Clustering time series based on probability distributions across temporal granularities

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 28, 2021

Abstract

With more and more time series data being collected at much finer temporal resolution, for a longer length of time, and for a larger number of individuals/entities, time series clustering research is getting a lot of traction. The sort of noisy, patchy, uneven, and asynchronous time series that is typical in many disciplines limits similarity searches among these lengthy time series. In this work, we suggest a method for overcoming these constraints by grouping time series based on probability distributions over cyclic temporal granularities. Cyclic granularities are temporal deconstructions of a time period into units such as hour-of-the-day, work-day/weekend, and so on, and can be helpful for detecting repeating patterns. Looking at probability distributions across cyclic granularities results in an approach that is robust to missing or noisy data, aids in dimension reduction, and ensures small pockets of similar repeated behaviours. The proposed method was tested using a collection of residential electricity customers. The simulated and empirical evidence demonstrates that our method is capable of producing meaningful clusters.

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

*Email: Sayani.Gupta@monash.edu

1 Introduction

Time-series clustering is the process of unsupervised partitioning of n time-series data into k ($k < n$) groups such that homogeneous time-series are grouped together based on a certain similarity measure. The time-series features, length of time-series, representation technique, and, of course, the purpose of clustering time-series all influence the suitable similarity measure or distance strategy to a meaningful level. The three primary methods to time series clustering (Liao (2005)) are algorithms that operate directly with distances or raw data points in the time or frequency domain (distance-based), with features derived from raw data (feature-based), or indirectly with models constructed from raw data (model-based) (model-based). The efficacy of distance-based techniques is highly dependent on the distance measure utilised. Defining an appropriate distance measure for the raw time series may be a difficult task since it must take into account noise, variable lengths of time series, asynchronous time series, different scales, and missing data. Commonly used Distance-based similarity measures as suggested by a review of time series clustering approaches (Aghabozorgi et al. (2015)) are Euclidean, Pearson's correlation coefficient and related distances, Dynamic Time Warping, Autocorrelation, Short time series distance, Piecewise regularisation, cross-correlation between time series, or a symmetric version of the Kullback–Liebler distances (Liao (2007)). Euclidean distance and DTW are often used in time series clustering. When it comes to time-series clustering accuracy, the Euclidean distance beats DTW, but DTW has its own advantages (Corradini (2001)).Euclidean distance requires time series of equal length. while DTW can assist cluster time series of varying lengths (Ratanamahatana & Keogh (2005)), only if there are no missing observations.

We are motivated by the residential smart meter data. These long time series are asynchronous, with varying time lengths for different houses and missing observations and characterised by noisy and patchy behavior that can quickly become overwhelming and hard to interpret, requiring summarizing the large number of customers into pockets of similar energy behavior. Choosing probability distributions instead of raw data seems to be a natural way to analyze these types of data sets. Hence this paper proposes a distance metric based on Jensen-Shannon distances between probability distributions across significant

cyclic granularities. Cyclic temporal granularities, which are temporal deconstructions of a time period into units such as hour-of-the-day, work-day/weekend, can be useful for measuring repetitive patterns in large univariate time series data. Since cyclic granularities are considered instead of linear granularities, the resulting clusters are expected to group customers that have similar repetitive behaviors. Below are some of the benefits of our method, which will be detailed in further depth in subsequent sections.

- Some clustering algorithms become problematic with the very high dimensionality of the time series resulting from the frequency at which they are recorded and the length of time for which they are observed. We can efficiently cluster long length time series by reducing dimensionality by characterising through probability distributions;
- By utilising Jensen-Shannon distances, we are evaluating the distance between two distributions rather than raw data, which is less susceptible to missing observations and outliers compared to other traditional distance measures;
- While most clustering algorithms produce clusters similar across just one temporal granularity, this technique takes a broader approach to the problem, attempting to group observations with similar forms across all key cyclic granularities. Because cyclic granularities are used rather than linear granularities, clustering would group consumers who exhibit similar repeating behaviour over many cyclic granularities where patterns are predicted to be important.
- It is reasonable to define a time series based on its degree of trend and seasonality and to take these characteristics into account while clustering it. The change in data structure by considering probability distributions across cyclic granularities ensures there is no trend and seasonal fluctuations are handled separately. Thus there is no need to de-trend or de-seasonalize the data prior to performing the clustering method. For similar reasons, there is no need to exclude holiday or weekend routines.

Background and motivation

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. . Larger data sets include greater uncertainty about customer behavior due to growing variety of customers. Households vary in size, location, and amenities such as solar panels, central heating, and air conditioning. The behavioural patterns differ amongst customers due to many temporal dependencies. Some households use a dryer, while others dry their clothes on a line. Their weekly profile may reflect this. They may vary monthly, with some customers using more air conditioners or heaters than others, while having equivalent electrical equipment and weather circumstances. Some customers are night owls, while others are morning larks. Day-off energy use varies depending on whether customers stay home or go outside. Age, lifestyle, family composition, building attributes, weather, availability of diverse electrical equipment, among other factors, make the task of properly segmenting customers into comparable energy behaviour a fascinating one. This challenge is worsened when all we know about our consumers is their energy use history (Ushakova & Jankin Mikhaylov (2020)). To safeguard the customers' privacy, it is probable that such information is not accessible. Also, energy suppliers may not always update client information, such as property features, in a timely manner. Thus, there is a growing need to have research that examines how much energy usage heterogeneity can be found in smart meter data and what are some of the most common power consumption patterns, rather than explaining why consumption differs.

Related work

A multitude of papers have emerged around smart meter time series clustering for deepening our knowledge of consumption patterns. 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 . Motlagh et al. (2019) suggests that the time feature extraction islimited by the type of noisy, patchy, and unequal time-series common in residential datasets and addresses model-based clustering by transforming the series into other other objects such as structure or set of parameters which can be more easily characterised and clustered. (Chicco & Akilimali 2010) addresses information theory-based clustering such as Shannon or Renyi entropy and its variations. Melnykov (2013) disucusses how outliers, noisy observations and scattered observations can complicate estimating mixture model parameters and hence the partitions.

Given the limitations of the similarity measures in dealing with large volumes of this complicated time series data, we present a similarity measure based on probability distributions that seems to be a more organic option for coping with time series data with aforementioned characteristics. 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 (Wickham et al. (2011)) 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.

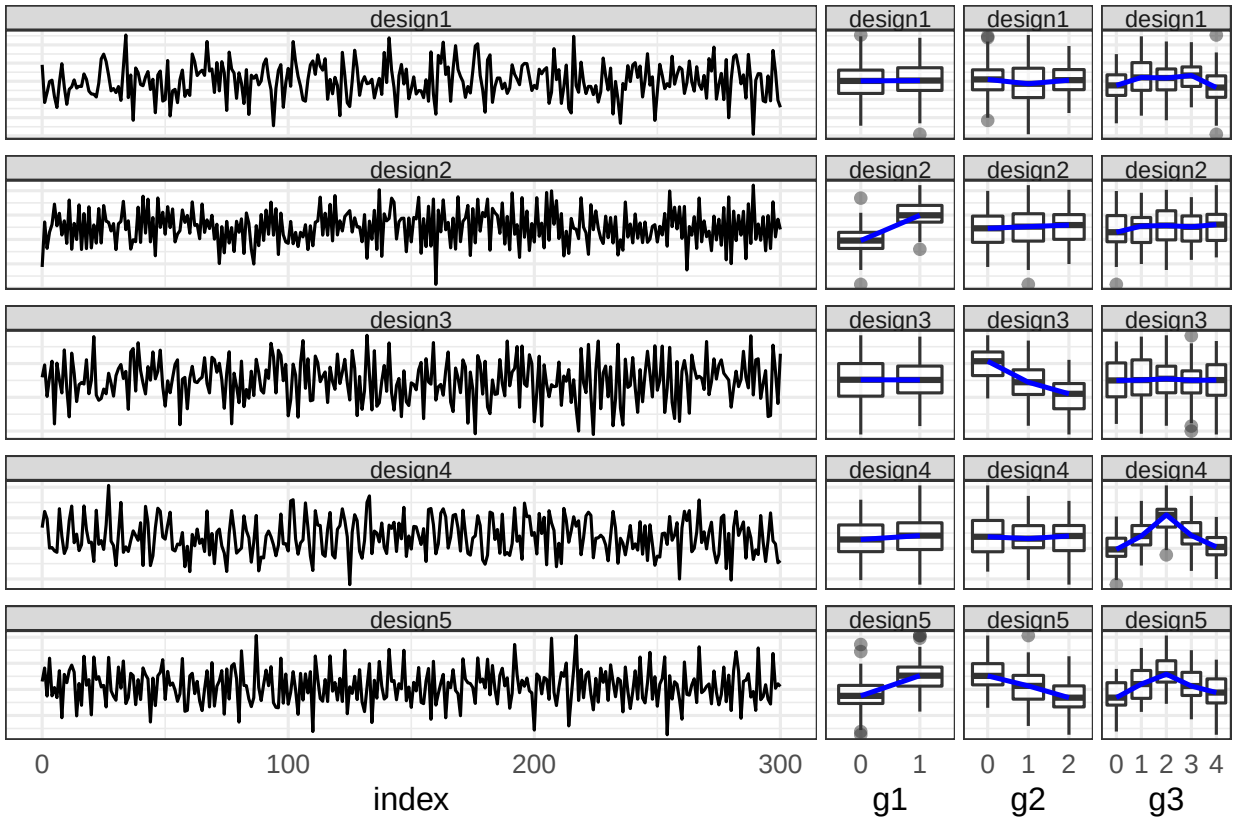


Figure 1: The linear (left) and cyclic (right) representation of the measured variable is shown. In this scenario, all of $g1$, $g2$ and $g3$ changes across at least one design. Also, it is not possible to comprehend these patterns across cyclic granularities or group similar series just by looking at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

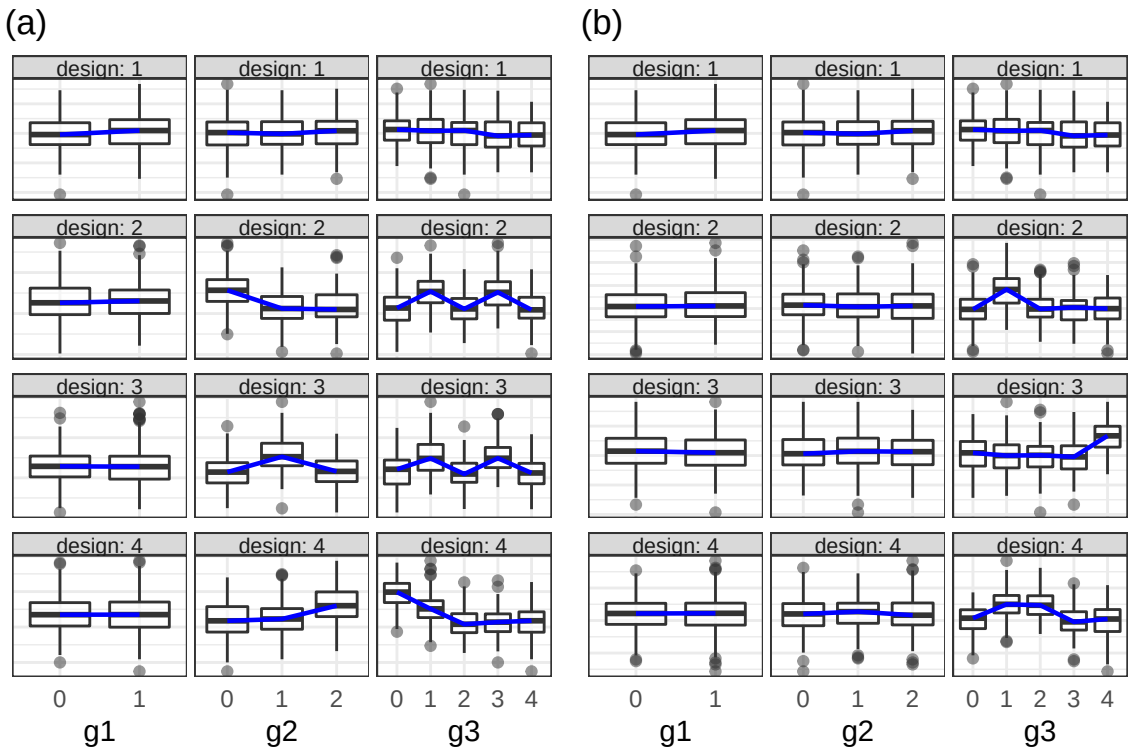


Figure 2: For the left scenario $g1$, $g2$ would change across atleast one design but $g3$ change remains same across all design. For the right one, only $g3$ changes across different designs.

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.

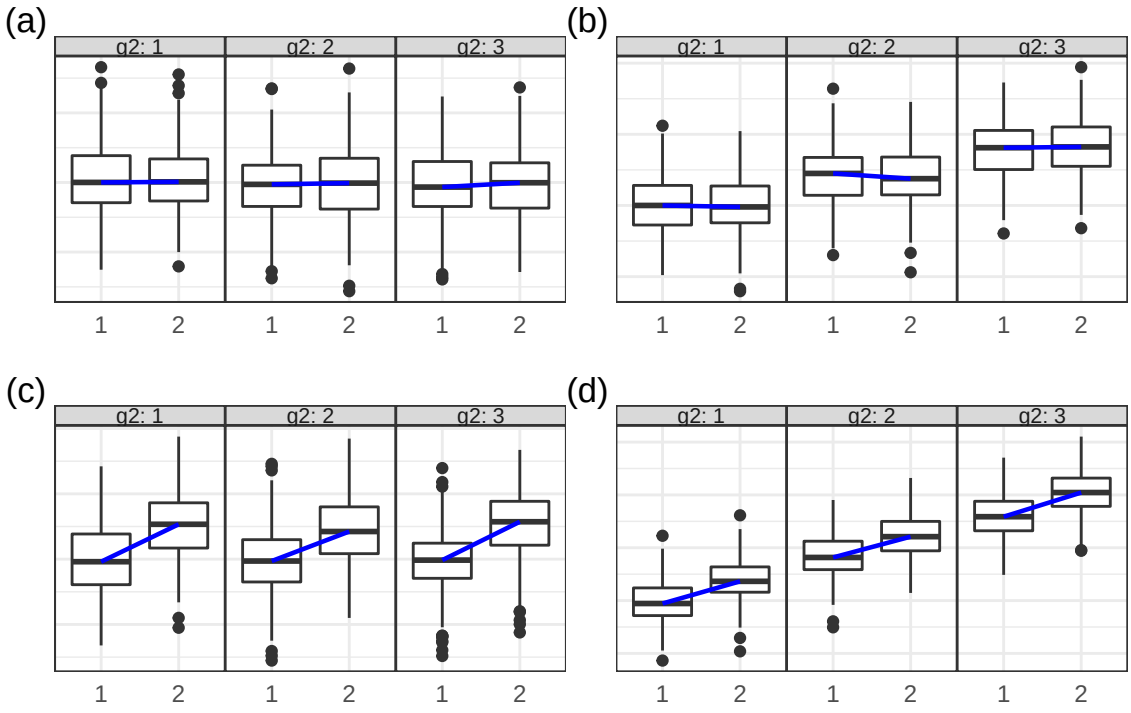


Figure 3: Design-1 (a) has no change in distributions across different categories of $g1$ or $g2$, while Design-2 (b) and Design-3 (c) change across only $g1$ and $g2$ respectively. Design-4 (d) changes across categories of both $g1$ and $g2$.

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.

Pre-processing steps

P1. We randomly select a sample of 600 customers which do not have any implicit missing values and filtered their data for the year 2013.

P2. 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. This implies that for most customers, there is some interesting pattern across these three granularities.

P3. Remove customers from this list 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.

P4. 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 which we run our prototype search. There

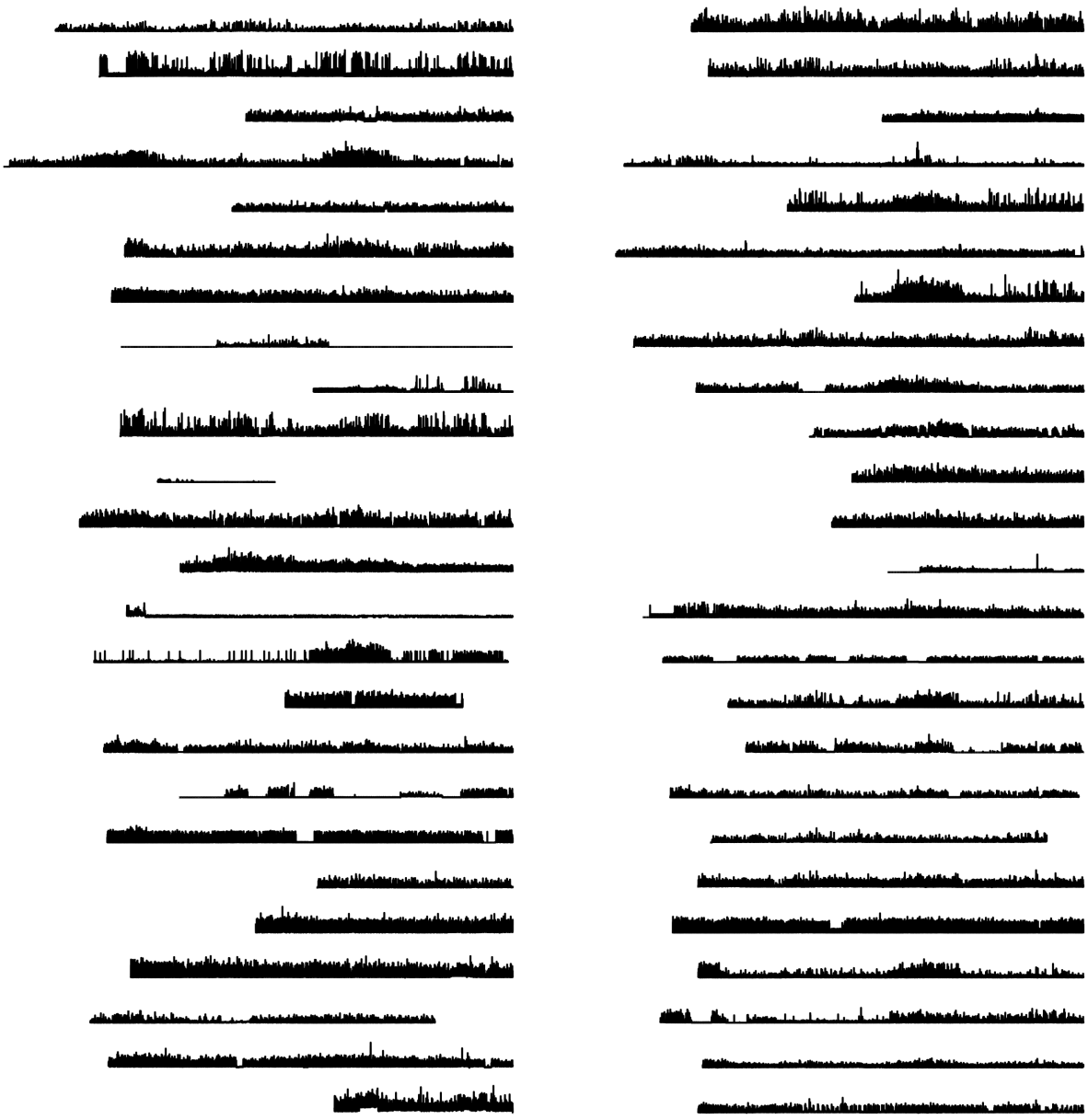


Figure 4: The raw data for 50 households are shown. It looks like there is a lot of missing values and unequal length of time series along with asynchronous periods for which data is observed. No insightful behavioral pattern could be discerned from this view other than when the customer is not at home.

are different methods of going ahead with this. For example, one approach could be to use any Non-linear dimensionality reduction technique like MDS or PCA and project the data in a 2-dimensional space. One can then look at few "anchor" customers which are far apart in the 2D space and pick few neighbors for each of the anchor customers. Paradoxically, the curse of dimensionality inverts for dimension reduction, resulting in an excessive amount of observations near the center of the distribution. This affects visualizations made on low-dimensional projections. [Ursula's paper] explains why points tend to be away from the center in the high-dimensional space, but crowd the center in low-dimensional projections, it is helpful to consider the projected volume relative to high-dimensional volume.

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.

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.

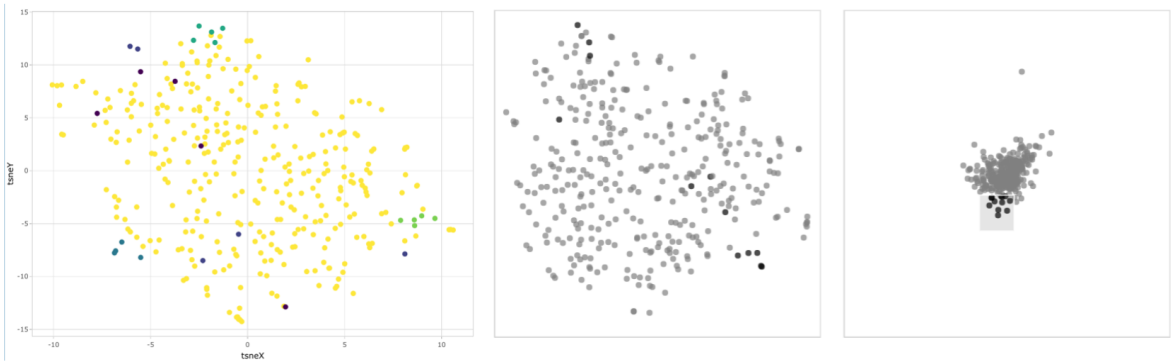
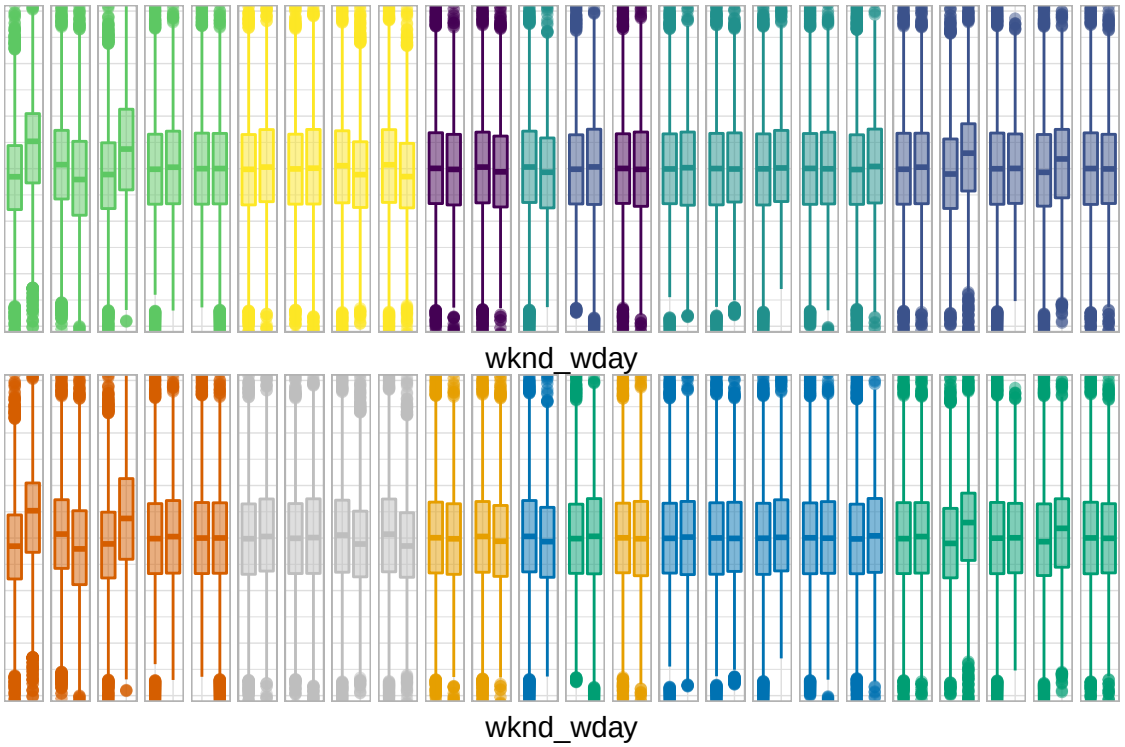


Figure 5: Instance selection using tours and projecting the points in a lower dimensional tsne cloud.

```
## Joining, by = c("customer_id", "group")
```

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



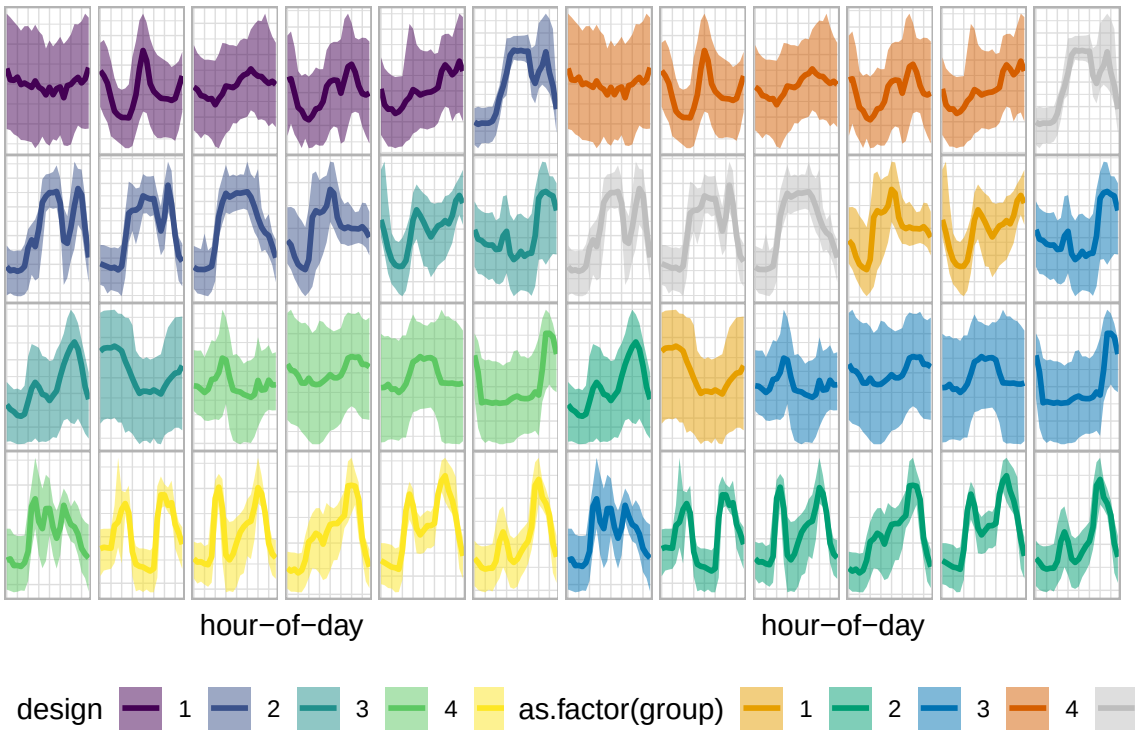


Figure 6: The energy demand distribution for the prototype designs (left) and clustering (right) is given for each hour of the day. Customer design grouping and clustering strategies complement one other well. For a one-to-one comparison, both sides have the same order of consumers. Design-4 has a low morning peak and a strong evening peak. Because other members of group-1 have higher late-night usage, group-2 adopts this design and assigns customer B to it rather than group-1.

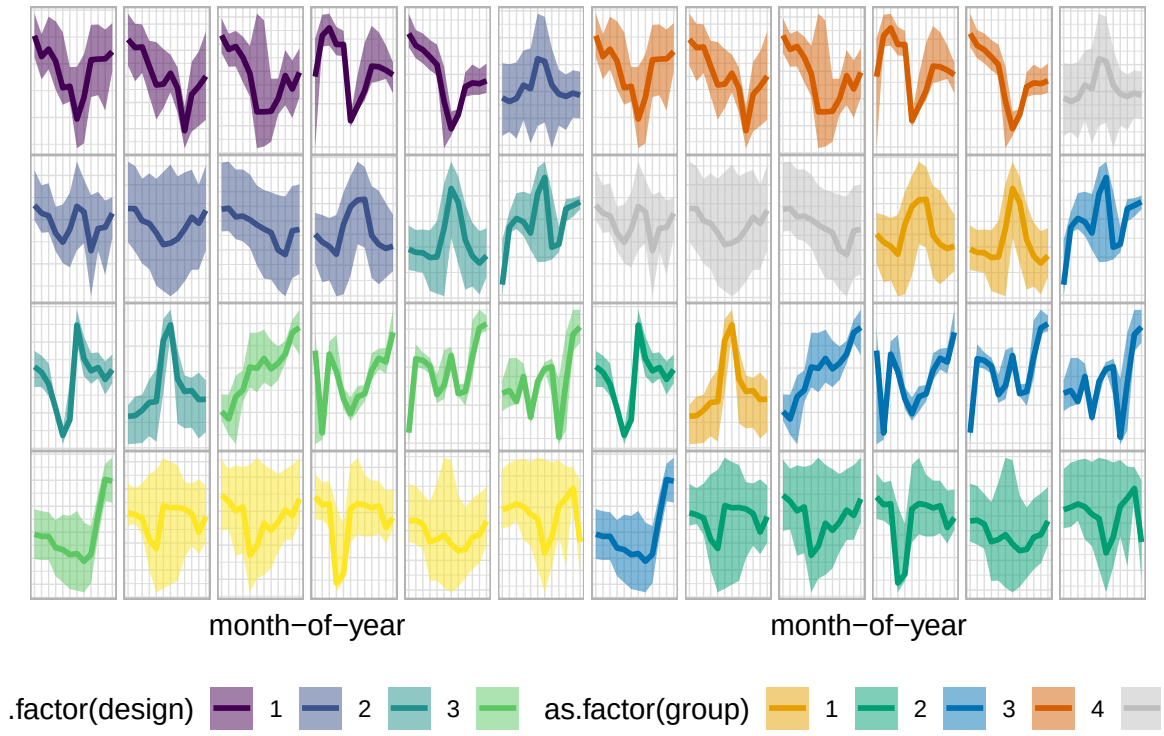


Figure 7: The energy demand distribution for the prototype designs (left) and clustering (right) are shown. The customer grouping by design and the clustering approach largely match, except for some exceptions. For example, in group 3, consumers are categorised by greater usage in the middle of the year, which corresponds to winter in Australia. This is a better grouping in terms of similar month-of-year pattern than originally proposed by the prototype design.

4.2 JS-based clustering

The 24 prototypes are clustered using the methodology described in **??**. The distribution of electricity demand for the selected 24 customers across hour-of-day and month-of-year are shown in the right panel of Figures 6 and 7 respectively. The median is shown by a line, and the shaded region shows the area between the 25^{th} and 75^{th}. Customers are placed in the same order on both sides for both figures to simplify one-to-one comparison. All customers with the same color represent the same design (left) or cluster (right). Overall, the clustering method has been able to capture the shapes in each designs quite well, irrespective of the fact that the designs are only selected based on the 50^{th} percentile with Designs 1, 2, 3, 4, and 5 correspond to groups 4, 5, 1, 3, 2 and, in most situations.Because our method uses distribution across hod, moy, and wknd-wday, there may be some mismatches from the designs; nonetheless, visualising individual customers can help us determine whether the grouping by clustering is better than the planned grouping through designs. In a few circumstances, for example, as shown in Figures 6 and 7, the clustering approach outperforms the design. 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.

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 **??** 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 group a strong early morning and late night hours. These consumers may be flexible students or elderly retirees who are night owls. Their day time usage has high variability. 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, but typically very less variability in other months.

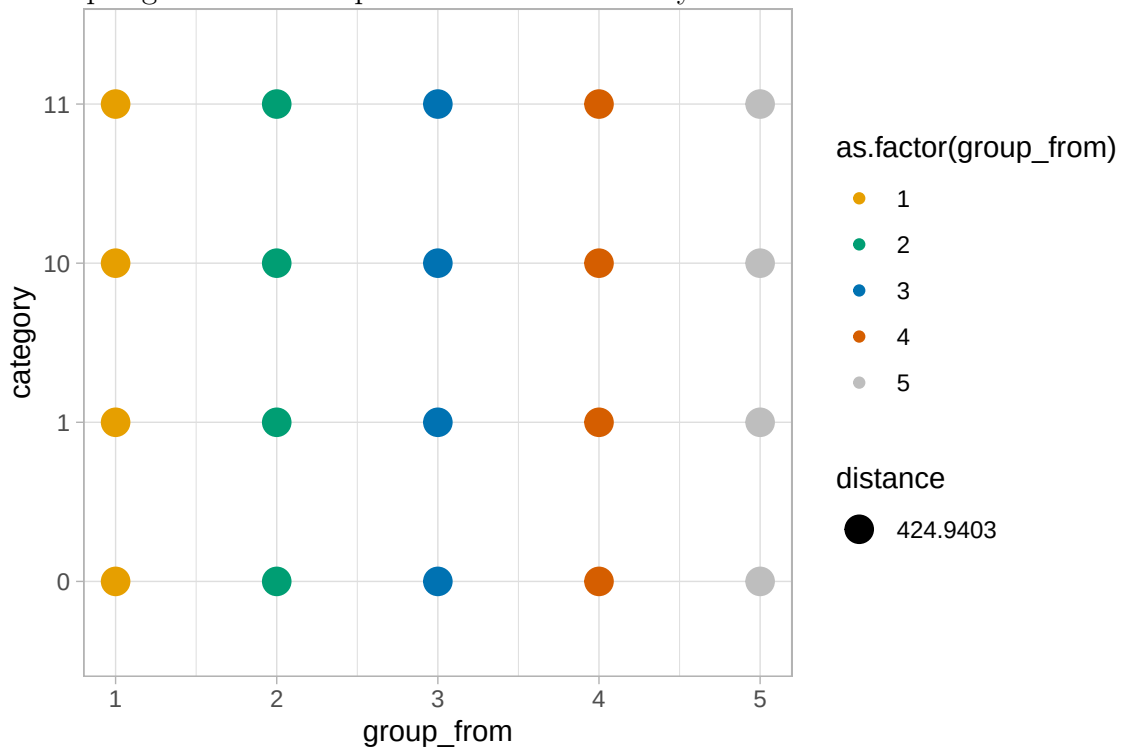
Group 2: They appear to work 9-5, get up and perform morning activities from 7-10, and then go. Evenings are busier than mornings as people return home to cook supper and perform other activities. During the fall (March-June), these consumers' average energy

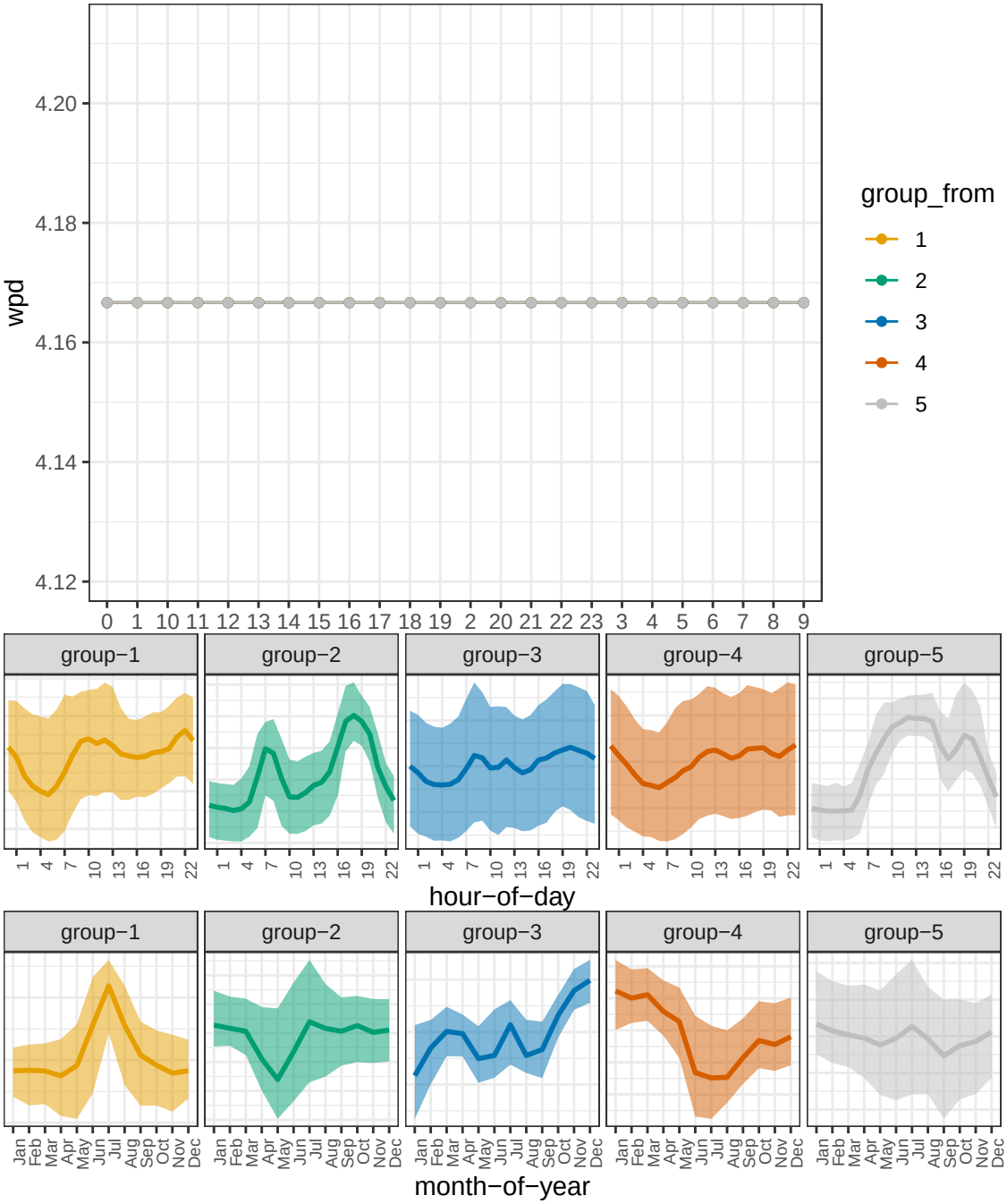
use declines, before rising again during the winter. However, energy behaviour varies more in the winter than in the summer and a bit in the fall months.

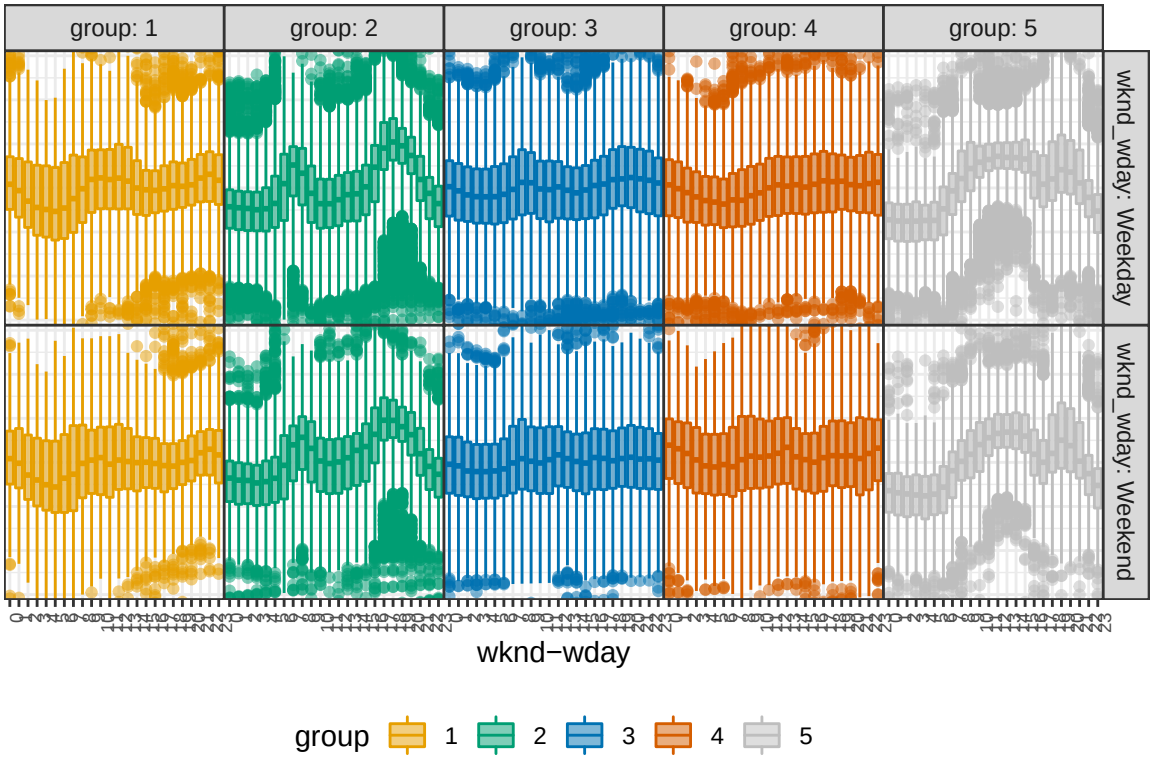
Group 3: Presence of children or stay-at-home parents is indicated by the group's almost equivalent morning, afternoon and evening median use and variability across all hours. They have low usage for winter months (or maybe due to use of gas heaters, it is not reflected in electricity usage). It seems that the users in this group are more concerned about the comfort and quality of life than the cost of electricity as we see their electricity demand peaking up in the spring and summer.

Group 4: These users have an opposite monthly profile to Group 3, with more usage during summer months (probably due to increased use of air conditioners) and demand decreasing across the year with least demand in winter months. The users in this cluster are sensitive to high temperature, but insensitive to lower temperature or their heater usage is not reflected in their electricity usage.

Group 5: Presence of children or stay-at-home parents is indicated by Group-4's almost equivalent morning, afternoon and evening profile. This group is more typical in their monthly energy behavior with more energy usage in winter and moderately decreasing for fall and spring months. The pattern in the variability is similar to that of the median.




wpd
4.20
4.18
4.16
4.14
4.12
0 1 10 11 12 13 14 15 16 17 18 19 2 20 21 22 23 3 4 5 6 7 8 9
group_from
1
2
3
4
5
group−1
group−2
group−3
group−4
group−5
1 4 7 10 13 16 19 22
hour−of−day
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
month−of−year
group
group−1
group−2
group−3
group−4
group−5



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 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.

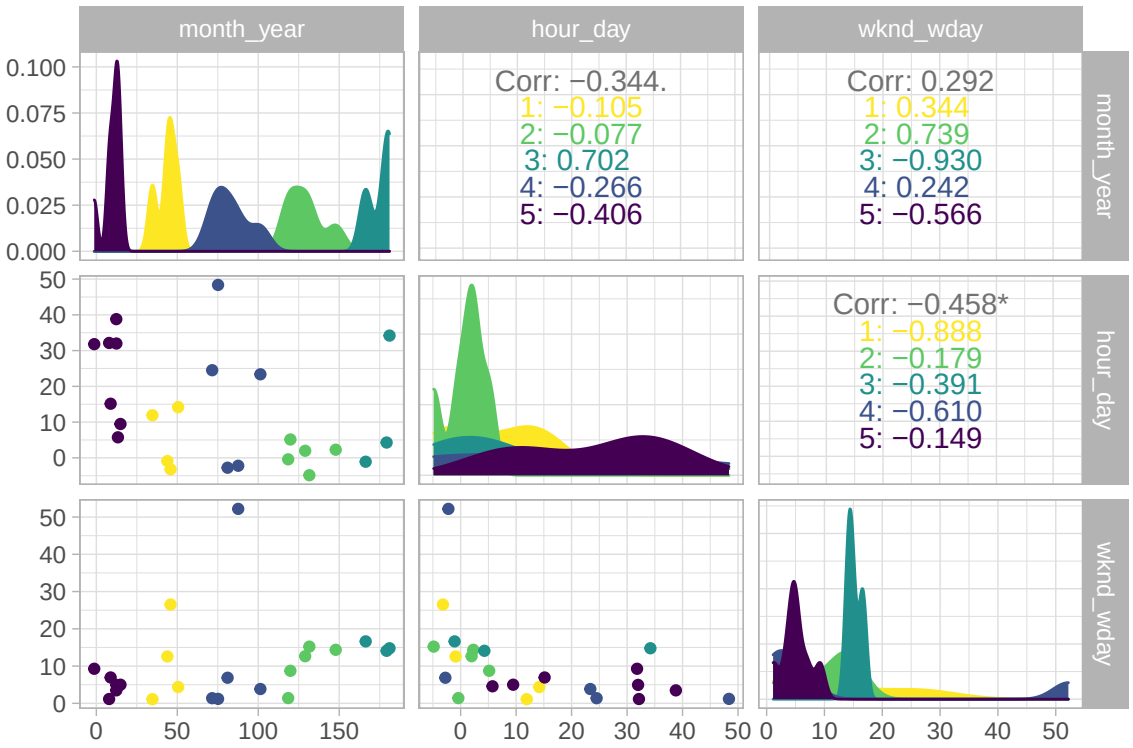


Figure 8: A ggpairsplot and parallel coordinate plot are used to depict each of the 24 customers. The ggpairs plot four distinct clusters across the month-of-year, which are less prominent across the hour-of-day and wknd-wday. The parallel coordinate plot ranks the variables in order of importance, indicating 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.

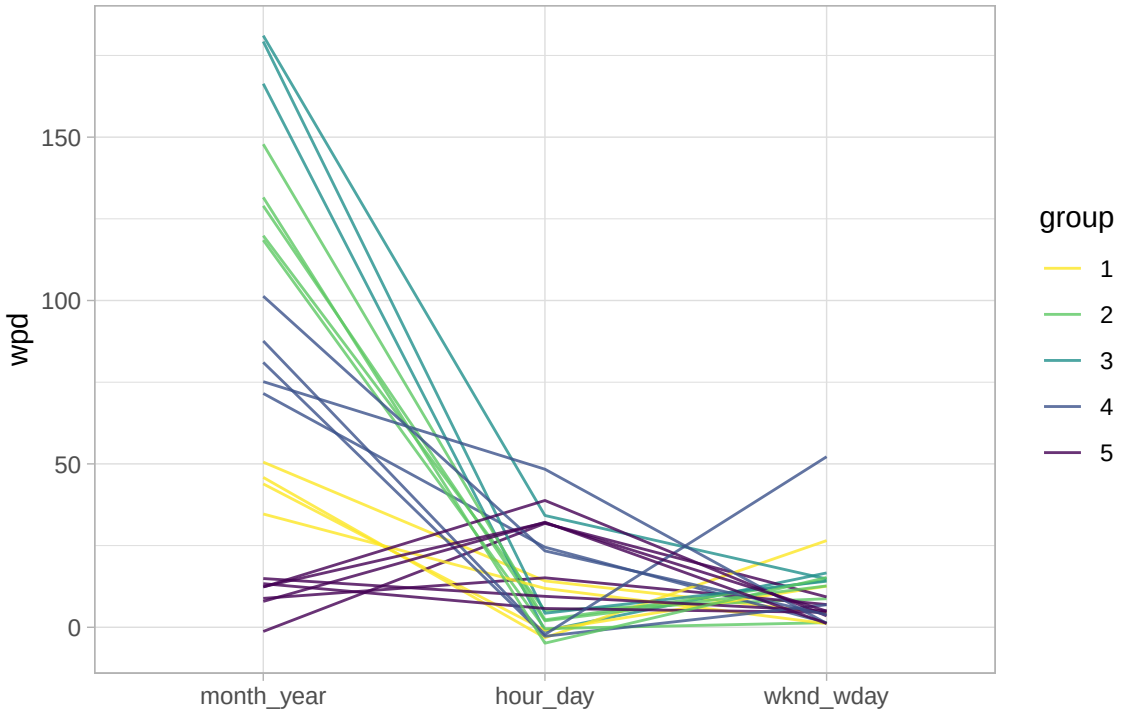


Figure 9: A parallel coordinate plot with the three significant cyclic granularities used for wpd-based clustering. The variables are sorted according to their separation across classes (rather than their overall variation between classes). This means that moy is the most important variable in distinguishing 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.

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.

```
## <ggproto object: Class ScaleDiscrete, Scale, gg>
##     aesthetics: colour
##     axis_order: function
##     break_info: function
##     break_positions: function
##     breaks: waiver
##     call: call
##     clone: function
##     dimension: function
##     drop: TRUE
##     expand: waiver
##     get_breaks: function
##     get_breaks_minor: function
##     get_labels: function
##     get_limits: function
##     guide: legend
##     is_discrete: function
##     is_empty: function
##     labels: waiver
##     limits: NULL
##     make_sec_title: function
##     make_title: function
##     map: function
```

```
##     map_df: function
##     n.breaks.cache: NULL
##     na.translate: TRUE
##     na.value: NA
##     name: waiver
##     palette: function
##     palette.cache: NULL
##     position: left
##     range: <ggproto object: Class RangeDiscrete, Range, gg>
##         range: NULL
##         reset: function
##         train: function
##         super:  <ggproto object: Class RangeDiscrete, Range, gg>
##     rescale: function
##     reset: function
##     scale_name: ochre
##     train: function
##     train_df: function
##     transform: function
##     transform_df: function
##     super:  <ggproto object: Class ScaleDiscrete, Scale, gg>
```

References

Aghabozorgi, S., Seyed Shirkhorshidi, A. & Ying Wah, T. (2015), 'Time-series clustering – a decade review', *Inf. Syst.* **53**, 16–38.

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.

Corradini, A. (2001), Dynamic time warping for off-line recognition of a small gesture vocabulary, *in* 'Proceedings IEEE ICCV workshop on recognition, analysis, and tracking of faces and gestures in real-time systems', IEEE, pp. 82–89.

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* .

Liao, T. W. (2005), 'Clustering of time series data—a survey', *Pattern recognition* **38**(11), 1857–1874.

Liao, T. W. (2007), 'A clustering procedure for exploratory mining of vector time series', *Pattern Recognition* **40**(9), 2550–2562.

Melnykov, V. (2013), 'Challenges in model-based clustering', *Wiley Interdiscip. Rev. Comput. Stat.* **5**(2), 135–148.

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.

Ratanamahatana, C. A. & Keogh, E. (2005), Multimedia retrieval using time series representation and relevance feedback, *in* 'International Conference on Asian Digital Libraries', Springer, pp. 400–405.

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

Ushakova, A. & Jankin Mikhaylov, S. (2020), 'Big data to the rescue? challenges in analysing granular household electricity consumption in the united kingdom', *Energy Research & Social Science* **64**, 101428.

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.

Wickham, H., Cook, D., Hofmann, H., Buja, A. et al. (2011), 'tourr: An r package for exploring multivariate data with projections', *Journal of Statistical Software* **40**(2), 1–18.