Clustering based on probability distributions with application on residential customers

Sayani Gupta *
Department of Econometrics and Business Statistics, Monash University
and
Rob J Hyndman
Department of Econometrics and Business Statistics, Monash University
and
Dianne Cook
Department of Econometrics and Business Statistics, Monash University

October 13, 2021

Abstract

Clustering elements based on behavior across time granularities

Keywords: data visualization, statistical distributions, time granularities, calendar algebra, periodicties, grammar of graphics, R

*Email: Sayani.Gupta@monash.edu

0.1 js-based clustering

The distribution of electricity demand for the selected 24 customers across hour-of-day is shown in 1. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. According to our prototype selection method, each row represents a distinct shape of daily load, and all customers in the same row should have similar daily profile.After clustering these consumers, all customers with the same colour represent the same group. Except for customer id 8269176, there is unanimity across design and groups. Even though their daily form resembles Group 2, our clustering approach places them in Group 3 (which is design 4 in our case). Because our method uses hod, moy, and wkndwday, there may be some mismatches depending on only one variable.
The distribution of electricity demand for the selected 24 customers across month-of-year is shown in Figure 2. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. Intriguingly, customer id 8269176 appears to be in the appropriate group (Group 3) because its monthly profile is more similar to Group 3 than Design 4. So, while our clustering approach failed to place the customer in the correct hour-of-day group, it did so for month-of-year. When contrasting the moy to hod profile, there are greater behavioural differences across customers within a group.

Characterization of clusters is the final stage of a cluster analysis. If we are convinced that we have identified a collection of clusters that can be distinguished from one another, we should intend to characterize them more formally, both statistically and qualitatively. We quantify them by producing statistics for each variable. We may look at the findings in graphs, and enhance our qualitative descriptions of the groupings.
We discover 4 qualitative clusters of varying shapes in the distribution of all consumers in the first panel of Figure 3. Group-1 includes consumers who work 9-5, get up and conduct morning activities from 7-10am, and then depart. Then they return home in the evening to cook supper and perform other activities, giving the evening a greater peak than the morning. Group 2 is the group that rushes out of the house in the morning to get to work. They only return at night and do all activities at night, so there is no morning peak. The third one has a strong early morning and late night hours. These consumers may be flexible students or elderly retirees who are night owls. Presence of children or

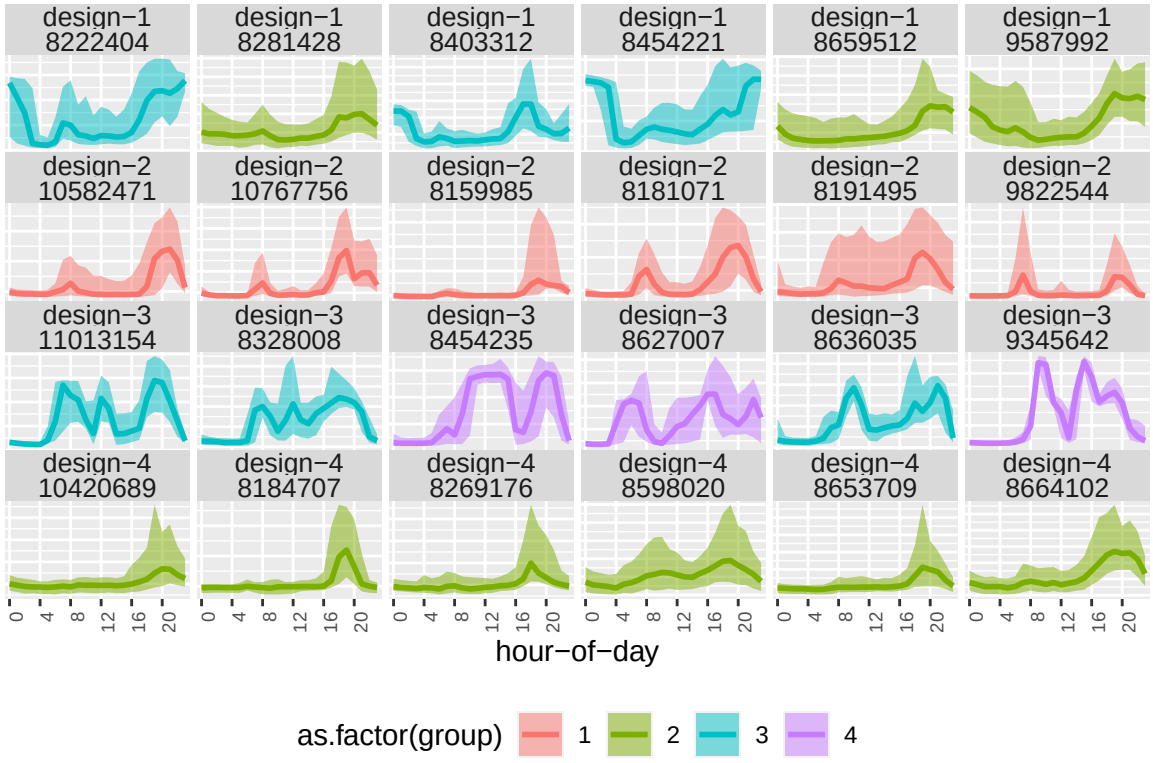

Figure 1: The distribution of electricity demand for the selected 24 customers across hour-of-day. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. According to our prototype selection method, each row represents a distinct shape of daily load, and all customers in the same row should have similar daily profile.After clustering these consumers, all customers with the same colour represent the same group. Except for customer id 8269176, there is unanimity across design and groups. Even though their daily form resembles Group 2, our clustering approach places them in Group 3 (which is design 4 in our case). Because our method uses hod, moy, and wkndwday, there may be some mismatches depending on only one variable.

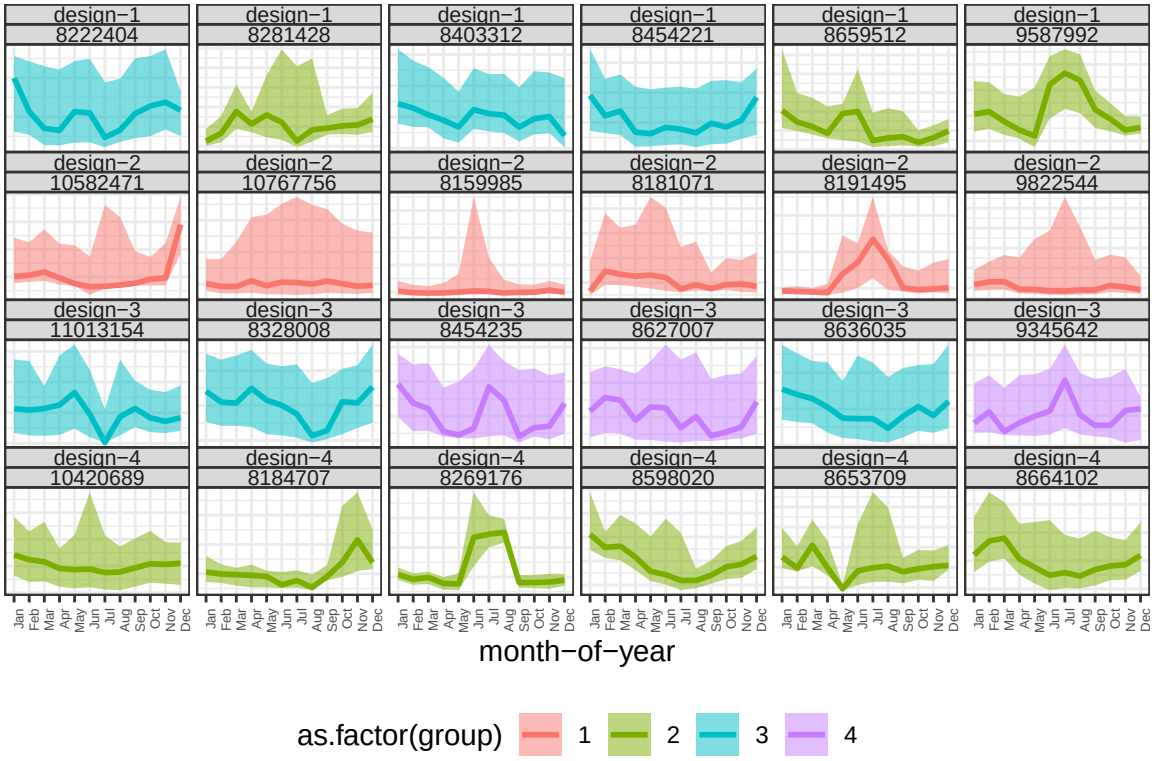

Figure 2: The distribution of electricity demand for the selected 24 customers across month-of-year. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. Intriguingly, customer id 8269176 appears to be in the appropriate group (Group 3) because its monthly profile is more similar to Group 3 than Design 4. So, while our clustering approach failed to place the customer in the correct hour-of-day group, it did so for month-of-year. When contrasting the moy to hod profile, there are greater behavioural differences across customers within a group.

stay-at-home parents is indicated by Group-4's almost equivalent morning, afternoon and evening profile. All of this may be validated with further information about the customer. The second panel of Figure 3 shows that month-of-year qualitative clusters are not as distinguishable as hour-of-day. Group 2 is the most distinct and uses the most power during the summer, possibly owing to the use of air conditioners. Group 4 has a flat profile, indicating no significant month-to-month changes. Groups 1 and 3 have heaters on in the winter but consume less energy in the summer. Since gas is not available in all of NSW LGAs, it is possible that customers' heater usage is recorded in electricity rather than gas.

The third panel of Figure 3 shows that the wknd-wday groups exhibit no significant changes across clusters, indicating that they may be a nuisance variable for these consumers.

The plotting scales are not displayed since we want to emphasise comparable shapes rather than scales. A customer in the cluster may have low daily or total energy usage, but their behaviour may be quite similar to a customer with high usage. That places them in the same group.

0.2 wpd-based clustering

```
## List of 1
##  $ legend.position: chr "bottom"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

We discover 4 qualitative clusters of varying shapes in the distribution of all consumers in the first panel of Figure 3. Group-1 includes consumers who work 9-5, get up and conduct morning activities from 7-10am, and then depart. Then they return home in the evening to cook supper and perform other activities, giving the evening a greater peak than the morning. Group 2 is the group that rushes out of the house in the morning to get to work. They only return at night and do all activities at night, so there is no morning peak. The third one has a strong early morning and late night hours. These consumers may be flexible students or elderly retirees who are night owls. Presence of children or

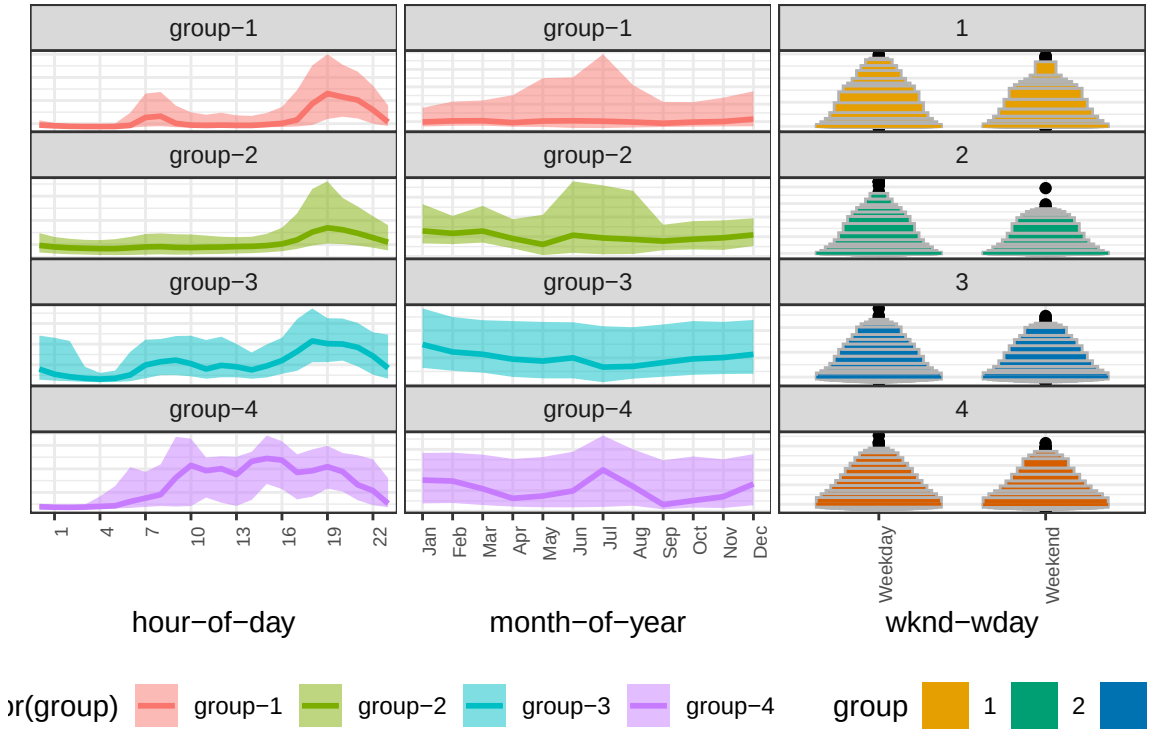

Figure 3: The distribution of electricity demand for the clusters across hour-of-day. The median is represented by a line and the shaded region represents the area between 25^{th} and 75^{th} percentile. Each cluster is characterised by unique shape across the granularity it is plotted against. For wknd-wday differences across different groups are not distinct suggesting that it might not be that important a variable to distinguish different clusters. This fact we will be re-established when we see the importance of each granularity throough the parallel coordinate plot.

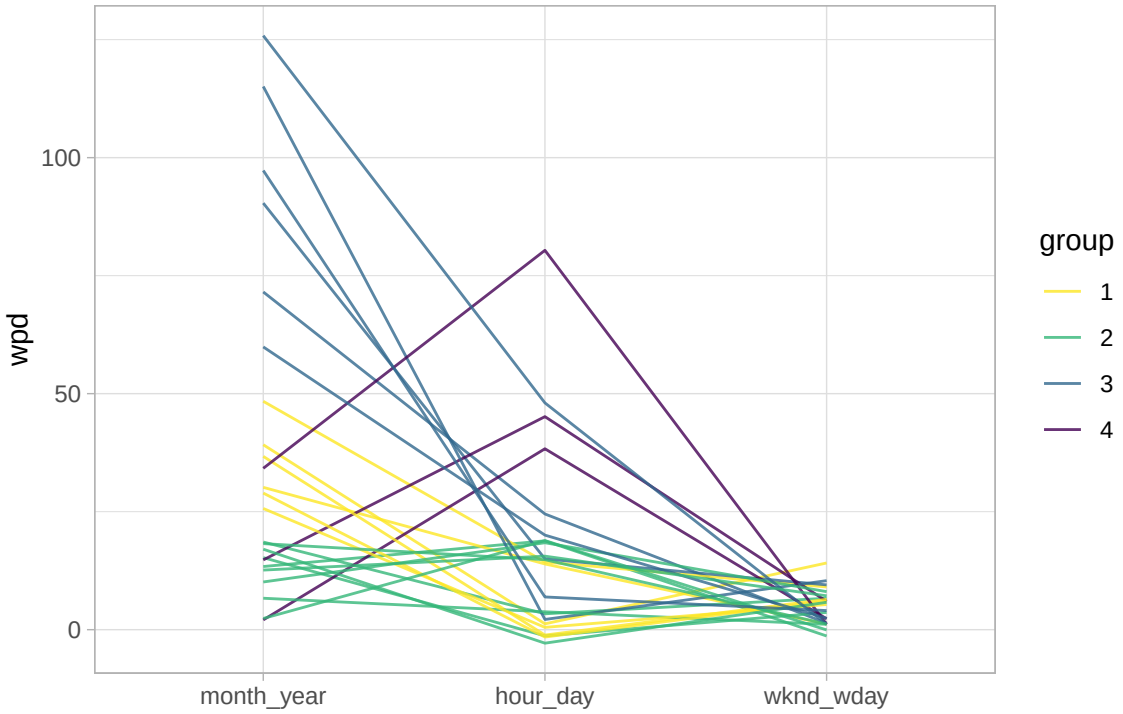

Figure 4: A parallel coordinate plot with the three significant cyclic granularities used for wpd-based clustering. The variables are sorted according to their separation across classes (rather than their overall variation between classes). This means that moy is the most important variable in distinguishing the designs followed by hod and wkndwday. It can be observed that cluster 3 and 4 are distinguished by moy while 1 and 2 are distinguished by hod. Here, wkndwday is still acting as the nuisance variable.

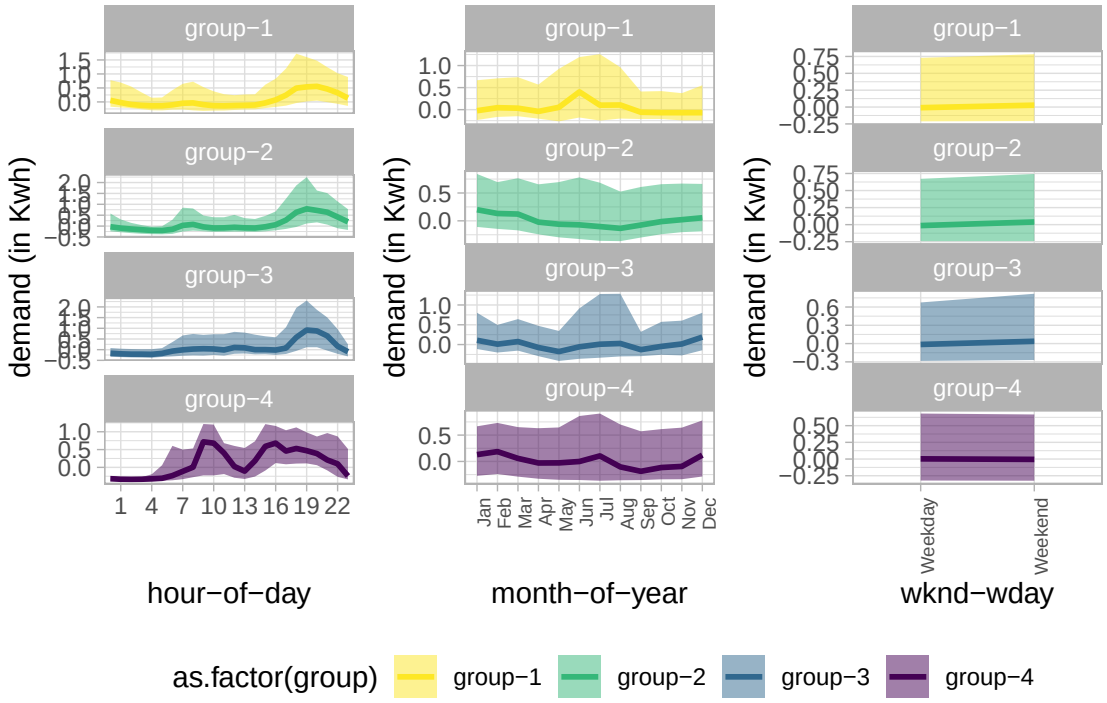

Figure 5: The distribution of electricity demand for the clusters across hour-of-day. The median is represented by a line and the shaded region represents the area between 25^{th} and 75^{th} percentile. Each cluster is characterised by unique shape across the granularity it is plotted against. For wknd-wday differences across different groups are not distinct suggesting that it might not be that important a variable to distinguish different clusters. This fact we will be re-established when we see the importance of each granularity throough the parallel coordinate plot.

stay-at-home parents is indicated by Group-4's almost equivalent morning, afternoon and evening profile. All of this may be validated with further information about the customer. The second panel of Figure 3 shows that month-of-year qualitative clusters are not as distinguishable as hour-of-day. Group 2 is the most distinct and uses the most power during the summer, possibly owing to the use of air conditioners. Group 4 has a flat profile, indicating no significant month-to-month changes. Groups 1 and 3 have heaters on in the winter but consume less energy in the summer. Since gas is not available in all of NSW LGAs, it is possible that customers' heater usage is recorded in electricity rather than gas.

The third panel of Figure 3 shows that the wknd-wday groups exhibit no significant changes across clusters, indicating that they may be a nuisance variable for these consumers.

The plotting scales are not displayed since we want to emphasise comparable shapes rather than scales. A customer in the cluster may have low daily or total energy usage, but their behaviour may be quite similar to a customer with high usage. That places them in the same group.