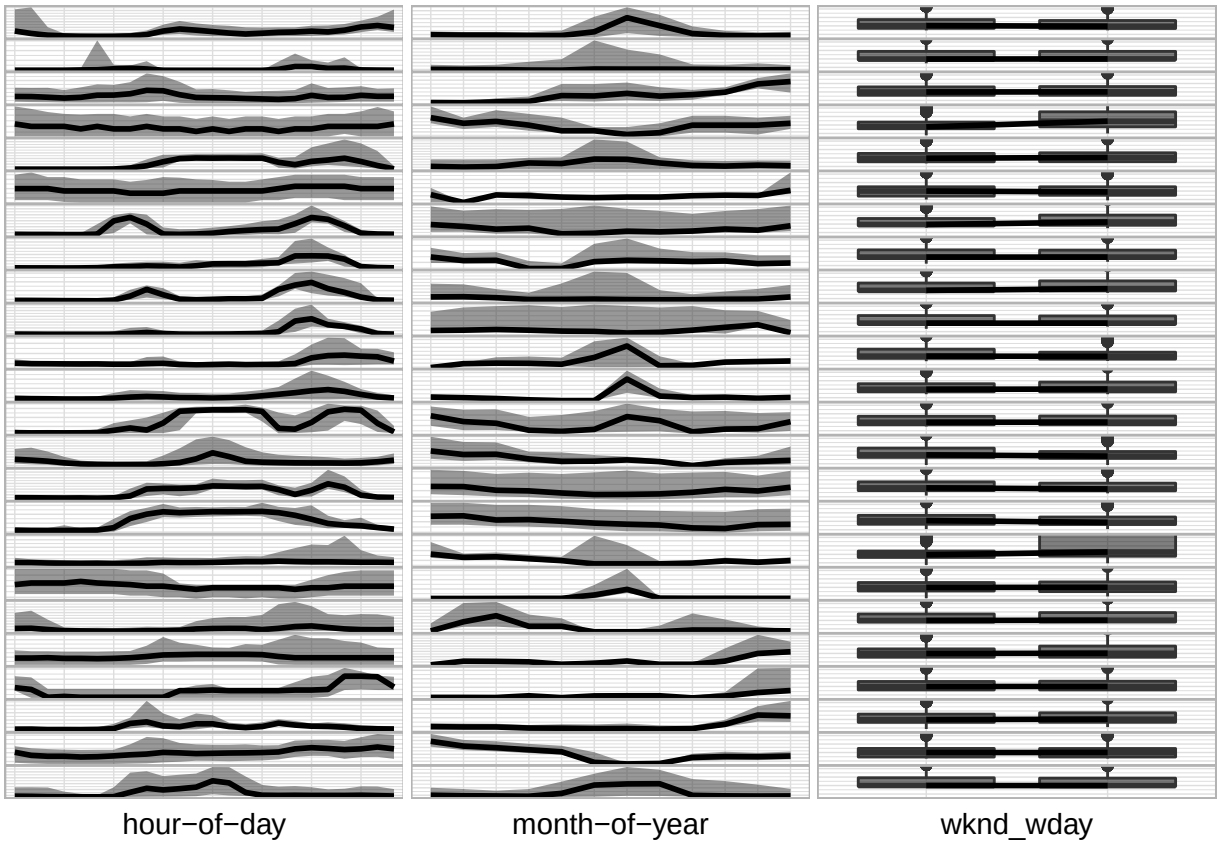

Figure 1: The distribution across moy, hod and wkndwday for the selected designs. Few are similar in their hod pattern, while others are similar in moy behavior. Some customers have distinct behavior as compared to all other customers.For example, although patterns across wkndwday do not look distinctly different for most households, there is one household for whom weekend behavior is standing out from the rest.

Clustering

Using JS-based distances

The 24 prototypes are first clustered using the methodology described in **??**. We chose the optimal number of clusters using (@Henning-2013) as 5. The distribution of electricity demand for the selected 24 customers across hour-of-day and month-of-year are shown in **??** respectively. The median is shown by a line, and the shaded region shows the area between the 25^{th} and 75^{th}. All customers with the same color represent the same design (left) or cluster (right). The plotting scales are not displayed since we want to emphasize

comparable shapes rather than scales. A customer in the cluster may have low daily or total energy usage, but their behavior may be quite similar to a customer with high usage.

```
## Joining, by = "customer_id"
```

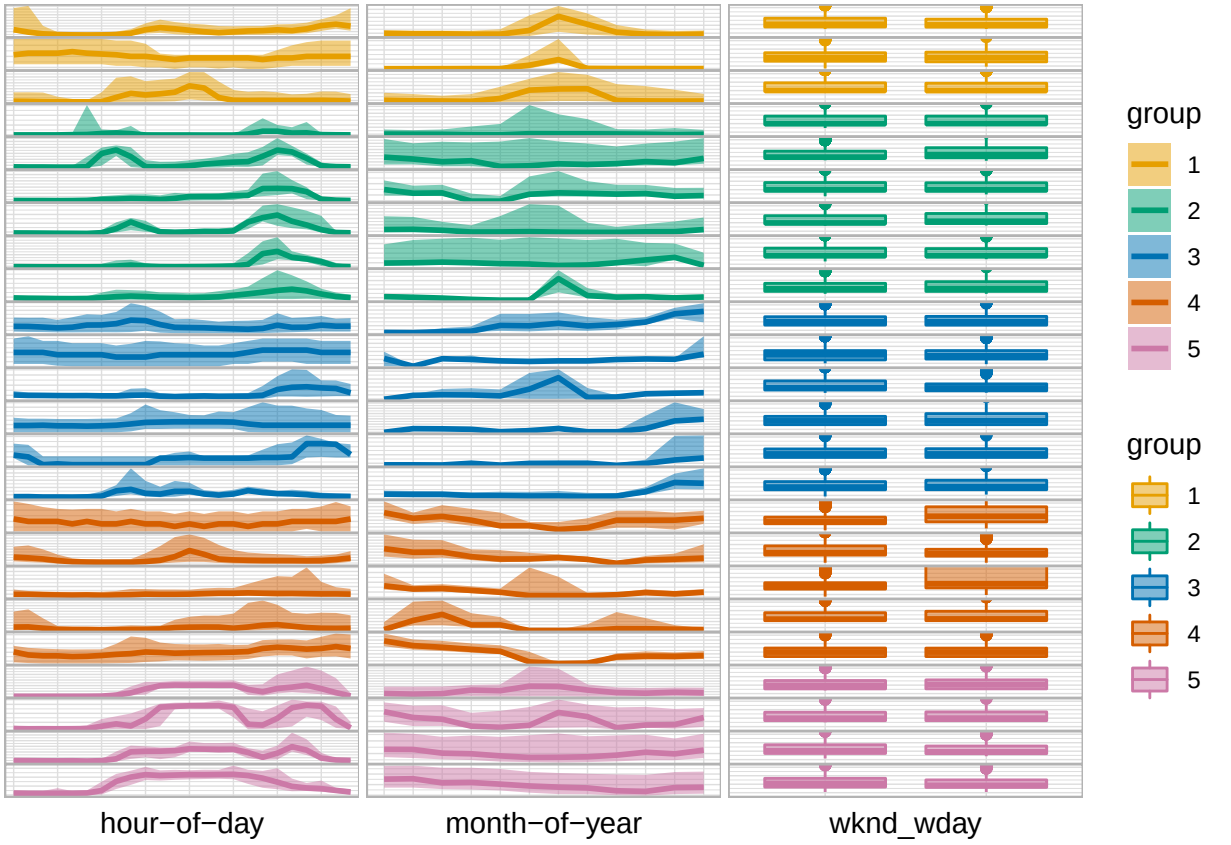

Figure 2: The distribution across moy, hod and wkndwday of each customer is shown. The same color represents same group.Our clustering methodology is useful for grouping together comparable distributions over hod and moy. Of course, certain consumers in each group have distributions that differ from the distributions for other customers in the same group. However, it appears that the goal of finding similar distribution across significant cyclic granularities has been established.

Using wpd-based distances

A parallel coordinate plot with the three significant cyclic granularities used for wpd-based clustering. The variables are sorted according to their separation across classes (rather than their overall variation between classes). This means that *moy* is the most important variable in distinguishing the designs followed by *hod* and *wkndwday*. It can be observed that clusters are well separated by *moy*, while *hod* and *wkndwday* are not useful distinguishing the clusters produced with this clustering method. The parallel coordinate plot ranks the variables in order of importance, indicating that the month-of-year is the most important in identifying clusters, whereas wkdn-wday is the least significant and has the least variability among the three variables. However, there is only one customer who has significant *wpd* across *wkndwday* and stands out from the rest of the customers. The ggpairs plot also shows five distinct clusters across the *moy*.

- Discussion

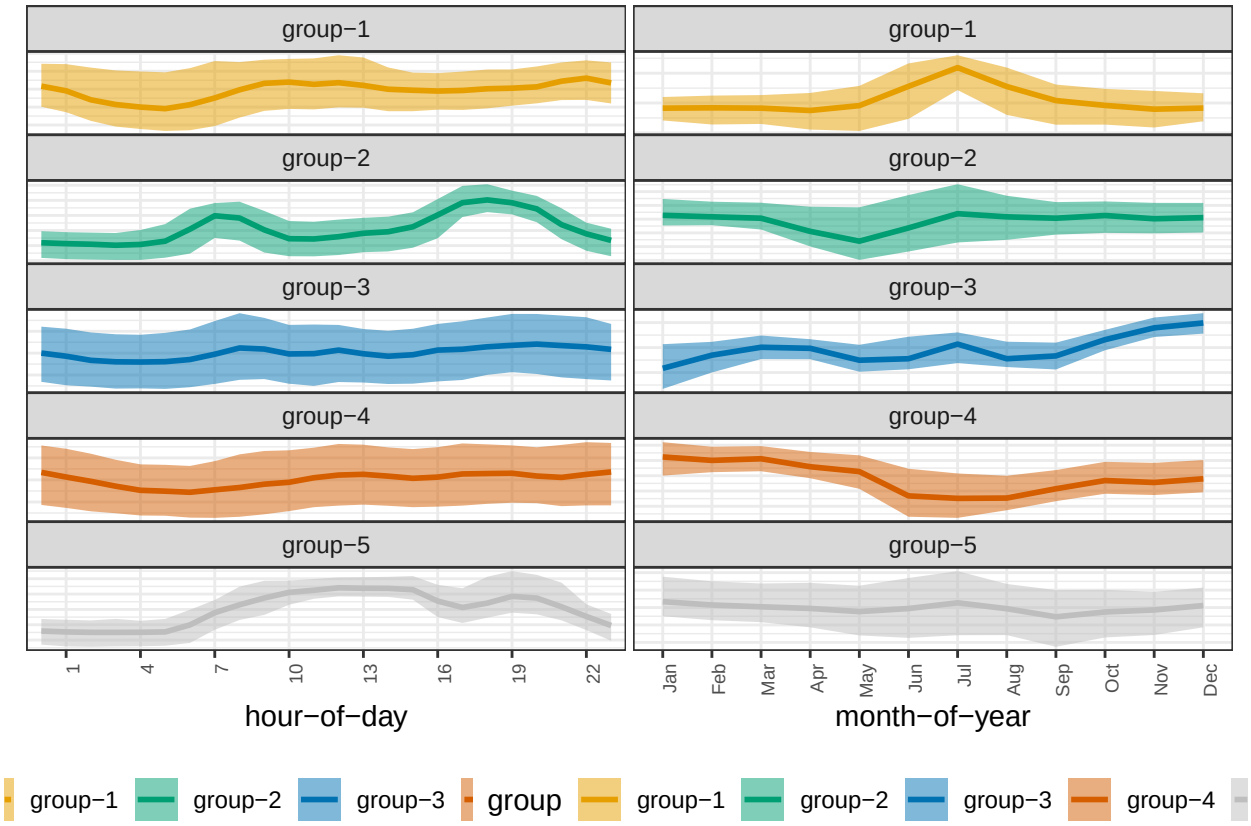

Figure 3: The distribution of electricity demand for the clusters across hour-of-day, month-of-year and wknd-wday. The median is represented by a line and the shaded region represents the area between 25^{th} and 75^{th} percentile. Group 2 and 5 have a stronger hour-of-day pattern, while group 1, 3, 5 have a month-of-year pattern. For wknd-wday differences across different groups are not distinct suggesting that it might not be that important a variable to distinguish different clusters.

As with any clustering method, things become much more complicated when we consider a larger data set with more uncertainty. The methodology run on several customers together might not be useful to given distinct shapes across granularities. Moreover, the groupings from two different appraoches lead to different results, both of which are useful but needs careful considering the context before choosing one over another. Also, note that the customers chosen here do not have a weekend-weekday effect, which might not be true for all customers in the data set. Ideally there are other groups of customers in the entire dataset which can act as prototypes in our problem. Ideally, if these prototypes are not outliers, we can use them for a classification problem for external validation of the clustering problem.

Notice that while looking for these customers

(can be used as classification method, problems in handling large dataset)