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. In most cases, electricity data is expected to have multiple seasonal patterns like daily, weekly or annual. We do not learn about these repetitive behaviors from the linear view because too many measurements all squeezed in that representation. 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). The study begins by subsetting a data set along all dimensions of interest using data filtering and prototyping. By grouping the prototypes using our methods and assessing their meaning, the study hopes to unravel some of the heterogeneities observed in energy usage data. Because our application does not employ additional customer data, we cannot explain why consumption varies, but rather try to identify how it varies.

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.

Prototype selection

Supervised learning uses a training set of known information to categorize new events through instance selection. Instance selection (@olvera2010review) is a method of rejecting instances that are not helpful for classification. This is analogous to subsampling the population along all dimensions of interest such that the sampled data represents the primary features of the underlying distribution. 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. There are several ways to approach the prototype selection. Following @Fan2021-bq's idea of picking related examples (neighbours) for each instance (anchor), we can first 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. Tours can reveal variable separation that was hidden in a single variable display. Hence we perform a linked tour with t-SNE layout (using the R package @R-liminal) to identify customers who are more likely to have distinct patterns across the variables studied. 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.

Clustering

JS-based distances

The 24 prototypes are first clustered using the methodology described in **??**. We chose the optimal number of clusters using (@Hennig2014-ah) 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 thecluster (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.

Cluster characterization is a crucial aspect of cluster analysis. Figure ref{fig:combined-groups-js} depicts the summarized distributions across groups and assists us in characterizing each cluster. All of these may be validated only with further information about the customer. Figure \ref{fig: Groups 2 and 5 show a stronger hour-of-day pattern, whereas groups 1, 3, and 5 show a month-of-year pattern. Differences in wknd-wday between groups are not discernible, implying that it may not be a relevant variable in distinguishing various clusters.

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 wknd-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*.

Entire data subset

Things become far more complicated when we consider a larger data set with more uncertainty, as they do with any clustering problem. Summarizing distributions across clusters with varied or outlying customers can result in a shape that does not represent the group. Furthermore, combining hetergeneous customers may result in similar-looking final clusters that are not effective for visually differentiating them. It is also worth noting that the wknd-wday behavior in the given case does not characterize any cluster. This, however, will not be true for all of the customers in the data set. If more extensive prototype selection is used, resulting in more comprehensive prototypes in the data set, this method might be used to classify the entire data set into these prototype behaviors. However, the goal of this section was to have a few customers that have significant patterns over one or more cyclic granularities, apply our clustering methodology to cluster them, and demonstrate that the method produces useful clusters.

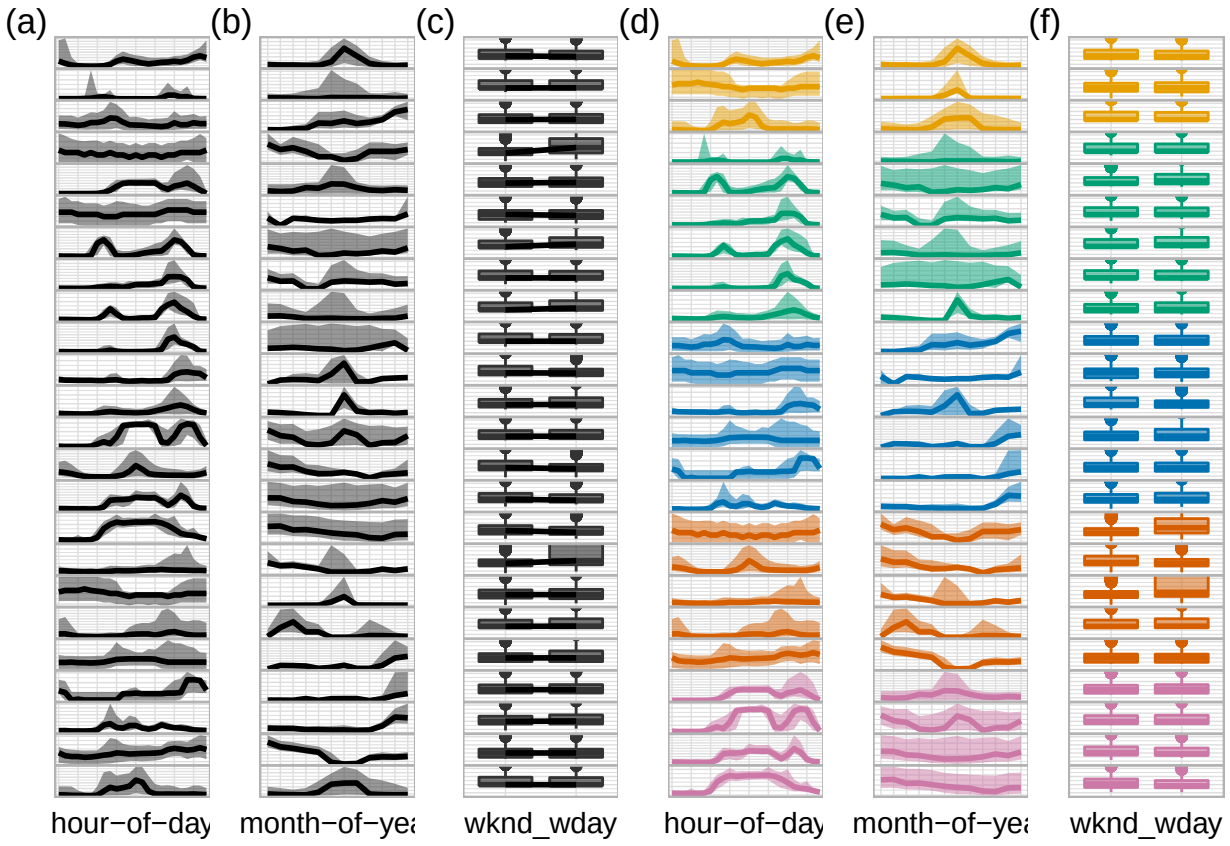


Figure 1: The distribution of selected consumers over hod (a, d), moy (b, e), and wkndwday (d, f). In each case, the same colour denotes the same group in plots (d), (e), (f) and are placed together to facilitate comparison. That means the customer orderings are different for (a, b, c) and (d, e, f). Our clustering methodology is useful for grouping similar distributions over hod and moy. Of course, certain customers in each group have distributions that differ from those of other members in the same group. However, it appears that the aim of grouping comparable distributions over considered variables has been accomplished to some extent.

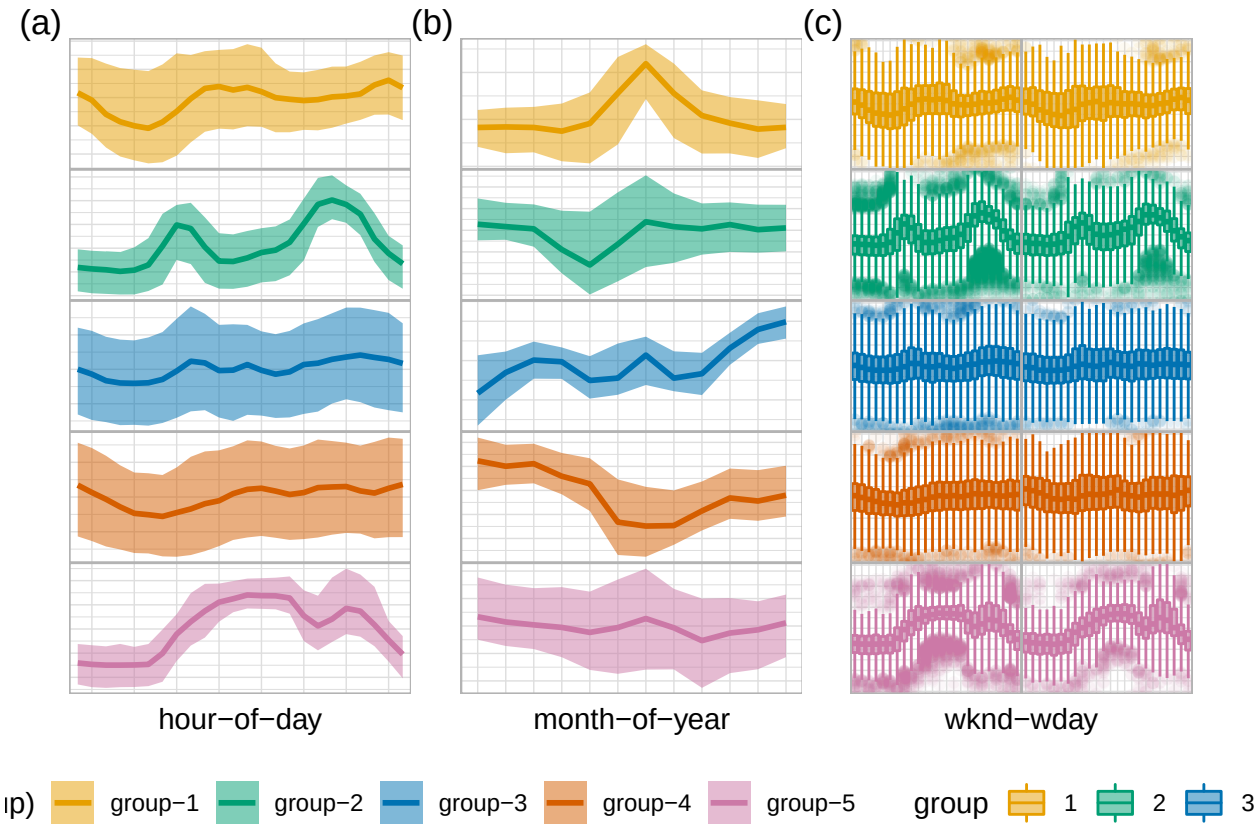
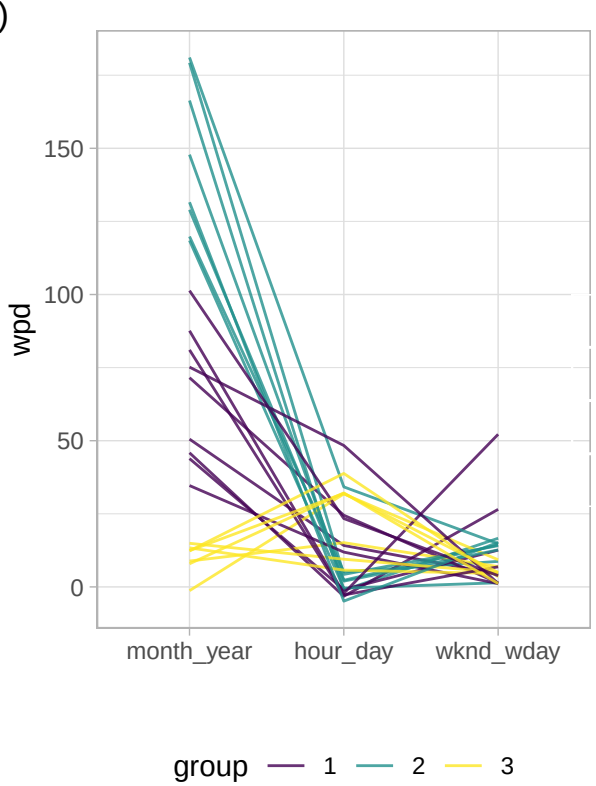


Figure 2: The distribution of electricity demand for the clusters across hod (a), moy (b) and wkndwday (c). It seems like group 2 and 5 have a hod pattern across its members, while group 1, 3, 5 have a moy pattern. Wknd-wday variations across groups are not distinguishable, indicating that it is not 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.

(a)



(b)

	group	nobs	moy	hod	wnwd
1	1	9	71.55	11.91	4.38
2	2	8	139.70	2.13	14.24
3	3	7	12.33	31.80	4.95

Figure 3: Each of the 24 customers is represented by a parallel coordinate plot (a) with three wpd-based groupings. The plot shows that moy is the most important variable in identifying clusters, whereas wkdn-wday is the least significant and has the least fluctuation. One particular customer with high wpd across wknwday stands out in this display. Group 3 has a higher wpd for hod than moy or wkndwday. Group 2 has most discernible pattern across moy. Group 1 is a mixed group with strong patterns on atleast one of the three variables. All of these could be observed from the plot or the table (b) which shows median wpd values for each group.