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

November 1, 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. First step is to decide what we mean by similar. 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. So going back to the smart meter example, $f_i(v)$ is the distribution of customer 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, we need to decide how to measure similarities between these collections of univariate probability distributions. There are multiple ways to measure similarities between time series based on probability distributions, depending on the objective of the problem. This paper focuses on looking at the (dis) similarity between underlying distributions that may have resulted in different patterns across different cyclic temporal granularities, that eventually have resulted in the (dis) similarity between time series. It considers a methodology with two appraoches for finding distances between time series. Both of these appraoches may be useful in a practical context and, depending on the data set, may or may not propose the same customer classification. The obtained distances 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. The methodology is explained in the Figure 1 and each element of the pipeline is discussed.

- *Find significant granularities or harmonies*

(Gupta et al. 2021) proposes a method for choosing significant cyclic granularities and

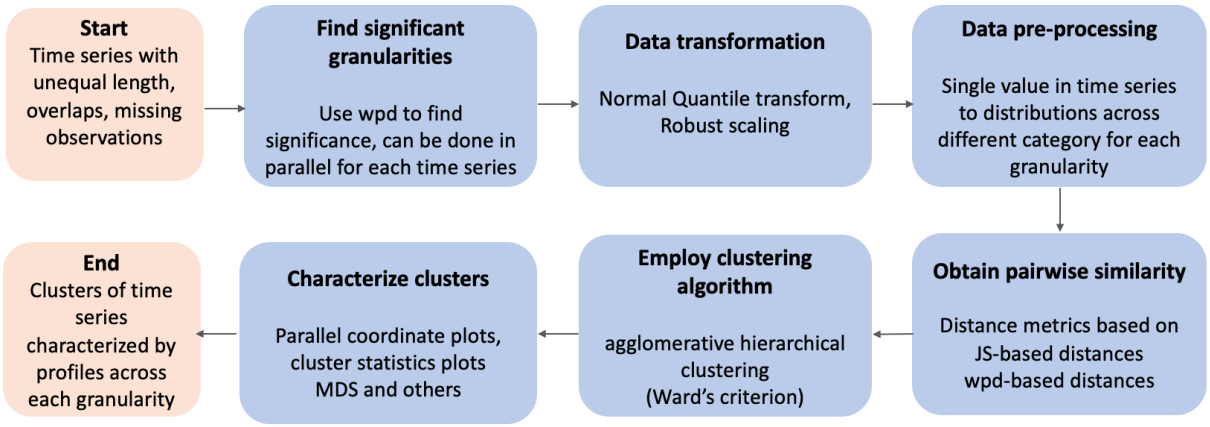


Figure 1: Flow chart illustrating the pipeline for methodology

harmonies, which is used in this work. We define "significant" granularities as those with significant distributional differences across categories. It is better to select only those granularities because it is expected that there would be some fascinating repetitive behaviour that we are interested in studying. It is worth noting that not all of the observations in the study may have the same set of important granularities. A method for selecting a list (S_c) of significant granularities for all observations may be as follows:

(a) eliminate from the comprehensive list the granularities that are inconsequential for all observations.

(b) consider only those granularities which are significant for most observations.

In both circumstances, there will be observations for which one or a few selected granularities are uninteresting. Even in that situation, having this group of observations that show no intriguing patterns over a granularity that frequently detects patterns may be useful. In contrast, if the granularities under consideration are indeed significant for a set of observations, distinct patterns could be detected while clustering them.

- *Data transformation*

Time series often have a somewhat skewed distribution and their ranges might vary greatly. It is helpful to do a statistical transformation on the data to bring all of them to the same range or normalize each series. For the JS-based approaches, two data transformation

techniques are utilised viz, Normal-Quantile Transform (NQT) and Robust scaling. NQT is a built-in transformation for computing *wpd*, which is the foundation of wpd-based distances.

Robust scaling The normalised i^{th} observation is denoted by $v_{norm} = \frac{v_t - p_{0.50}}{p_{0.75} - p_{0.25}}$, where v_t is the actual value at the t^{th} time point and $p_{0.25}$, $p_{0.50}$ and $p_{0.75}$ are the 25^{th}, 50^{th} and 75^{th} percentile of the time series for the i^{th} observation. v_{norm} has 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 The raw data for all observations is individually normal-quantile transformed (NQT) (Krzysztofowicz 1997), so that the transformed data follows a standard normal distribution. NQT will make the skewed distributions bell-shaped. As a result, determining which raw distribution was used is difficult using the modified distribution. Also, multimodality is disguised or inverted. This, however, is not a problem for implementation of this methodology as distributions are characterised by quantiles and the order of the quantiles is reserved under NQT.

- *Data pre-preprocessing*

Wang et al. (2020) introduced the tidy "tsibble" data structure to assist temporal data exploration and modeling. To start with, the measured variable for each key variable (observation in this context) is a time-indexed sequence of values for various measurement variables at each time point. This sequence, however, could be shown in several ways. A shuffle of the raw sequence may represent hourly consumption throughout a day, a week, or a year. Cyclic granularities like hour-of-day, hour-of-week can be expressed in terms of the index set in the "tsibble" data structure. But 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 induced by the multiple observations 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.

- *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 metric considers two time series to be similar if the distributions of each category of an individual cyclic granularity or combination of categories for interacting cyclic granularities are similar. In this study, the distribution for each category is characterised using deciles (can potentially consider any list of quantiles), and the distances between distributions are calculated using the Jensen-Shannon distances (Menéndez et al. (1997)), which are symmetric and thus could be used as a distance measure.

The sum of the distances between two series x and y in terms of cyclic granularity A is 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). After determining the distance between two series in terms of one granularity, we must combine them to produce a distance based on all significant granularities. 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 employed, which is also a distance metric since it is the sum of JS distances. In this approach, the variation in time series within each group is in magnitude rather than distributional pattern, while the variation between groups is only in distributional pattern across categories.

wpd-based distances

Compute weighted pairwise distances *wpd* (Gupta et al. (2021)) for all considered granularities for all observations. *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 normalised by appropriate parameters. A higher value of *wpd* indicates that some interesting pattern is expected, whereas a lower value would indicate otherwise.
Once we have chosen *wpd* as a relevant feature for characterizing the time series across one cyclic granularity, we have to decide how we combine differences between the multiple features (corresponding to multiple granularities) into a single number. The euclidean distance between them is chosen, with the granularities acting as variables and *wpd* representing the value under each variable. With this approach, we should expect the observations with similar *wpd* values to be clustered together. Thus, this approach is useful for grouping observations that have similar significance of patterns across different granularities. Similar significance does not imply similar pattern, which is where this technique varies from JS-based distances, which detect differences in patterns across categories.

- *Clustering algorithm*

With a way to obtain pairwise distances, any clustering algorithm can be employed that supports the given distance metric as input. A good comprehensive list of algorithms can be found in Xu & Tian (2015) based on traditional ways like partition, hierarchy or more recent approaches like distribution, density and others. We employ agglomerative hierarchical clustering in conjunction with Ward's criteria (XXX reference). Hierarchical cluster techniques fuse neighbouring points sequentially to form bigger clusters, beginning with a full pairwise distance matrix. The distance between clusters is described using a "linkage technique". This agglomerative approach successively merges the pair of clusters with the shortest between-cluster distance using Ward's linkage method. Hierarchical algorithms are one of the most widely used, can operate with data of any shape, has reasonable scalability, and the number of clusters is not needed as a parameter.

- *Characterization of clusters*

Cluster characterization, both quantitatively and qualitatively, is a crucial aspect in cluster analysis. Cook & Swayne (2007) lists numerous methods for characterising clusters. Listed below are a few techniques and R packages that are utilized in this study.

(a) *Parallel coordinate plots* (Wegman (1990)) are often used to visualise high-dimensional and multivariate data, allowing visual grouping and pattern detection.. A Parallel Coordinates Plot features parallel axes for each variable. Each axis is linked by lines. The axes' arrangement may affect the reader's interpretation of the data. Changing the axes may reveal patterns or relationships between variables for categorical variables. However, for categories with cyclic temporal granularities, preserving the underlying ordering is more desirable.

(b) *Scatterplot matrix* contains pairwise scatter plots of the p variables. Pairwise scatter plots are useful for figuring out how variables relate to each other and how factors determine the clustering.

(c) *Displaying cluster statistics* are useful when we have larger problems and it is difficult to read the Parallel coordinate plots due to congestion. (Dasu et al. (2005))

(d) *MDS, PCA and t-SNE* While all of them use a distance or dissimilarity matrix to construct a reduced-dimension space representation, their goals are diverse. PCA seeks 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. The t-SNE embeddings will compress data points which are close in high-dimensional space.

(e) *Tour* is a collection of interpolated linear projections of multivariate data into lower-dimensional space. As a result, the viewer may observe the high-dimensional data's shadows from a low-dimensional perspective.

The cluster characterization approach varies depending on the distance metric used. Parallel coordinate plots, scatter plot matrices, MDS or PCA are potentially useful ways

to characterize clusters using wpd-based distances. For JS-based distances, plotting cluster statistics is beneficial for characterization and variable importance could be displayed through parallel coordinate plots. This part of the work uses R packages `GGally` (Schloerke et al. (2021)), `Rtsne` (Krijthe (2015)), `ggplot2` (Wickham2009pk), `tour` (Wickham et al. (2011)), `stats` (R Core Team (2021)).

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 constructed independent replications of all data designs $R = 25, 250, 500$ to investigate if our proposed clustering method can discover distinct designs in small, medium, and big number of series. All designs employ $T = 300, 1000, 5000$ sample sizes to evaluate small, medium, and large sized series. Variations in method performance may be due to different jumps between categories. So a mean difference of $diff = 1, 2, 5$ is examined. The approaches' performance varies with the number of significant granularities. So all, few, and one major granularity scenarios are considered. Please see the Supplementary section (`https://github.com/Sayani07/paper-gracsR`) for the code and findings.

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. 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 instance where $g1$, $g2$, and $g3$ all contribute to design distinction. Meaning that at least one of the designs to be categorized will have significantly different patterns for each granularity. In Table 1 (top), we explore various distributions across categories (as shown in Table Table 1 (bottom). Figure 2 shows the simulated variable's linear and cyclic representations for each of these five designs. The structural difference in the time series variable is impossible to discern from the linear view. The shift in structure may be seen clearly in the distribution of cyclic granularities (Figure 2 (right)). The following scenarios use solely graphical displays across cyclic granularities to highlight distributional differences in categories.

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 $g3$ for few designs and $g1$ does not change across designs. The proposed design is shown in Figure 3(left).

(c) One signifiant granularity

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

Only one granularity is responsible for identifying the designs in this case. This is depicted in Figure 3 (right) where only $g3$ affects the designs significantly.

3.2.2 Interaction of granularities

The proposed methods could be extended when two granularities of interest interact and we want to group subjects based on the interaction of the two granularities. Consider a group that has a different weekday and weekend behavior in the summer but not in the winter. This type of combined behaviour across granularities can be discovered by evaluating the distribution across combinations of categories for different interacting granularities (wknd-wday and month-of-year in this example). As a result, in this scenario, we analyse a combination of categories generated from different distributions. Consider a case in which there are only two interacting granularities of interest, $g1$ and $g2$. In contrast to the previous situation, when we could study distributions across $l_{g_1} + l_{g_2} = 5$ separate categories, with interaction, we must evaluate the distribution of the $l_{g_1} * l_{g_2} = 6$ combination of categories. Consider the 4 designs in Figure 4, where various distributions are assumed for different

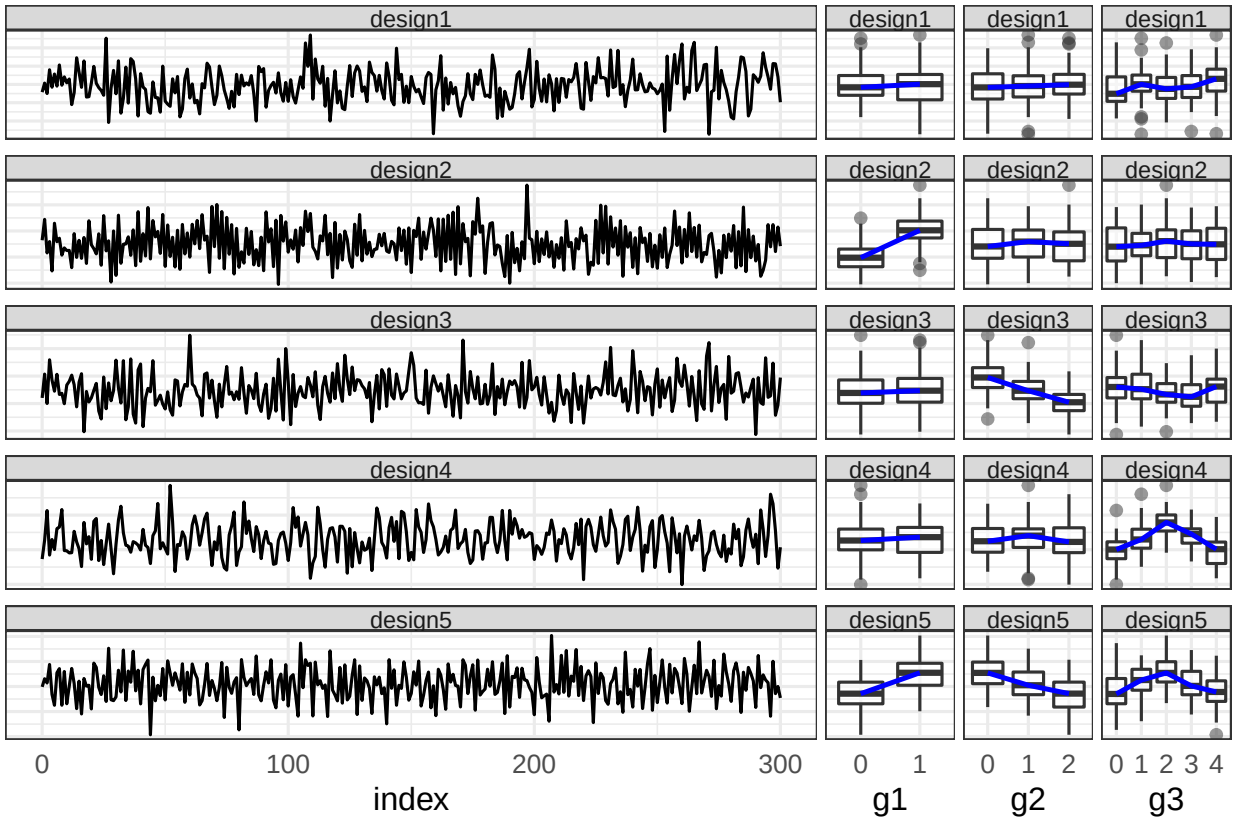


Figure 2: The linear (left) and cyclic (right) representation of the measured variable is shown. In this scenario, all of $g1$, $g2$ and $g3$ changes across at least one design. Also, it is not possible to comprehend these patterns across cyclic granularities or group similar series just by looking at the linear plots.

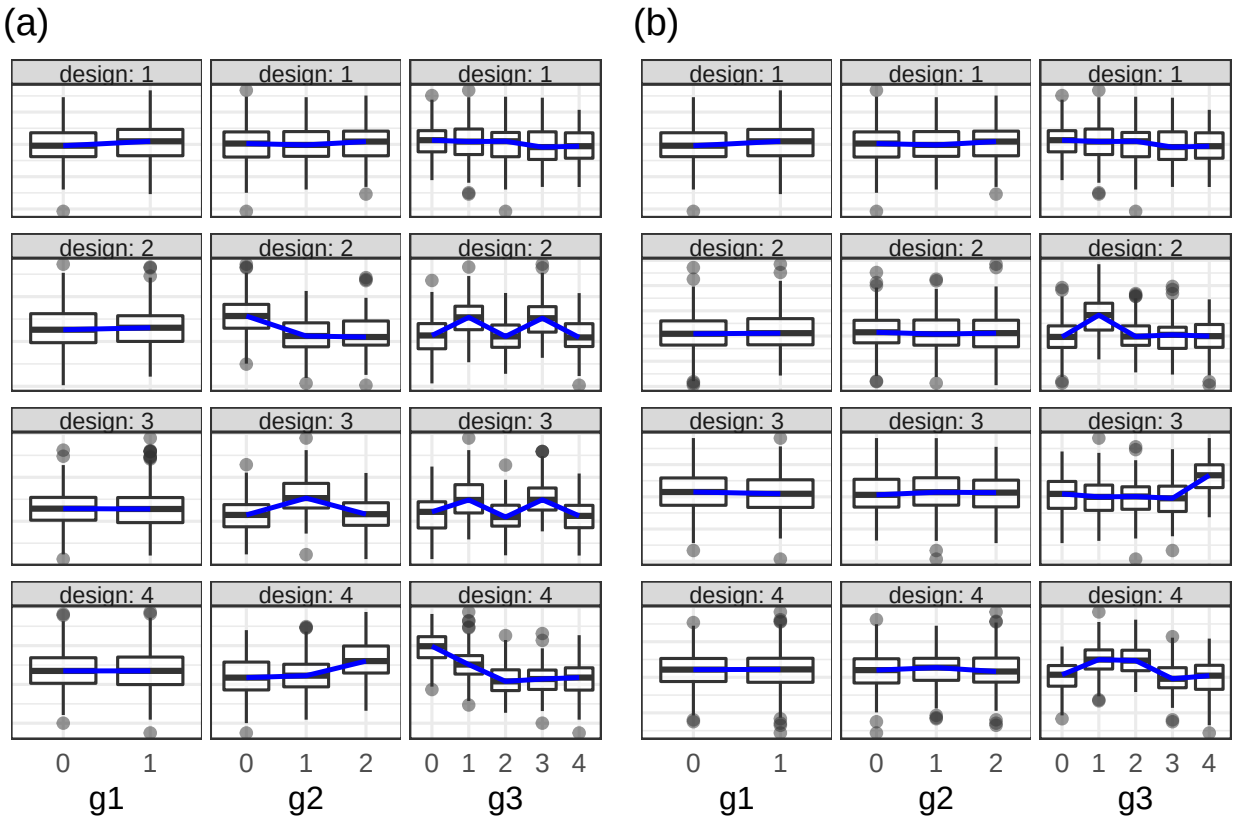


Figure 3: For the left scenario $g1$, $g2$ would change across atleast one design but $g3$ change remains same across all design. For the right one, only $g3$ changes across different designs.

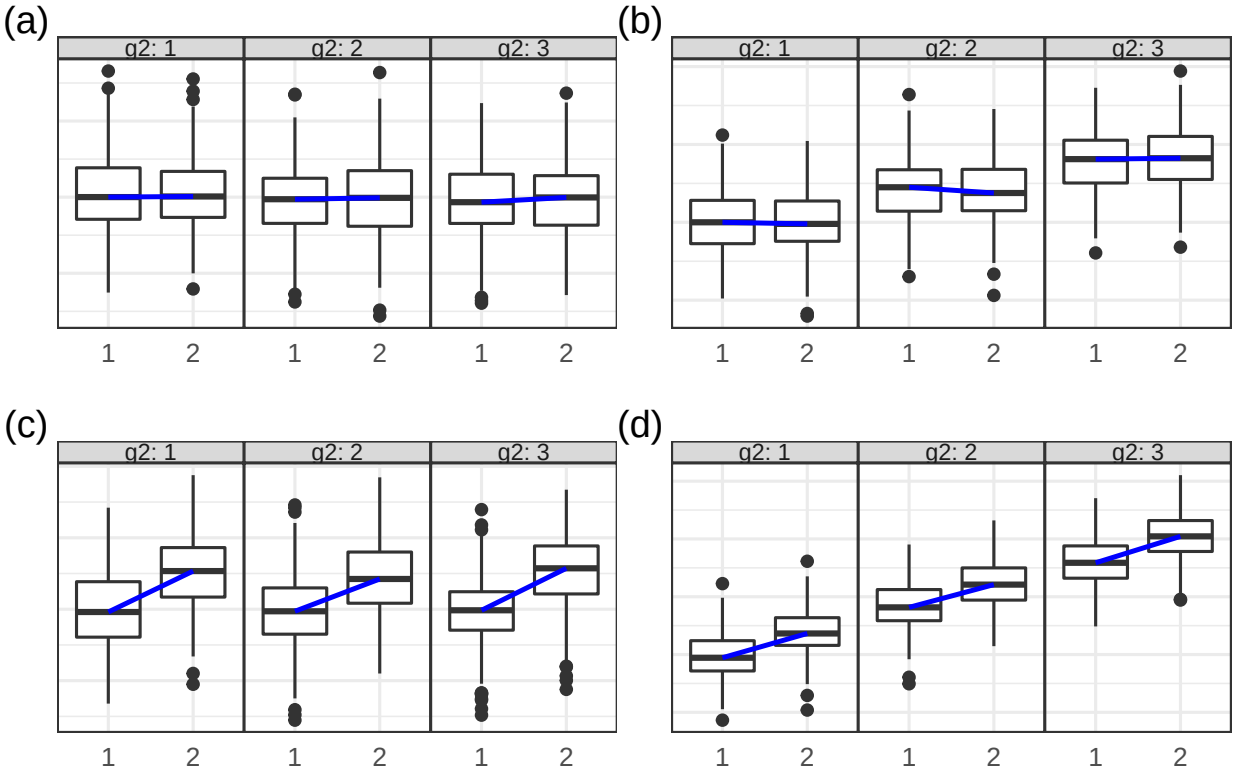


Figure 4: Design-1 (a) has no change in distributions across different categories of $g1$ or $g2$, while Design-2 (b) and Design-3 (c) change across only $g1$ and $g2$ respectively. Design-4 (d) changes across categories of both $g1$ and $g2$.

combinations of categories, resulting in different designs. Design-1 exhibits no change in distributions across $g1$ or $g2$, whereas Design-2 and Design-3 alter across only $g1$ and $g2$, respectively. Design-4 varies across both $g1$ and $g2$ categories. Design-3 and Design-4 appear similar based on their relative difference across consecutive categories, however Design-4 also changes across facets, unlike Design-3, which has all facets look the same.

3.3 Visual exploration of findings

All of the approaches were fitted to each data design and for each combination of the considered parameters. The formed clusters have to match the design, be well separated, and have minimal intra-cluster variation. It is possible to study these desired clustering traits visually in a more comprehensive way than just looking at index values. So we use MDS and parallel coordinate graphs to demonstrate the findings:

- In Figure 6, we tried to see how separated our clusters are. We observe that in all scenarios and for different mean differences, cluster is separated. However, the separation increases with increase in mean differences across scenarios. This is intuitive because with increasing difference between categories, it gets easier for the methods to correctly distinguish the designs. Results corresponding to Scenario (a) is shown here.

- Figure 5 depicts a parallel coordinate plot with the vertical bar showing total inter-cluster distances with regard to granularities $g1$, $g2$, and $g3$. For all simulation settings and scenarios, values are represented as a sequence of lines connected across each axis. One line in the figure, for example, shows the inter-cluster distances for one simulation scenario. The lines are not coloured by group since the purpose is to highlight the contribution of the factors to categorization rather than class separation. The first plot shows that no variable stands out in the clustering, but the following two designs show that {g1} and {g1, g2} have very low inter cluster distances, meaning that they did not contribute to the clustering. It is worth noting that these facts correspond to our original assumptions when developing the scenarios, which incorporate distributional differences over three (a), two (b), and one (c) significant granularities. Hence, Figure 5 (a), (b), and (c) validate the construction of scenarios (a), (b), and (c) respectively.

- The js-robust and wpd methods perform worse for $nT = 300$, then improve for higher nT evaluated in the study. Although, with the type of residential load data sets, a complete year of load is the minimum requirement to capture predicted differences in winter and summer profiles, for example. Even if the data is only available for a month, nT is expected to be at least 1000 with half-hourly data. Hence, practically this is not a challenge as long as the performance is promising for higher $nT = 300$.

- For smaller difference between categories, it is seen that method js-nqt would preform better js-robust in our study sample. In our study sample, method js-nqt outperforms method js-robust for smaller differences between categories. More testing, however, is required to corroborate this.

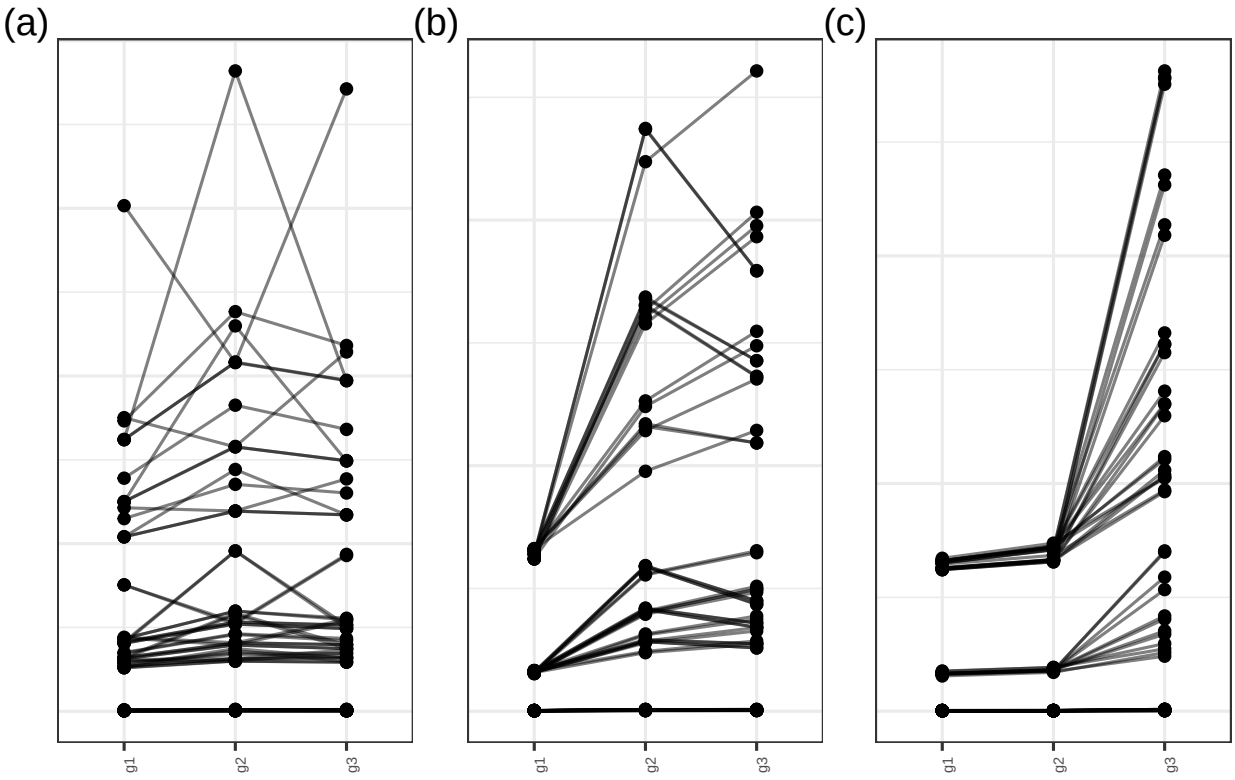


Figure 5: The parallel coordinate plot illustrates the total inter-cluster distances for granularities $g1$, $g2$, and $g3$. One line in the figure depicts the inter-cluster distances for a single simulation scenario. While the first plot indicates that no variable stands out during clustering, the next two designs demonstrate that g1 and g1, g2 have extremely low inter-cluster distances, indicating that they did not contribute to clustering. It is worth emphasising that these facts are consistent with our initial assumptions when designing the scenarios.

For more detailed plots and tables, please refer to the supplementary paper.

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. In most

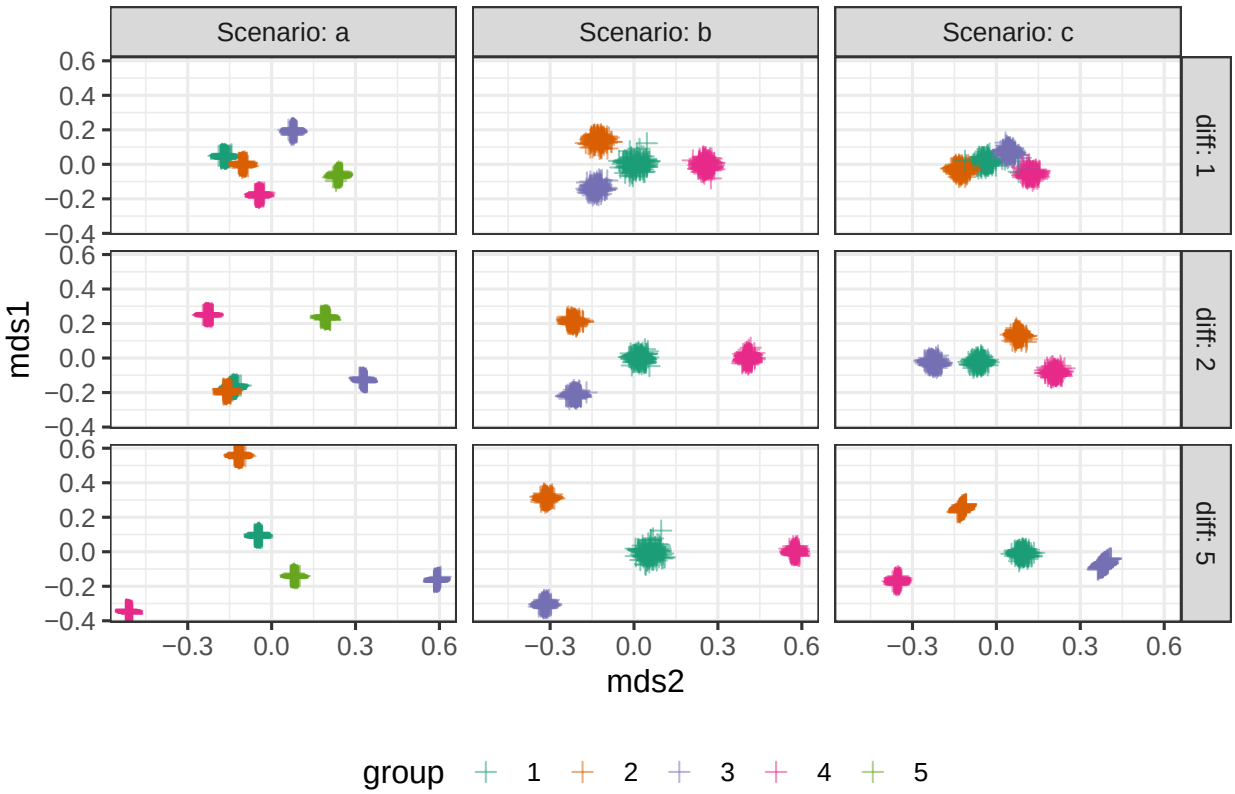


Figure 6: Cases are projected through MDS and the first two dimensions are shown to display the relative position of clusters for different simulation parameters. The rows of the grid represent the different mean differences in designs and the columns represent different designs. It can be observed that clusters become more compact and separated for higher distributional differences between categories.

cases, electricity data is expected to have multiple seasonal patterns like daily, weekly or annual. We do not learn about these repetitive behaviors from the linear view because too many measurements all squeezed in that representation. Hence we transition into looking at cyclic granularities, that can potentially provide more insight on their repetitive behavior. The raw data for these consumers is of unequal length, with varying start and finish dates. Because our proposed methods evaluate probability distributions rather than raw data, neither of these data features would pose any threat to our methodology unless they contained any structure or systematic patterns. Additionally, there were missing values in the database but further investigation revealed that there is no structure in the missingness (see Supplementary paper for raw data features and missingness). The study begins by subsetting a data set along all dimensions of interest using data filtering and prototyping. By grouping the prototypes using our methods and assessing their meaning, the study hopes to unravel some of the heterogeneities observed in energy usage data. Because our application does not employ additional customer data, we cannot explain why consumption varies, but rather try to identify how it varies.

Data filtering and variable selection

- Choose a smaller subset of randomly selected 600 customers with no implicit missing values for 2013.

- Obtain *wpd* for all cyclic granularities considered for these customers. It was found that `hod` (hour-of-day), `moy` (month-of-year) and `wkndwday` (weeknd/weekday) are coming out to be significant for most customers. We use these three granularities while clustering.

- Remove customers whose data for an entire category of a significant granularity is empty. For example, a customer who does not have data for an entire month is excluded because their monthly behaviour cannot be analyzed.

- Remove customers whose energy consumption is 0 in all deciles. These are the clients whose consumption is likely to remain essentially flat and with no intriguing repeated patterns that we are interested in studying.

Prototype selection

Supervised learning uses a training set of known information to categorize new events through instance selection. Instance selection (Olvera-López et al. (2010)) is a method of rejecting instances that are not helpful for classification. This is analogous to subsampling the population along all dimensions of interest such that the sampled data represents the primary features of the underlying distribution. Instance selection in unsupervised learning has received little attention in the literature, yet it could be a useful tool for evaluating model or method performance. There are several ways to approach the prototype selection. Following Fan et al. (2021)'s idea of picking related examples (neighbours) for each instance (anchor), we can first use any dimensionality reduction techniques like MDS or PCA to project the data into a 2D space. Then pick a few "anchor" customers who are far apart in 2D space and pick a few neighbors for each. Unfortunately, this does not ensure that consumers with significant patterns across all variables are chosen. Tours can reveal variable separation that was hidden in a single variable display better than static projections. Hence we perform a linked tour with t-SNE layout using the R package `liminal` (Lee (2021)) to identify customers who are more likely to have distinct patterns across the variables studied. Please see the Supplementary article for further details on how the prototypes are chosen. Figure 7 (a, b, c) shows the raw time plot, distribution across `hod`, `moy` and `wkndwday` for the set of chosen 24 customers. Few of these customers have similar distribution across `moy` and some are similar in their `hod` distribution.

4.1 Clustering

Cluster characterization is a crucial aspect of cluster analysis. The 24 prototypes are clustered using the methodology described in 2 and results are reported below. In the following plots, the median is shown by a line, and the shaded region shows the area between the 25^{th} and 75^{th}. All customers with the same color represent same clustered groups. Groups by JS-based distances and wpd-based distances are colored differently as they represent different groupings. The plotting scales are not displayed since we want to emphasize comparable shapes rather than scales. The idea is that a customer in a cluster may have low total energy usage, but their behavior may be quite similar to a customer

with high usage with respect to shape or significance across cyclic granularities.

4.1.1 JS-based distances

For clustering based on JS-based distances, we chose the optimal number of clusters using (Hennig (2014)) as 5. The distribution of electricity demand for the selected 24 customers across hod, moy and wdwn are shown in Figure 7 (d, e, f). Our methodology is useful for grouping similar distributions over hod and moy and they are placed closely for easy comparison. Of course, certain customers in each group have distributions that differ from other members in the same group. However, it appears that the aim of grouping comparable distributions over considered variables has been accomplished to some extent. Figure 8 shows the summarized distributions across 5 groups and assists us in characterizing each cluster. Figure 8 shows Groups 2 and 5 show a stronger hod pattern with a typical morning and evening peak, whereas groups 1, 3, and 5 show a moy pattern with higher usage in winter months. Differences in wknd-wday between groups are not discernible, implying that it may not be a relevant variable in distinguishing various clusters.

4.1.2 wpd-based distances

We chose the optimal number of clusters using (Hennig (2014)) as 3. 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*. There is only one customer who has significant *wpd* across *wkndwday* and stands out from the rest of the customers. Group 3 has a higher wpd for hod than moy or wkndwday. Group 2 has the most distinct pattern across moy. Group 1 is a mixed group that has strong patterns on at least one of the three variables. The findings vary from js-based clustering, yet it is a helpful grouping.

Things become far more complicated when we consider a larger data set with more uncertainty, as they do with any clustering problem. Summarizing distributions across clusters with varied or outlying customers can result in a shape that does not represent the group. Furthermore, combining hetergeneous customers may result in similar-looking final

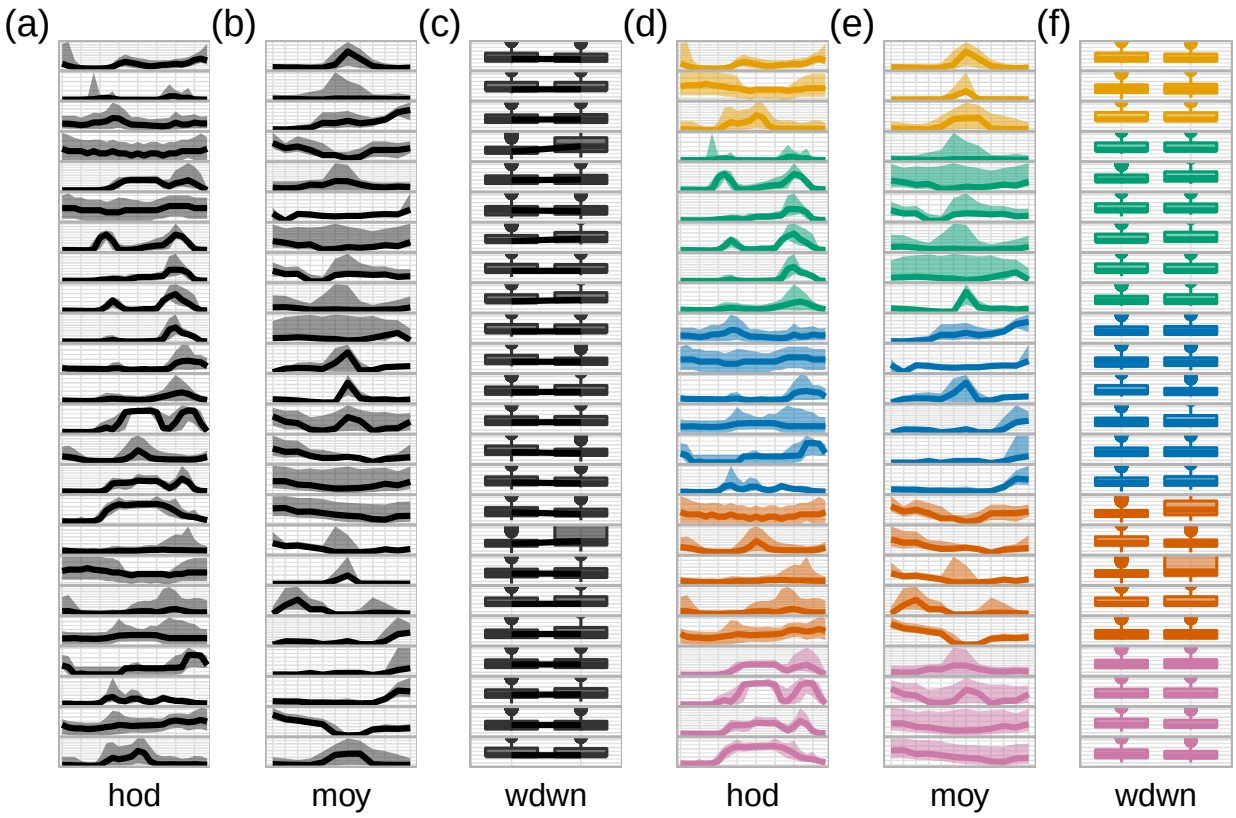


Figure 7: The distribution of selected consumers over hod (a, d), moy (b, e), and wkndwday (d, f). In each case, the same colour denotes the same group in plots (d), (e), (f) and are placed together to facilitate comparison. That means the customer orderings are different for (a, b, c) and (d, e, f). Our clustering methodology is useful for grouping similar distributions over hod and moy. Of course, certain customers in each group have distributions that differ from those of other members in the same group. However, it appears that the aim of grouping comparable distributions over considered variables has been accomplished to some extent.

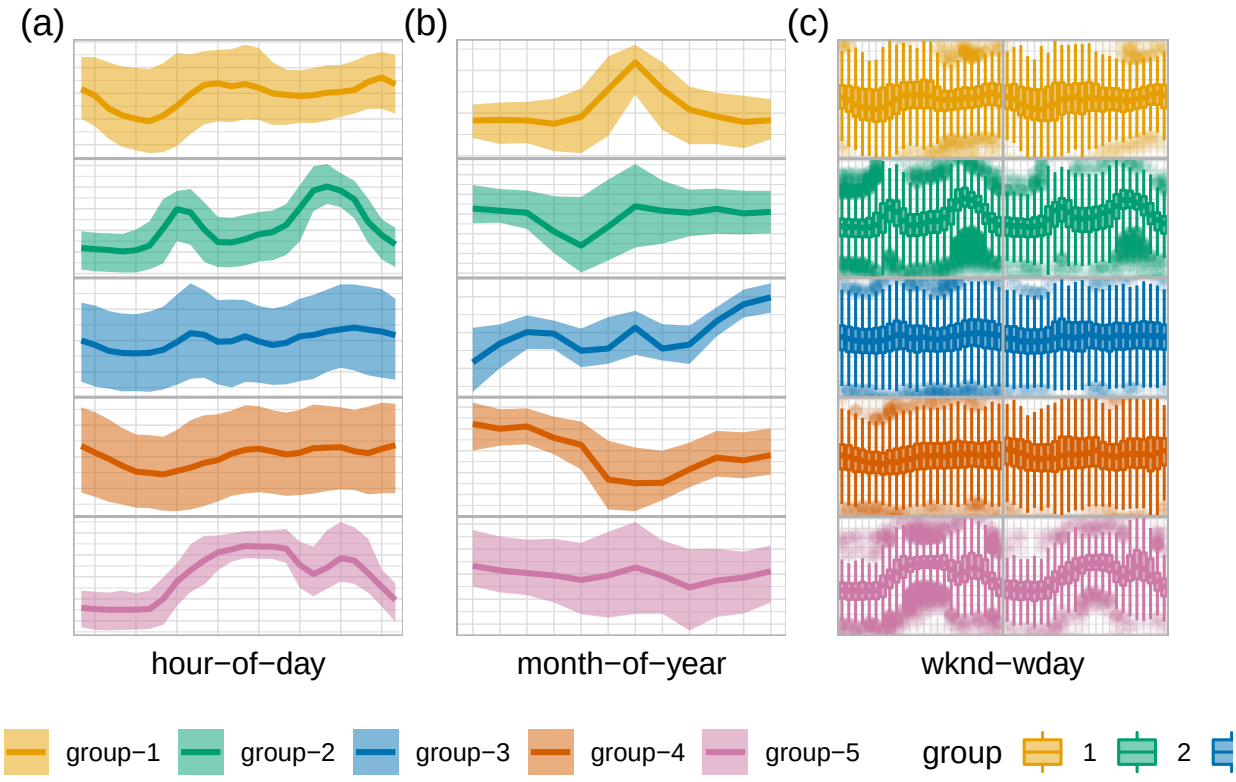
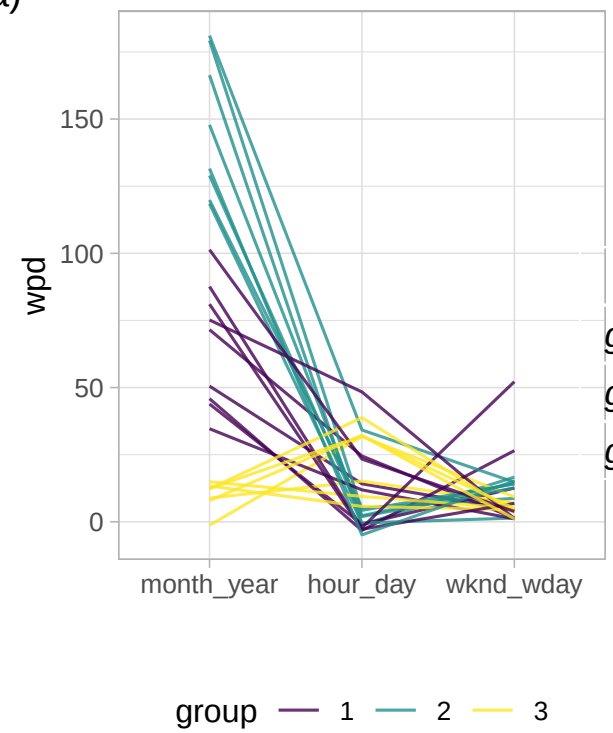


Figure 8: The distribution of electricity demand for the clusters across hod (a), moy (b) and wkndwday (c). It seems like group 2 and 5 have a hod pattern across its members, while group 1, 3, 5 have a moy pattern. Wknd-wday variations across groups are not distinguishable, indicating that it is not a critical variable for clustering. It is helpful to compare the summarised distributions of groups to that of individuals to confirm that the most of individuals in the group have the same characterisation.

(a)



(b)

	nobs	moy	hod	wdwr
group−1	9	71.55	11.91	4.38
group−2	8	139.7	2.13	14.24
group−3	7	12.33	31.8	4.95

Figure 9: Each of the 24 customers is represented by a parallel coordinate plot (a) with three wpd-based groupings. The plot shows that moy is the most important variable in identifying clusters, whereas wkdn-wday is the least significant and has the least fluctuation. One particular customer with high wpd across wknwday stands out in this display. Group 3 has a higher wpd for hod than moy or wkndwday. Group 2 has most discernible pattern across moy. Group 1 is a mixed group with strong patterns on atleast one of the three variables. All of these could be observed from the plot or the table (b) which shows median wpd values for each group.

clusters that are not effective for visually differentiating them. It is also worth noting that the wknd-wday behavior in the given case does not characterize any cluster. This, however, will not be true for all of the customers in the data set. If more extensive prototype selection is used, resulting in more comprehensive prototypes in the data set, this method might be used to classify the entire data set into these prototype behaviors. However, the goal of this section was to have a few customers that have significant patterns over one or more cyclic granularities, apply our clustering methodology to cluster them, and demonstrate that the method produces useful clusters.

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

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. (2005), Grouping multivariate time series: A case study, *in* 'Proceedings of the IEEE Workshop on Temporal Data Mining: Algorithms, Theory and Applications, in conjunction with the Conference on Data Mining, Houston', Citeseer, pp. 25–32.

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* .

Hennig, C. (2014), How many bee species? a case study in determining the number of clusters, *in* 'Data Analysis, Machine Learning and Knowledge Discovery', Springer International Publishing, pp. 41–49.

Krijthe, J. H. (2015), *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.15.
URL: *https://github.com/jkrijthe/Rtsne*

Krzysztofowicz, R. (1997), 'Transformation and normalization of variates with specified distributions', *J. Hydrol.* **197**(1-4), 286–292.

Lee, S. (2021), *liminal: Multivariate Data Visualization with Tours and Embeddings*. R package version 0.1.2.
URL: *https://CRAN.R-project.org/package=liminal*

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.

Menéndez, M. L., Pardo, J. A., Pardo, L. & Pardo, M. C. (1997), 'The Jensen-Shannon divergence', *J. Franklin Inst.* **334**(2), 307–318.

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.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

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.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A. & Crowley, J. (2021), *GGally: Extension to 'ggplot2'*. R package version 2.1.1.
URL: *https://CRAN.R-project.org/package=GGally*

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 & 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.

Xu, D. & Tian, Y. (2015), 'A comprehensive survey of clustering algorithms', *Annals of Data Science* **2**(2), 165–193.