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 20, 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

1 Introduction

Large spatio-temporal data sets, both from open and administrative sources, offer up a world of possibilities for research. One such data sets for Australia is the Smart Grid, Smart City (SGSC) project (2010–2014) available through Department of the Environment and Energy. The project provides half-hourly data of over 13,000 household electricity smart meters distributed unevenly from October 2011 to March 2014. Raw data of these asynchronous time series can quickly become overwhelming and hard to interpret, requiring summarizing the large number of customers into pockets of similar energy behavior. Electricity utilities can utilize the smart meter usage patterns to develop targeted tariffs for individual groups and alleviate the problem of volatility in production by capitalizing on the flexibility of consumers.

Larger data sets include greater uncertainty about customer behavior due to growing variety of customers. Households are distributed geographically, and have different demographic properties such as the existence of solar panels, central heating or air conditioning. Multiple temporal dependencies define the behavioral patterns, which vary from customer to customer. Some families, for example, use a dryer to dry their clothing, while others hang them to dry on a line. This may be reflected in their weekly profile. They may have monthly variations where some customers are more prone to use air conditioners or heaters than others despite the existence of comparable electrical equipment and being subjected to similar weather conditions. The variations in behavior may occur on a regular basis, with some consumers being night owls and others being morning larks. Day-off energy usage may vary depending on whether consumers choose to remain at home or engage in outdoor activities. These various causes of variations in energy behavior, ranging from age, lifestyle, and family composition to building characteristics, weather, presence of various electrical equipments and others, make the problem of effectively segmenting consumers into similar energy behavior a particularly intriguing one.

This problem becomes more difficult when there is no other information on the customers besides their time series of energy usage. It is likely that such additional information is not available in order to protect the customers' confidentiality. Furthermore, it is not guaranteed that energy providers would always update customer profiles, such as property

characteristics, in a timely manner whenever they change. So this study does not attempt to explain why consumption differs. Instead, the work looks at how much energy-use heterogeneity can be uncovered in smart meter data and what are some of the most typical electricity use patterns by simply using the time series.

This is similar to a stochastic approach (Motlagh et al. (2019)) to clustering, which proposes interpreting electricity demand as a random process and extracting time-series characteristics, or a model of the series, to enable unsupervised clustering. Unsupervised clustering is only as good as the features that are extracted/selected or the distance metrics that were utilized. Well-designed additional features may collect characteristics that default features cannot. Based on the underlying structure of the temporal data, this article offers new distance metric and features for clustering and applies them to actual smart-meter data. Firstly, the distance metric is based on probability distribution, which in our knowledge is the first attempt to cluster smart meter data using probability distributions. These recorded time series are asynchronous, with varying time lengths for different houses and missing observations. Taking probability distributions helps to deal with such data, while helping with dimension reduction in one hand but not losing too much information due to aggregation. Secondly, we recognise that most clustering algorithms only provide hourly energy profiles during the day, but this approach provides a wider approach to the issue, seeking to group consumers with similar shapes over all important cyclic granularities. Since cyclic granularities are considered instead of linear granularities, clustering would group customers that have similar repetitive behavior across more than one cyclic granularities across which patterns are expected to be significant.

Related work

A typical clustering technique includes the following steps: (a) establishing distance (dissimilarity) and similarity through feature or model extraction and selection; and (b) selecting the clustering algorithm design. Distance measures could be time-domain based or frequency-domain (fast Fourier transform). A model-based clustering works by transforming the series into other other objects such as structure or set of parameters which can be more easily characterised and clustered (Motlagh et al. (2019)). (Chicco & Akilimali 2010) addresses information theory-based clustering such as Shannon or Renyi entropy and

its variations. The essential temporal characteristics of the curves are defined or extracted using feature-based clustering. Tureczek & Nielsen (2017) conducted a systematic study of over 2100 peer-reviewed papers on smart meter data analytics. None of the 34 articles chosen for their emphasis use Australian smart meter data. The most often used algorithm is K-Means. Using K-Means without considering time series structure or correlation results in inefficient clusters. Principal Component Analysis (PCA) or Self-Organizing Maps (SOM) eliminate correlation patterns and decrease feature space, but lose interpretability. To reduce dimensionality, several studies use principal component analysis or factor analysis to pre-process smart-meter data before clustering (Ndiaye & Gabriel (2011)). Other algorithms utilised in the literature include k-means variants, hierarchical approaches, and greedy k-medoids. Time series data, such as smart metre data, are not well-suited to any of the techniques mentioned in Tureczek & Nielsen (2017). Only one study (Ozawa et al. 2016) identified time series characteristics using Fourier transformation, which converts data from time to frequency and then uses K-Means to cluster by greatest frequency .

The remainder of the paper is organized as follows: Section 2 provides the clustering methodology introducing the features and distance metrics. Section 3 shows data designs to validate our methods and draw comparisons against several methods. Section 4 discusses the application of the method to a subset of the real data. Finally, we summarize our results and discuss possible future directions in Section 5.

2 Clustering methodology

The proposed methodology aim to leverage the intrinsic temporal data structure hidden in time series data. The foundation of our method is unsupervised clustering algorithms based exclusively on the time-series data. The similarity measure is the most essential ingredient of time series clustering. The (dis) similarity measure in this paper focuses on looking at the (dis) similarity between underlying distributions that may have resulted in different patterns across different cyclic temporal granularities. It is worth noting that when studying these similarities, a variety of objectives may be pursued. One objective could be to group time series with similar shapes over all relevant cyclic granularities. In this scenario, the variation in customers within each group is in magnitude rather than

shape, while the variation between groups is only in shape. There are distance measures are used for shape-based clustering [Ding et al. 2008; Wang et al. 2013] and many more but none of them look at the probability distributions while computing similarity. Moreover, most distance measures offer similar shape across just one dimension. For example, we often see "similar" daily energy profiles across hours of the day, but we suggest a broader approach to the problem, aiming to group consumers with similar distributional shape across all significant cyclic granularities. Another purpose of clustering could be to group customers that have similar differences in patterns across all major cyclic granularities, capturing similar jumps across categories regardless of the overall shape. For example, in the first goal, similar shapes across hours of the day will be grouped together, resulting in customers with similar behaviour across all hours of the day, whereas in the second goal, any similar big-enough jumps across hours of the day will be clubbed together, regardless of which hour of the day it is. Both of these objectives may be useful in a practical context and, depending on the data set, may or may not propose the same customer classification. Depending on the goal of clustering, the distance metric for defining similarity would be different. These distance metrics could be fed into a clustering algorithm to break large data sets into subgroups that can then be analyzed separately. These clusters may be commonly associated with real-world data segmentation. However, since the data is unlabeled a priori, more information is required to corroborate this. This section presents the work flow of the methodology:

- *Data preparation*

Wang et al. (2020) introduced the tidy "tsibble" data structure to support exploration and modeling of temporal data comprising of an index, optional key(s), and measured variables. For each key variable, the raw smart meter data is a sequence that is indexed by time and comprises values of several measurement variables at each time point. This sequence, though, could be depicted in a variety of ways. A shuffling of the raw sequence could reflect the distribution of hourly consumption over a single day, while another could indicate consumption over a week or a year. These temporal deconstructions of a time period into units such as hour-of-day, work-day/weekend are called cyclic temporal granularities. All cyclic granularities can be expressed in terms of the index set and could be

augmented with the initial tsibble structure (index, key, measurements). It is worthwhile to note that the data structure changes while transporting from linear to cyclic scale of time as multiple observations of the measured variable would correspond to each category of the cyclic granularities. In this paper, quantiles are chosen to characterize the distributions for each category of the cyclic granularity. So, each category of a cyclic granularity corresponds to a list of numbers which is essentially few chosen quantiles of the multiple observations.

- *Finding significant cyclic granularities or harmonies*

These cyclic granularities are useful for exploring repetitive patterns in time series data that get lost in the linear representation of time. It is advantageous to consider only those cyclic granularities across which there is a significant repetitive pattern for the majority of customers or noteworthy in an electricity-behavior context. In that case, when the customers are grouped, we can expect to observe some interesting patterns across the categories of the cyclic granularities considered. (Gupta et al. 2021) proposes a way to select significant cyclic granularities and harmonies which is used for this paper.

- *Individual or combined categories of cyclic granularities as DGP*

The existing work on clustering probability distributions assumes we have an iid sample $f_1(v), \ldots, f_n(v)$, where $f_i(v)$ denotes the distribution from observation i over some random variable $v = \{v_t : t = 0, 1, 2, \ldots, T-1\}$ observed across T time points. In our work, we are using i as denoting a customer and the underlying variable as the electricity demand. So $f_i(v)$ is the distribution of household i and v is electricity demand. In this work, instead of considering the probability distributions of the linear time series, we assume that the measured variables across different categories of any cyclic granularity are from different data generating processes. Hence, we want to be able to cluster distributions of the form $f_{i,A,B\ldots,N_C}(v)$, where A, B represent the cyclic granularities under consideration such that $A = \{a_j : j = 1, 2, \ldots J\}$, $B = \{b_k : k = 1, 2, \ldots K\}$ and so on. We consider individual category of a cyclic granularity (A) or combination of categories for interaction of cyclic granularities (for e.g. $A * B$) to have a distribution. For example, let us consider we have two cyclic granularities of interest, $A = 0, 1, 2, \ldots, 23$ representing hour-of-day

and $B = \{Mon, Tue, Wed, \ldots, Sun\}$ representing day-of-week. Each customer i consist of a collection of probability distributions. In case individual granularities (A or B) are considered, there are $J = 24$ distributions of the form $f_{i,j}(v)$ or $K = 7$ distributions of the form $f_{i,k}(v)$ for each customer i. In case of interaction, $J * K = 168$ distributions of the form $f_{i,j,k}(v)$ could be conceived for each customer i.
As a result, a distance between collections of these univariate probability distributions is required. Depending on the objective of the problem, there could be many approaches to considering such distances. This paper considers two approaches, which are explained in the next segment.

- *Distance metrics*

Considering each individual or combined categories of cyclic granularities as a data generating process lead to a collection of conditional distributions for each customer i. The (dis) similarity between each pair of customers should be obtained by combining the distances between these collections of conditional distributions such that the resulting metric is a distance metric, which could be fed into the clustering algorithm. Two types of distance metric is considered:

JS-based distances

This distance matrix considers two objects to be similar if every category of an individual cyclic granularity or combination of categories for interacting cyclic granularities have similar distributions. In this study, the distribution for each category is characterized using deciles and the distances between distributions are computed by using the Jensen-Shannon distance, which is symmetric and hence could be used as a distance measure.
The total distance between two elements x and y is then defined as

$$S_{x,y}^{A} = \sum_{j} D_{x,y}(A)$$

(sum of distances between each category j of cyclic granularity A) or

$$S_{x,y}^{A*B} = \sum_{j} \sum_{k} D_{x,y}(A, B)$$

(sum of distances between each combination of categories (j, k) of the harmony (A, B). When combining distances from individual L cyclic granularities C_l with n_l levels,

$$S_{x,y} = \sum_l S_{x,y}^{C_l}/n_l$$

is used, which is also a distance metric being the sum of JS distances.

wpd-based distances

Compute weighted pairwise distances (*wpd*) for all considered granularities for all objects. *wpd* is designed to capture the maximum variation in the measured variable explained by an individual cyclic granularity or their interaction and is estimated by the maximum pairwise distances between consecutive categories normalized by appropriate parameters. A higher value of *wpd* indicates that some interesting pattern is expected, whereas a lower value would indicate otherwise.

Distance between elements is then taken as the euclidean distances between them with the granularities being the variables and *wpd* being the value under each variable. Since Euclidean distance is chosen, the observations with high values of features (*wpd* values) will be clustered together. The same holds true for observations with low values of features. Thus this distance matrix would be useful to group customers that have similar significance of patterns across different granularities.

- *Pre-processing steps*

Practically most problems will have a very skewed distribution, it is often helpful to bring them to a normal-like shape before clustering. Two data transformation techniques are employed for the JS-based methods and NQT is built-in transformation used for computation of *wpd*, which forms the basis of wpd-based distances.

Robust scaling Standardizing is a common scaling method that subtracts the mean from values and divides by the standard deviation, resulting in a conventional Gaussian probability distribution for an input variable (zero mean and unit variance). If the input variable includes outlier values, standardisation may become skewed or prejudiced. To address this, robust scaling methods could be utilized (value – median) / (p75 – p25)) which results in a variable with a zero mean and median, as well as a standard deviation of one, while the outliers are still there with the same relative connections to other values.

Normal-Quantile transform First as a data pre-processing step to make all assymetrical real world variables more symmetric, we perform a quantile-normal transform on the data. This makes sure that the CDF of the resulting variable in Gaussian. The original data is ranked in ascending order and the probabilities $P(Y <= y(i)) = i/(n+1)$ are attached to $y(i)$, in terms of their ranking order. A NQT based transformation is applied by computing from a standard normal distribution a variable $\eta(i)$, which corresponds to the same probability $P(\eta < \eta(i)) = i/n+1$. By doing this, the new variables $\eta(i)$ will be marginally distributed according to standard Normal, N(0,1). NQT will transform the positively and negatively skewed distribution to a similar bell-shaped. From the transformed distribution, it is difficult to understand that raw distribution was of which shape. Also, multimodality gets hidden or magnitude get reversed with NQT. But deciles from the distribution will move in a similar manner as the raw distribution and hence the final distance matrix seem to be unaffected. Hence, this could be used.

- *Clustering algorithm*

In the analysis of energy smart meter data, K-Means or hierarchical clustering are often employed. These are simple and effective techniques that work well in a range of scenarios. For clustering, both employ a distance measure, and the distance measure chosen has a major influence on the structure of the clusters. We employ agglomerative hierarchical clustering in conjunction with Ward's criteria (XXX reference). The pair of clusters with minimum between-cluster distance are are sequentially merged in this using this agglomerative algorithms. A good comprehensive list of algorithms can be found in @.Xu2015-ja. We can possibly employ any clustering method that supports the given distance metric as input.

- *Characterization of clusters*

Characterization of clusters both statistically and qualitatively is an important stage of a cluster analysis. A potential way is to look at the findings from all the groups in graphs, and enhance our qualitative descriptions of the groupings. Cook & Swayne (2007) provides several ways to characterize clusters.

(a) Parallel coordinate plot: Parallel coordinate plot (Wegman (1990)) are widely used to display high-dimensional and multivariate data, allowing visual grouping to detect patterns. In a Parallel Coordinates Plot, each variable has its own axis, which are all parallel. Each axis is connected by a series of lines. That is, each line is made up of connected points on each axis. The order of the axes can affect how the reader interprets the data. This is because adjacent variables are more easily perceived than non-adjacent variables and rearranging the axes can reveal patterns or correlations between variables. Scattered plots of the p variables are arranged in a scatterplot matrix. It's a neat way to show multiple relationships at once, and it allows us to compare all the plots at once.

(b) Scatterplot matrix: The scatter plot matrix (draftsman's plot) is a matrix that comprises pairwise scatter plots of the p variables.Pairwise scatter plots are excellent for determining relationships between variables and determining which factors have contributed the most to clustering.

(c) Plotting cluster statistics: For larger problems, parallel coordinate plots may become cluttered and difficult to read, therefore we may opt to display cluster statistics instead. (Dasu et al. (n.d.))

(d) MDS, PCA and t-SNE: While all of the techniques examine a matrix of distances or dissimilarities to give a representation of the data points in a reduced-dimension space, their goals are not the same. The principal component analysis (Johnson & Wichern 2002) attempts to retain data variance. Multidimensional scaling (Borg & Groenen (2005)) seeks to maintain the distances between pairs of data points, with an emphasis on pairings of distant points in the original space. t-SNE, on the other hand, is concerned with preserving neighbourhood data points. Close data points in high-dimensional space will be condensed in the t-SNE embeddings.

(e) Tour: A tour (**?**) is a collection of interpolated linear projections of multivariate data into a lower-dimensional space. The sequence is seen as a dynamic visualization, enabling the viewer to observe the shadows cast by the high-dimensional data in a lower-dimensional view.

Depending on the distance measure utilized for the study, the cluster characterization technique will differ. Clusters that utilize wpd-based distances are characterised using multi-dimensional scaling and parallel coordinate displays. For JS-based distances, the distribution across major granularities may be presented to ensure that the goal of similar shapes within clusters and distinct shapes across clusters is met. This technique may potentially make advantage of multi-dimensional scaling.

3 Validation

To validate the clustering approaches, we create data designs that replicate prototype behaviors that might be seen in electricity data contexts. We spiked several attributes in the data to see where one method works better than the other and where they might give us the same outcome or the effect of missing data on the proposed methods. Three circular granularities $g1$, $g2$ and $g3$ are considered with categories denoted by $g10, g11$, $g20, g21, g22$ and $g30, g31, g32, g33, g34$ and levels $l_{g_1} = 2$, $l_{g_2} = 3$ and $l_{g_3} = 5$. These categories could be integers or some more meaningful labels. For example, the granularity "day-of-week" could be either represented by $0, 1, 2, \ldots, 6$ or $Mon, Tue, \ldots, Sun$. Here categories of $g1$, $g2$ and $g3$ are represented by $\{0, 1\}$, $\{0, 1, 2\}$ and $\{0, 1, 2, 3, 4\}$ respectively. A continuous measured variable v of length T indexed by $\{0, 1, \ldots T-1\}$ is simulated such that it follows the structure across $g1$, $g2$ and $g3$. We created independent replications $R = \{25, 250, 500\}$ of all data designs to see if our proposed clustering approaches can detect distinct designs in various groups for small, medium and large number of series. A sample size of $T = \{300, 1000, 5000\}$ is used in all designs to test small, medium and large sized series. The methods could preform differently with different jumps between consecutive categories. So a mean difference of $diff = \{1, 2, 5\}$ for corresponding categories are also considered. The performance of the methods can vary with different number of significant granularities. So scenarios with all, few and just one significant granularities are considered. The code for creating these designs and the detailed results can be found in the Supplementary section (link to github repo).

3.1 Data generating processes

Each category or combination of categories from $g1$, $g2$ and $g3$ are assumed to come from the same distribution, a subset of them from the same distribution, a subset of them from separate distributions, or all from different distributions, resulting in various data designs. As the methods ignore the linear progression of time, there is little value in adding time dependency in the data generating process. It is often reasonable to construct a time series using properties such as trend, seasonality, and auto-correlation. However, when examining distributions across categories of cyclic granularities, these time series features are lost or addressed independently by considering seasonal fluctuations through cyclic granularities. Because the time span during which an entity is observed in order to ascertain its behavior is not very long, the behavior of the entity will not change drastically and hence the time series can be assumed to remain stationary throughout the observation period. If the observation period is very long (for e.g more than 3 years), property, physical or geographical attributes might change leading to a non-stationary time series. But such a scenario is not considered here and the resulting clusters are assumed to be time invariant in the observation period. The data type is set to be "continuous," and the setup is assumed to be Gaussian. When the distribution of a granularity is "fixed", it means distributions across categories do not vary and are considered to be from N (0,1). The mean of different categories are altered in the "varying" designs, leading to varying distributions across categories.

3.2 Data designs

3.2.1 Individual granularities

Scenario (a): All signifiant granularities

Consider the scenario when all three granularities $g1$, $g2$, and $g3$ are responsible for distinguishing the designs. This implies that the patterns across each granularity will change significantly for at least one among the to-be-grouped designs. We consider different distributions across categories (as in Table 1 top) that will lead to different designs (as in Table 1 below). Figure 1 shows the linear and cyclic representation of the simulated variable under these five designs. As could be seen from the plot, it is impossible to decipher the struc-

Table 1: For Scenario (a), distributions of different categories (top), 5 designs resulting from different distributions across categories (below)

granularity	Varying distributions
g1	g10 ~ N(0, 1), g11 ~ N(2, 1)
g2	g21 ~ N(2, 1), g22 ~ N(1, 1), g23 ~ N(0, 1)
g3	g31 ~ N(0, 1), g32 ~ N(1, 1), g33 ~ N(2, 1), g34 ~ N(1, 1), g35 ~ N(0, 1)

design	g1	g2	g3
design-1	fixed	fixed	fixed
design-2	vary	fixed	fixed
design-3	fixed	vary	fixed
design-4	fixed	fixed	vary
design-5	vary	vary	vary

tural difference in the time series variable just by looking at the linear view. The difference in structure becomes quite clear when we see the distribution across cyclic granularities. Hence, for the consequent scenarios, only graphical displays across cyclic granularities are provided to emphasize the difference in structure.

Scenario (b): Few significant granularities

This is the case where one granularity will remain the same across all designs. We consider the case where the distribution of v would vary across levels of $g2$ for all designs, across levels of $g1$ for few designs and $g3$ does not change across designs. So $g3$ is not responsible for distinguishing across designs. Figure **??**(left) shows the considered design.

(c) One signifiant granularity

Here only one granularity is responsible for distinguishing the designs. Designs change significantly only for the granularity $g3$. Figure **??**(right) shows this.

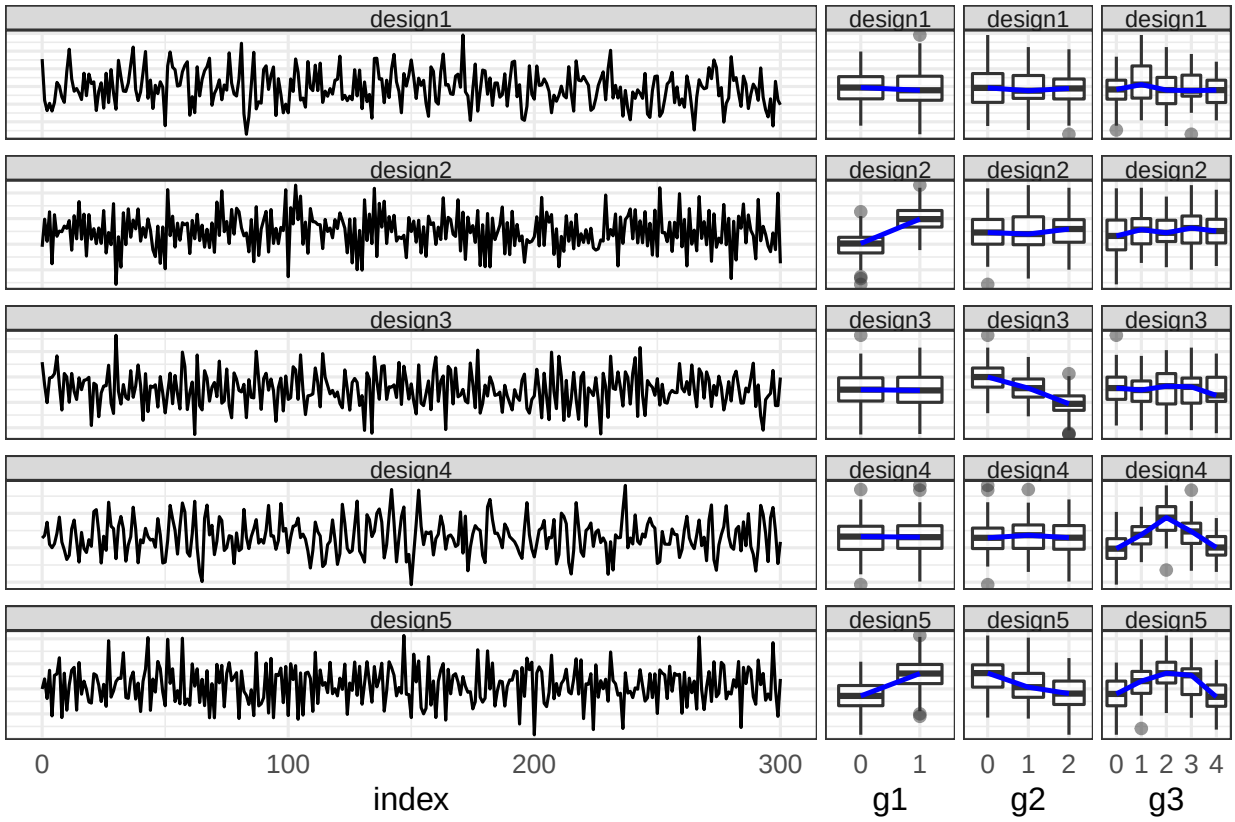

Figure 1: The linear (left) and cyclic (right) representation of the measured variable is shown. In this scenario, all of $g1$, $g2$ and $g3$ changes across at least one design. Also, it is not possible to comprehend these patterns across cyclic granularities or group similar series just by looking at the linear plots.

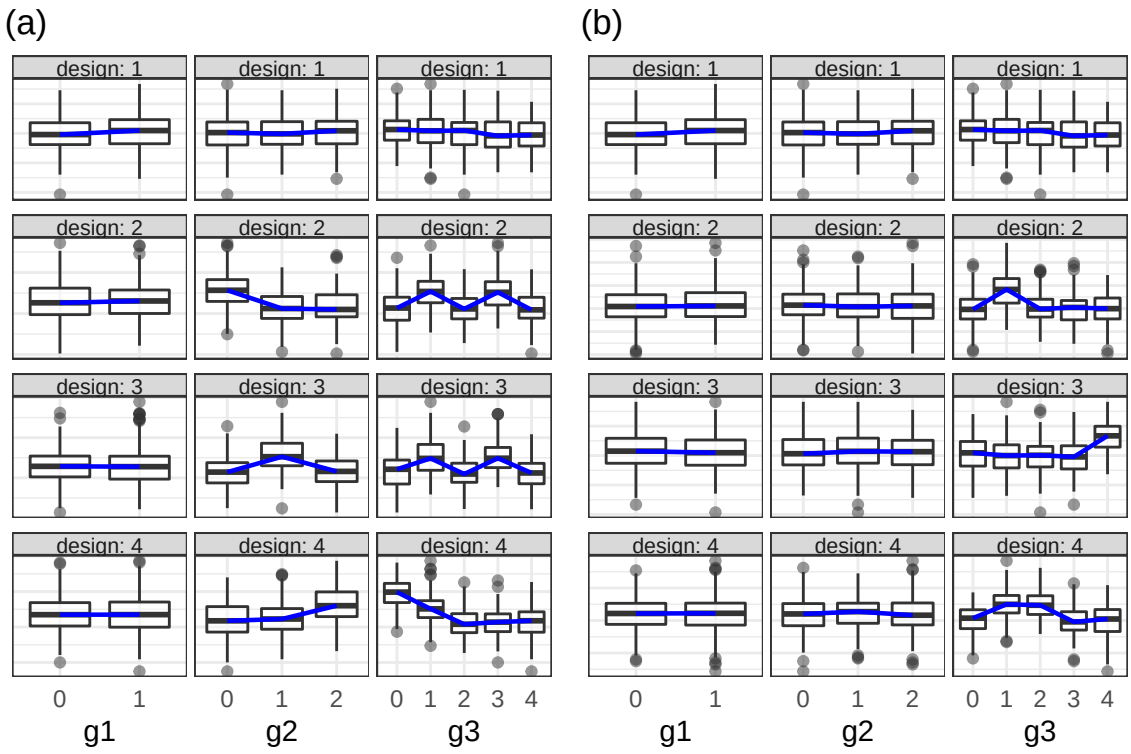

Figure 2: For the left scenario $g1$, $g2$ would change across atleast one design but $g3$ change remains same across all design. For the right one, only $g3$ changes across different designs.

3.2.2 Interaction of granularities

The proposed methods could be extended when two granularities of interest interact and we are interested to group subjects based on the interaction of the two granularities. For example, consider a group having a different weekday, weekend behavior in summer months, but not across winter. This type of joint behavior across granularities wknd-wday and month-of-year can be discovered by examining the distribution across combination of categories for different interacting granularities. Hence, in this scenario, we consider combination of categories to be generated from different distributions. For simplicity, consider a case with just two interacting granularities $g1$ and $g2$ of interest. As opposed to the last case, where we could examine distributions across $l_{g_1} + l_{g_2} = 5$ individual categories, with interaction, we need to examine the distribution of $l_{g_1} * l_{g_2} = 6$ combination of categories. Consider 4 designs in Figure 3 where different distributions are assumed for different combination of categories resulting in different designs. Design-1 has no change in distributions across $g1$ or $g2$, while Design-2 and Design-3 change across only $g1$ and $g2$ respectively. Design-4 changes across categories of both $g1$ and $g2$. Design-3 and Design-4 looks similar according to their relative difference between consecutive categories, but Design-4 also changes across facets, unlike Design-3 where all facets look the same.

3.3 Results

All the methods were fitted to each data designs and results are reported through confusion matrices. With increasing difference between categories, it gets easier for the methods to correctly distinguish the designs. For $mean_diff = 1$, the performances are pretty bad for js-robust methods and wpd method for lower nT. Although, with the kind of residential load datasets, a full year of load is the minimal requirement to capture expected variations in winter and summer profiles, for example. It is likely that nT would be at least 1000 with half-hourly data, even if data is only available just for a month. The performance is promising except when the number of observations for a customer is really small. For smaller difference between categories, it is expected that method js-nqt would preform better than the other two.

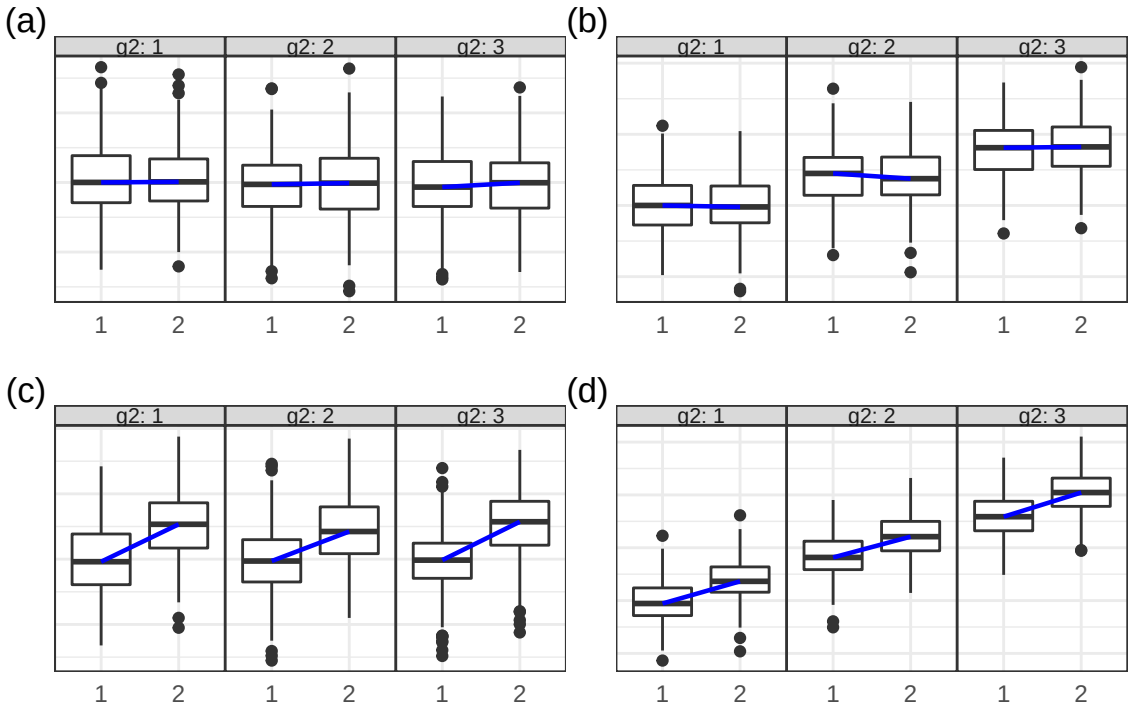

Figure 3: Design-1 (a) has no change in distributions across different categories of $g1$ or $g2$, while Design-2 (b) and Design-3 (c) change across only $g1$ and $g2$ respectively. Design-4 (d) changes across categories of both $g1$ and $g2$.

4 Application

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. It also provides demographic data for these customers most of which are missing and not utilized for the purpose of this paper. To maintain anonymity, the energy patterns could not be recognised at the person level, but rather by the geographical location of their dwelling and information about their Local Government Area.

In Figure 4, the time series of energy consumption is plotted along the y-axis against time from past to future for 50 sampled households. Each of these series correspond to a single customer. For each customer, the energy consumption is available at fine temporal resolution (every 30 minutes) for a long period of time (˜ 2 years) resulting in 27,000 (median) observations for each customer. Some customers' electricity use may be unavailable owing to power outages or improper recording, resulting in implied missing numbers in the database. For this data set it was found that out of 13,735 customers in total, 8,685 customers do not have any implicit missing observations, while the rest 5,050 customers had missing values. With further exploration, it was found that there is no structure in the missing-ness, that is missing observations can occur at any time point (see Appendix). Moreover, the data for these customers are characterized by unequal length, different start and end dates. Since our proposed methods consider probability distribution instead of raw data, both of these characteristics would not pose any threat to our methodology unless of course there is any structure or systematic patterns in them.
It can be expected that energy consumption vary substantially between customers, which is a reflection of their varied behavior owing to differences in profession, family size, geographical or physical characteristics. Since the linear time series plot has too many measurements all squeezed in this linear 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.

4.1 Prototype selection

In supervised learning, a training set containing previously known information is used to categorize new occurrences. Acceptable classification rates may be obtained by discarding instances which are not helpful for classification; this process is known as instance selection (Olvera-López et al. (2010)). This is similar to subsetting the population along all dimensions of importance such that the sampled data is representative of the main characteristics of the underlying distribution. Instance selection in unsupervised learning has received limited attention in the literature, but could serve as an useful way to sample evaluation data set to measure the performance of a model or method. One such procedure is suggested in Fan et al. (2021) that selects similar instances (neighbors) for each instance (anchor) and treats the anchor and its neighbors as the same class. In this section, a similar idea is used to select customers with prototype behaviors that serves as evaluation data sets for our proposed methodology.

First, we select the customers which do not have any implicit missing values and filtered their data for the year 2013. From this set, we randomly sample a set of 600 customers. We obtain the *wpd* for all cyclic granularities considered for these customers and found that `hod` (hour-of-day), `moy` (month-of-year) and `wkndwday` (weeknd/weekday) are coming out to be significant for most customers. This implies that for most customers, there is some interesting pattern across these three granularities. Potentially this could be done for the entire data set. This step is time consuming and hence has been only run for the 600 sampled customers. As a second step, we remove customers for which data in any category for the significant granularities are empty. For example, in this data set, if a customer do not have data for an entire month, they have been removed as their monthly behavior could not be studied in that case. Further, we also remove customers for which all the deciles of the energy consumption is zero. These are the customers whose consumption remain mostly flat and is expected to have no interesting repetitive patterns that is our interest of study. Finally, we are left with 356 customers. From this set we select 4 "anchor" customers which are far apart from each other and 5 neighboring customers for each of

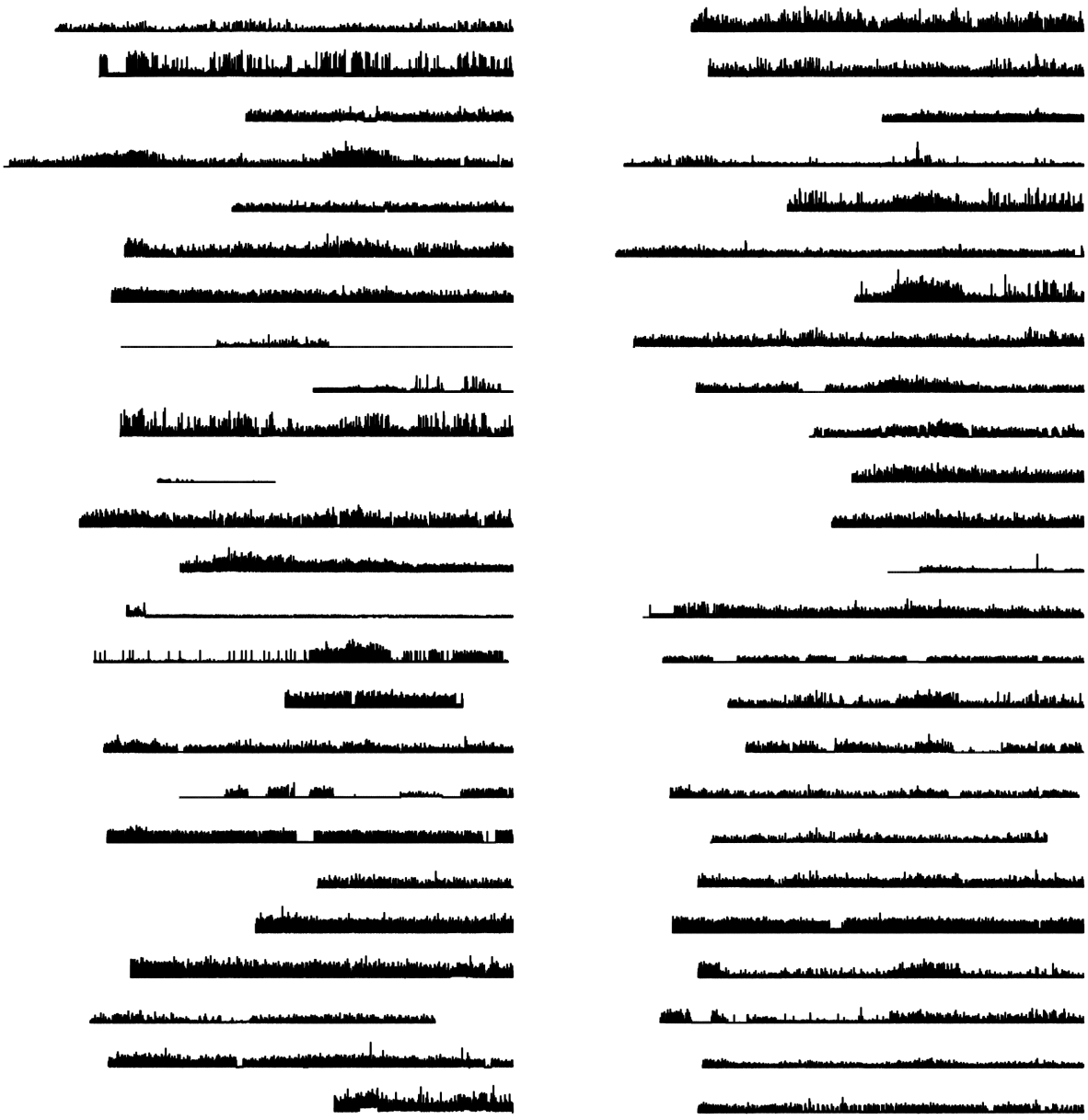

Figure 4: The raw data for 50 households are shown. It looks like there is a lot of missing values and unequal length of time series along with asynchronous periods for which data is observed. No insightful behavioral pattern could be discerned from this view other than when the customer is not at home.

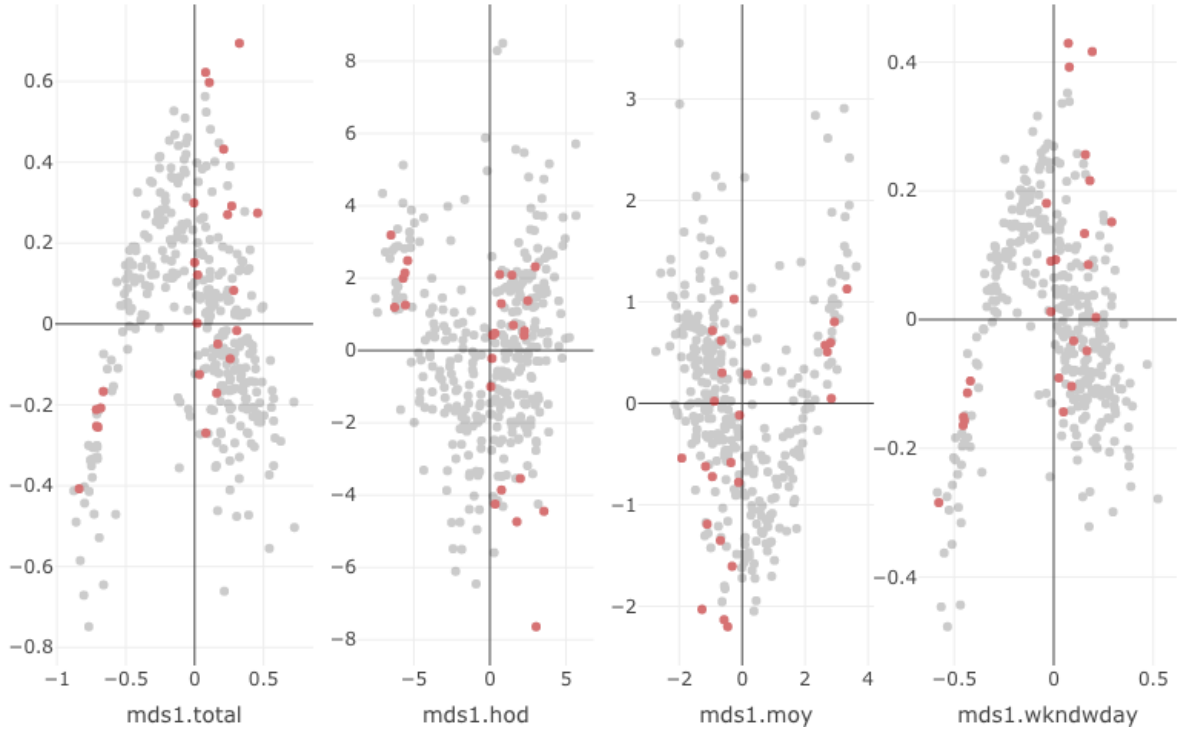

Figure 5: Instance selection with respect to 'hod' such that there are two distinct group and two overlapping groups.

these anchors. These selections were done using the granularity *hod* space. It is important to note that when we use our proposed methodologies, it is based on all dimensions `hod`, `moy` and `wkndwday`.Fig 5 shows the MDS of these 356 customers in a 2D space basis their distance on individual granularities and when all of them are combined. Our methodolgies are run on these 24 customers, which act as a way to evaluate the proposed methodologies.

4.2 JS-based clustering

The distribution of electricity demand for the selected 24 customers across hour-of-day is shown in 6. 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 color represent the same group. Except for customer id 8269176, there is unanimity across design

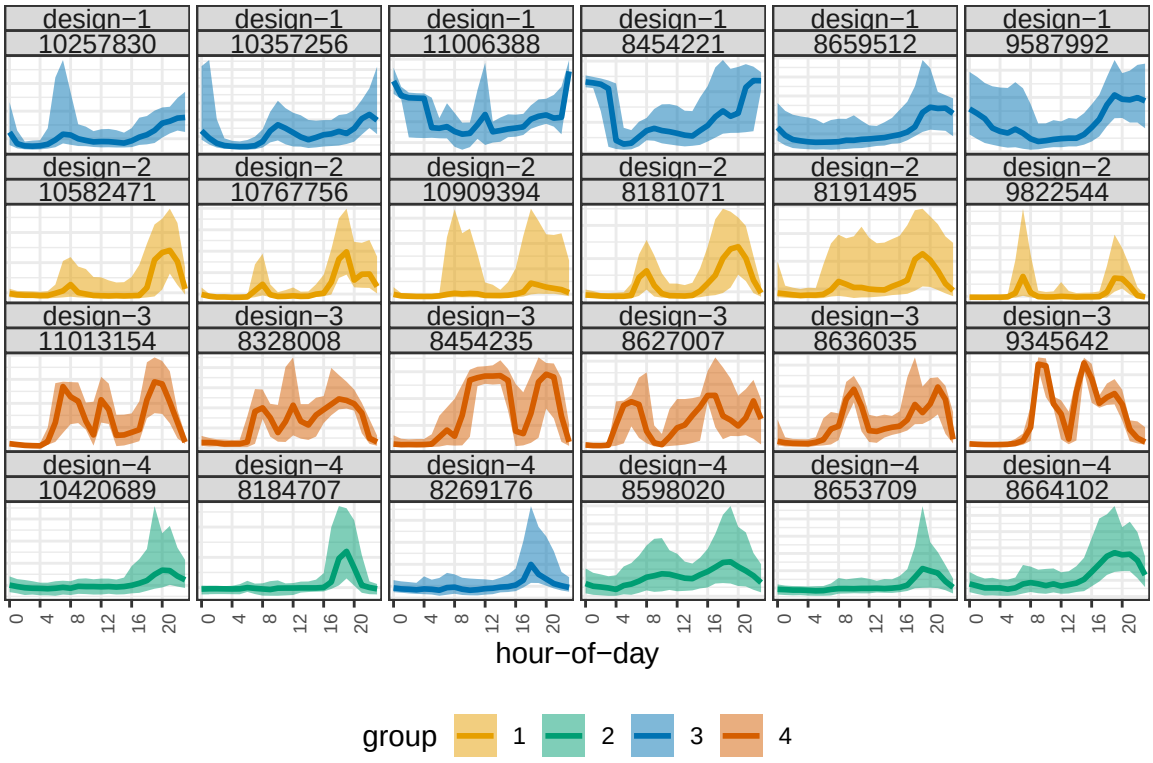

Figure 6: The distribution of electricity demand for the selected 24 customers across hour-of-day. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. According to our prototype selection method, each row represents a distinct shape of daily 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.

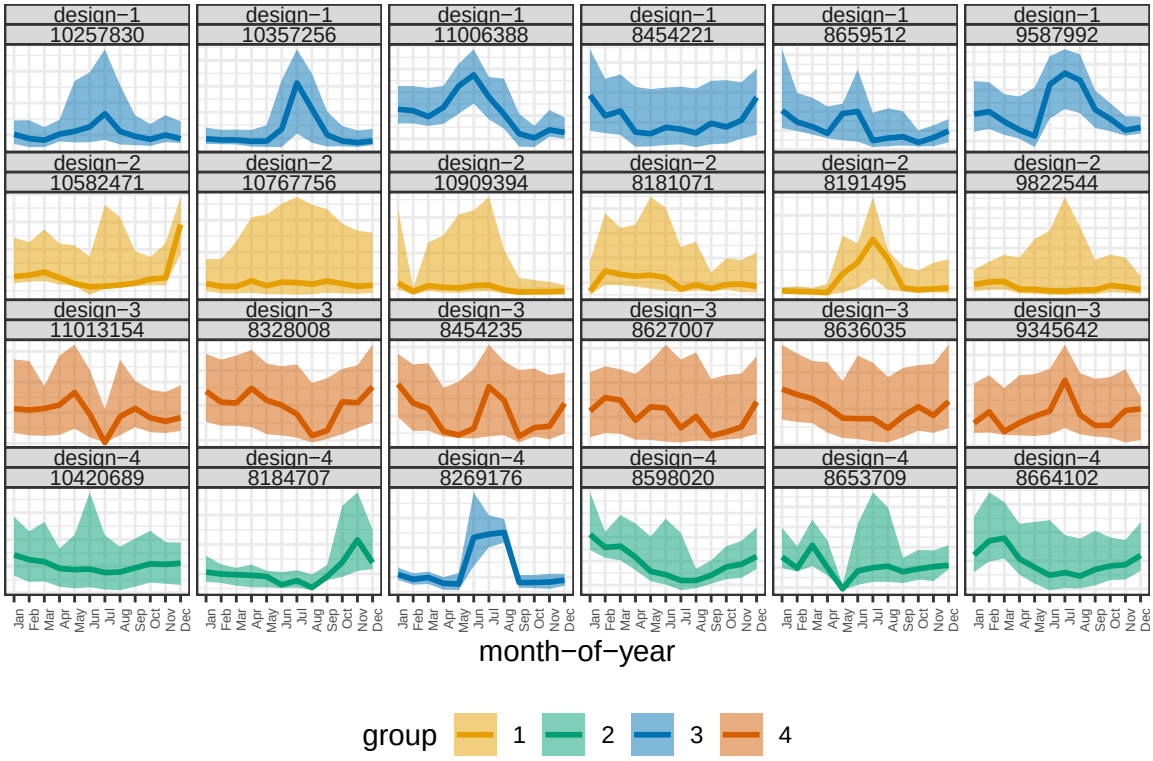

Figure 7: The distribution of electricity demand for the selected 24 customers across month-of-year. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. Intriguingly, customer id 8269176 appears to be in the appropriate group (Group 3) because 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.

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 7. 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 behavioral differences across customers within a group, which implies that `hod` has the minimum within-cluster variation.

Characterization of clusters both statistically and qualitatively is an important stage of a cluster analysis. A potential way is to look at the findings from all the groups in graphs, and enhance our qualitative descriptions of the groupings. Figure 8 shows the distribution of the summarized groups and help us to characterise each of the clusters. All of these may be validated with further information about the customer.

Group 1: This 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. These users in this cluster are insensitive to high temperature but have their heaters on in the winter months May-Aug.

Group 2: This 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. In addition, the users in this cluster are insensitive to high temperature, but sensitive to lower temperature.

Group 3: This group a strong early morning and late night hours. These consumers may be flexible students or elderly retirees who are night owls. They have heaters on in the winter but consume less energy in the summer. Winter usage may also be due to increased usage of heaters at night when they are up.

Group 4: Presence of children or stay-at-home parents is indicated by Group-4's almost

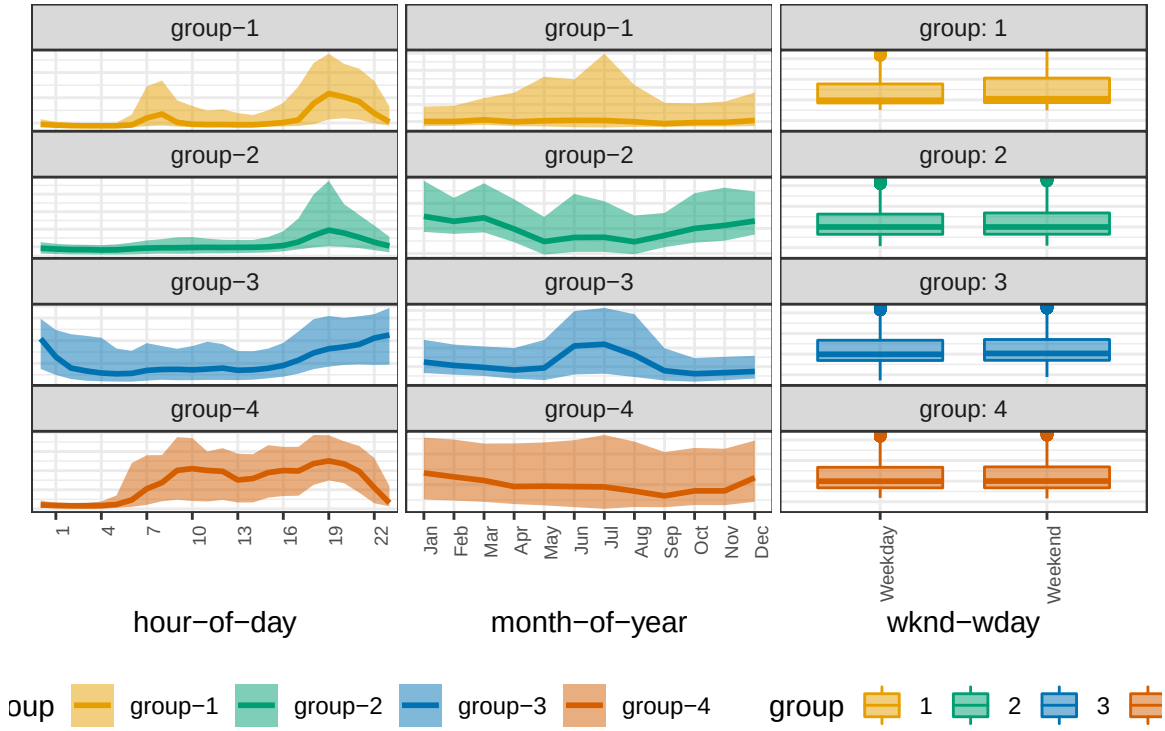

Figure 8: The distribution of electricity demand for the clusters across hour-of-day. The median is represented by a line and the shaded region represents the area between 25^{th} and 75^{th} percentile. Each cluster is characterised by unique shape across the granularity it is plotted against. For wknd-wday 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.

equivalent morning, afternoon and evening profile. They have a flat monthly profile, indicating the usage across all months are similar. It seems that the users in this group are more concerned about the comfort and quality of life than the cost of electricity as opposed to Group 1 and 3, who do not consume more electricity in the summer.

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 behavior may be quite similar to a customer with high usage. The third panel of Figure 8 shows that the wknd-wday groups exhibit no significant changes across clusters, indicating that they may be a nuisance variable for these consumers.

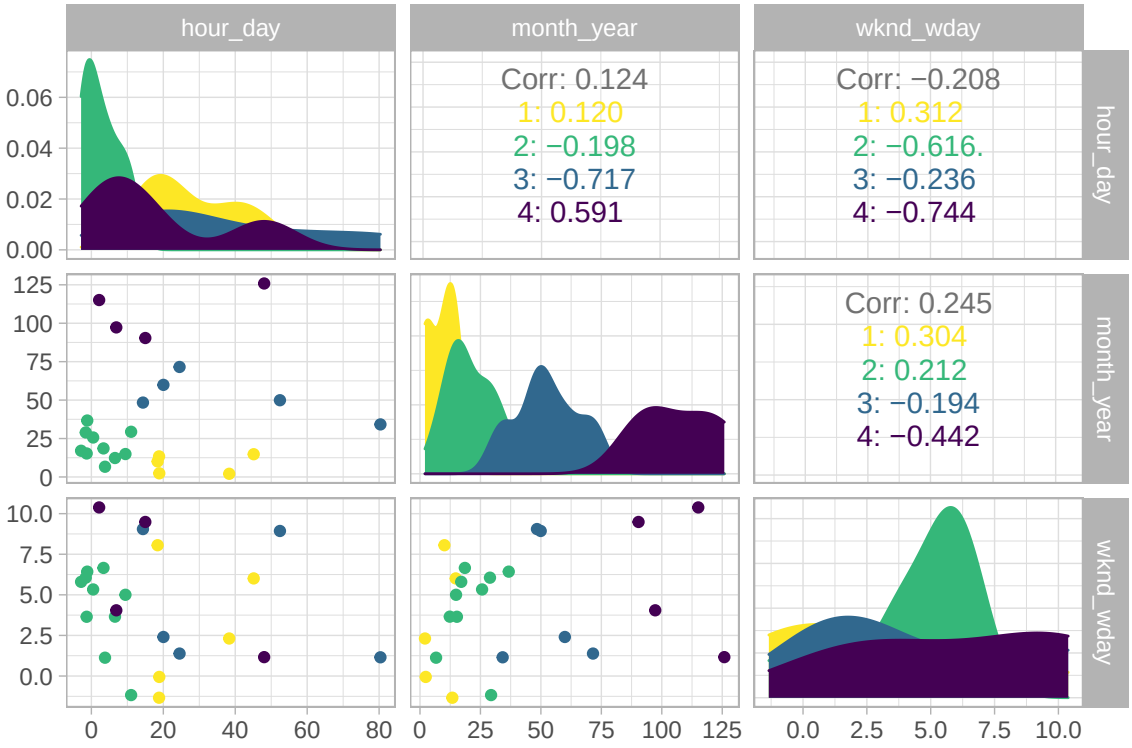

Figure 9: A ggpairsplot and parallel coordinate plot are used to depict each of the 24 customers. The ggpairs plot four distinct clusters across the month-of-year, which are less prominent across the hour-of-day and wknd-wday. The parallel coordinate plot ranks the variables in order of importance, indicating 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.

4.3 wpd-based clustering

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. The ggpairs plot four distinct clusters across the month-of-year, which are less prominent across the hour-of-day and wknd-wday. The parallel coordinate plot ranks the variables in order of importance, indicating that

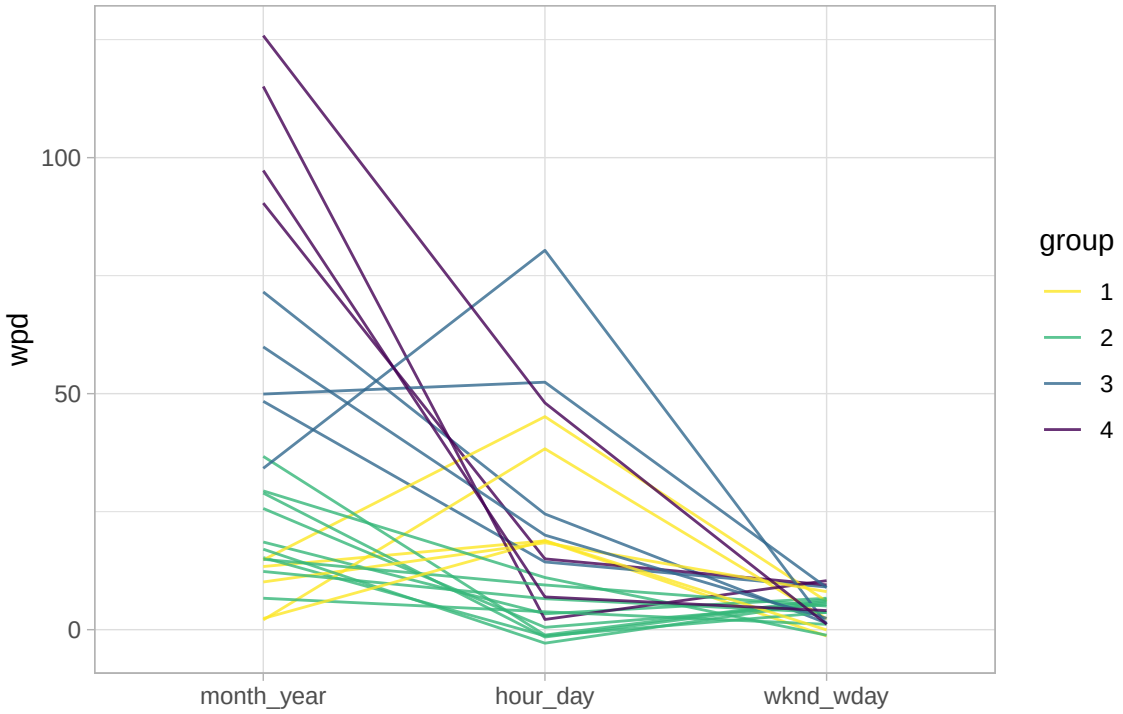

Figure 10: A parallel coordinate plot with the three significant cyclic granularities used for wpd-based clustering. The variables are sorted according to their separation across classes (rather than their overall variation between classes). This means that moy is the most important variable in distinguishing 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.

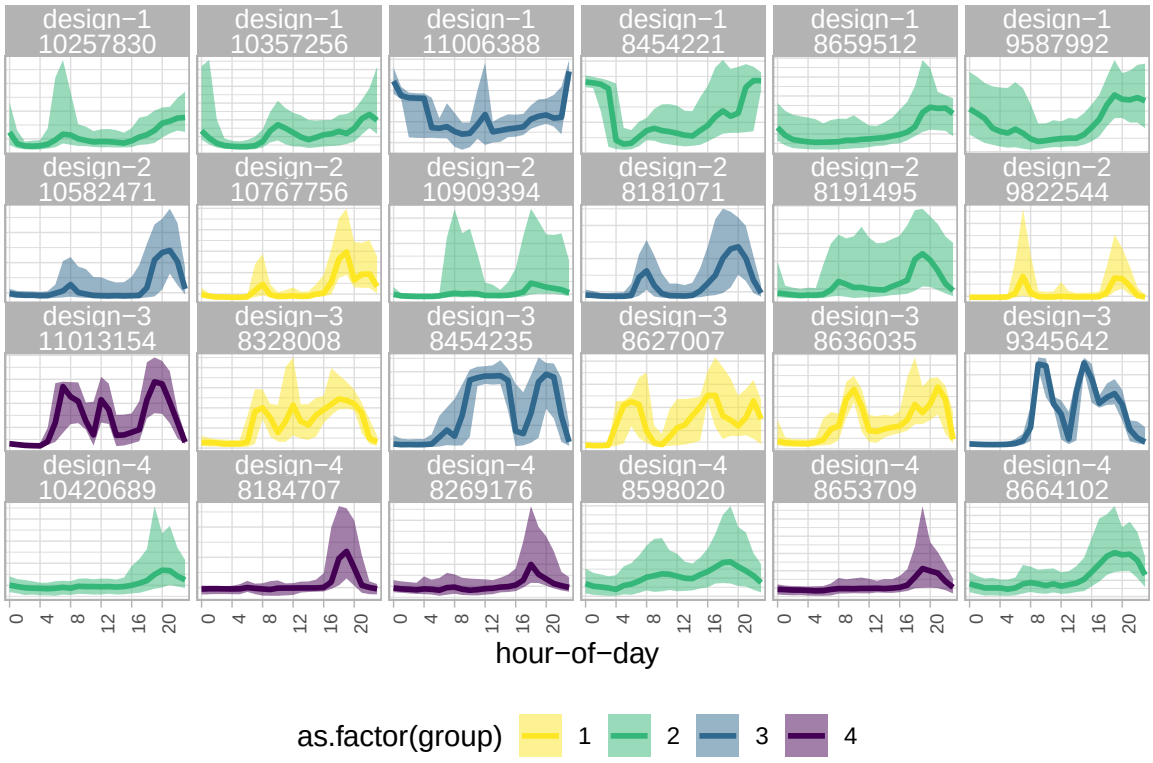

Figure 11: The distribution of electricity demand for the selected 24 customers across hour-of-day. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. According to our prototype selection method, each row represents a distinct shape of daily 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.

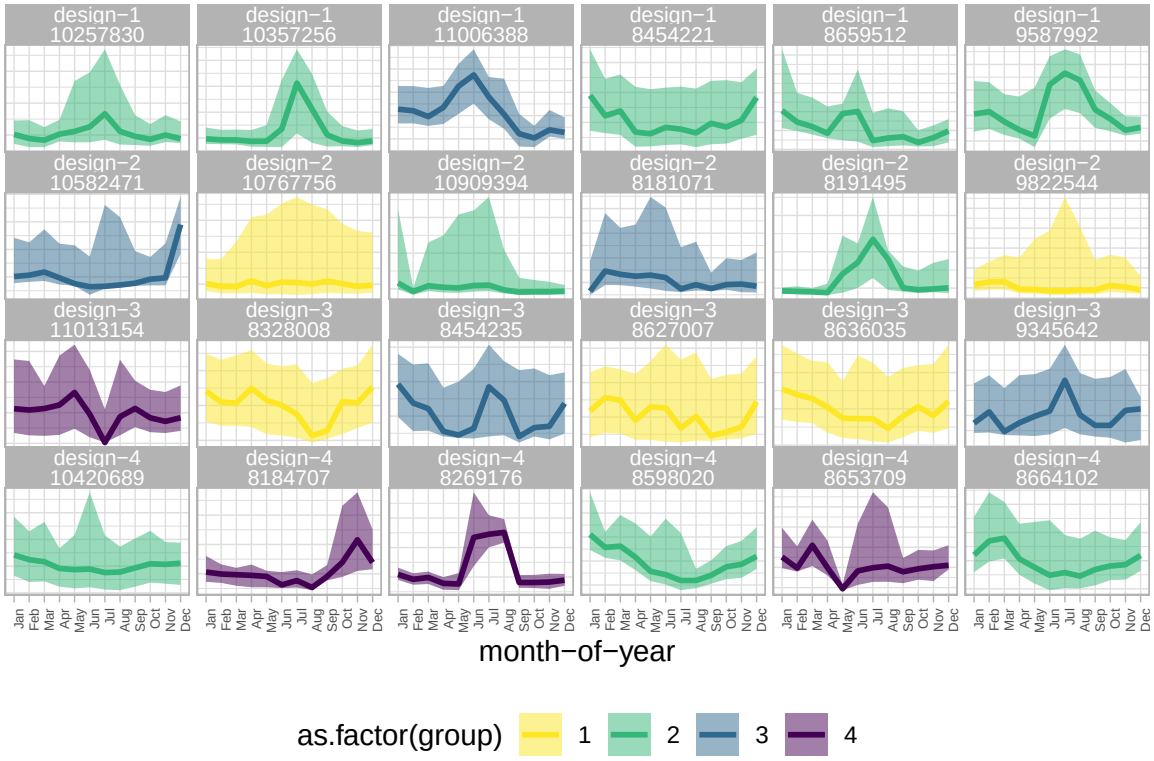

Figure 12: The distribution of electricity demand for the selected 24 customers across month-of-year. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. Intriguingly, customer id 8269176 appears to be in the appropriate group (Group 3) because 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.

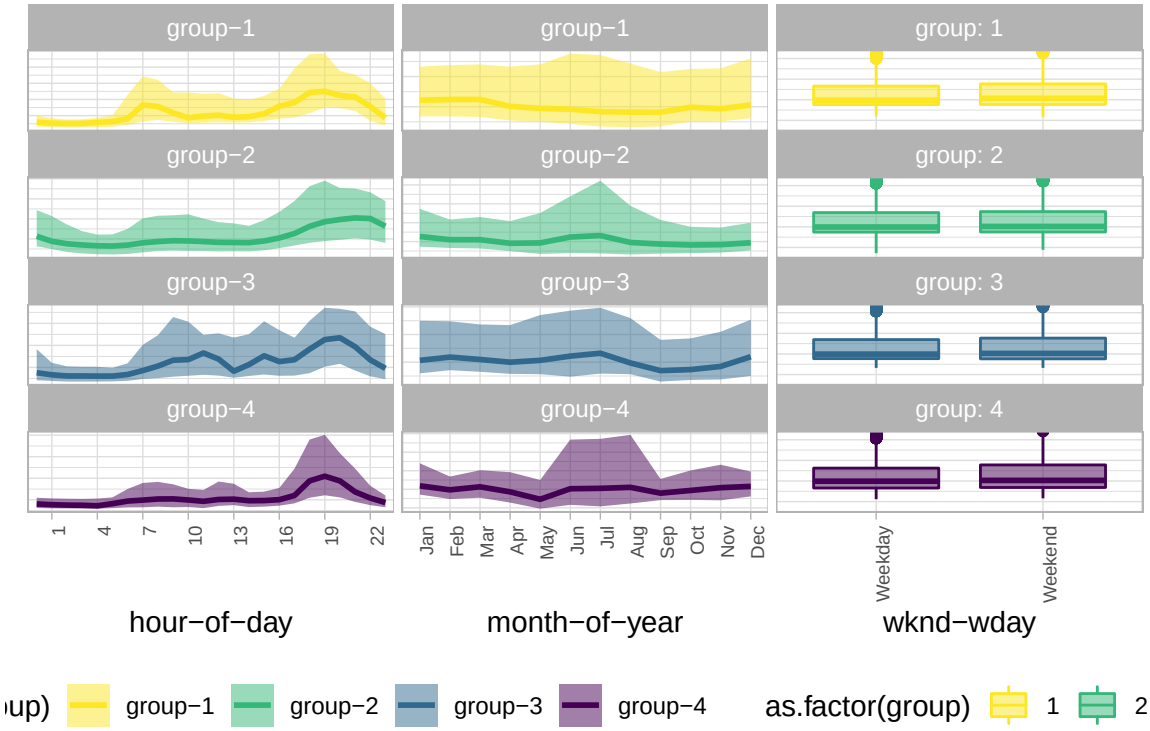

Figure 13: The distribution of electricity demand for the clusters across hour-of-day. The median is represented by a line and the shaded region represents the area between 25^{th} and 75^{th} percentile. Each cluster is characterised by unique shape across the granularity it is plotted against. For 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.

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.

4.4 Classifying the 356 customers

5 Discussion

We propose different clustering methodology for grouping noisy, patchy time series data available at a fine temporal scale. Depending on the aim of clustering, they produce different clustering. The clustering is done based on probability distributions of the time series variable measured across several cyclic granularirties. There is issue with scaling it up to many customers as anomalies need to be removed before such classification would be useful.

References

Borg, I. & Groenen, P. J. (2005), *Modern multidimensional scaling: Theory and applications*, Springer Science & Business Media.

Chicco, G. & Akilimali, J. S. (2010), 'Renyi entropy-based classification of daily electrical load patterns', *IET generation, transmission & distribution* **4**(6), 736–745.

Cook, D. & Swayne, D. F. (2007), *Interactive and Dynamic Graphics for Data Analysis: With R and Ggobi*, Springer, New York, NY.

Dasu, T., Swayne, D. F. & Poole, D. (n.d.), 'Grouping multivariate time series: A case study', `https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.7876&rep=rep1&type=pdf`. Accessed: 2021-10-20.

Fan, H., Liu, P., Xu, M. & Yang, Y. (2021), 'Unsupervised visual representation learning via Dual-Level progressive similar instance selection', *IEEE Trans Cybern* **PP**.

Gupta, S., Hyndman, R. J. & Cook, D. (2021), 'Detecting distributional differences between temporal granularities for exploratory time series analysis', *unpublished* .

Motlagh, O., Berry, A. & O'Neil, L. (2019), 'Clustering of residential electricity customers using load time series', *Appl. Energy* **237**, 11–24.

Ndiaye, D. & Gabriel, K. (2011), 'Principal component analysis of the electricity consumption in residential dwellings', *Energy Build.* **43**(2), 446–453.

Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F. & Kittler, J. (2010), 'A review of instance selection methods', *Artificial Intelligence Review* **34**(2), 133–143.

Ozawa, A., Furusato, R. & Yoshida, Y. (2016), 'Determining the relationship between a household's lifestyle and its electricity consumption in japan by analyzing measured electric load profiles', *Energy and Buildings* **119**, 200–210.

Tureczek, A. M. & Nielsen, P. S. (2017), 'Structured literature review of electricity consumption classification using smart meter data', *Energies* **10**(5), 584.

Wang, E., Cook, D. & Hyndman, R. J. (2020), 'A new tidy data structure to support exploration and modeling of temporal data', *Journal of Computational and Graphical Statistics* **29**(3), 466–478.

Wegman, E. J. (1990), 'Hyperdimensional data analysis using parallel coordinates', *Journal of the American Statistical Association* **85**(411), 664–675.