Clustering based on probability distributions with application on residential customers

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

October 11, 2021

Abstract

Clustering elements based on behavior across time granularities

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

*Email: Sayani.Gupta@monash.edu

1 Introduction

Large spatio-temporal data sets, both open and administrative, offer up a world of possibilities for energy and social science 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 and Data61 CSIRO. The project provides half-hourly data from over 13,000 household electricity smart meters for nearly two years. Electricity utilities can utilize these smart meter usage patterns of customers to develop targeted tariffs for individual groups and alleviate the problem of volatility in production by capitalizing on the flexibility of consumers.

The enormous quantity of data provides for greater individual level clarity and analysis. However, due to the growing variety of consumers, larger data sets include greater uncertainty about consumer behavior. Households are distributed geographically, and have different demographic properties such as the existence of solar panels, central heating or air conditioning. Multiple temporal dependencies define the behavioral patterns, which vary from customer to customer. Some families, for example, use a dryer to dry their clothing, while others hang them to dry on a line. This may be reflected in their weekly profile. They may have monthly variations where some customers are more prone to use air conditioners or heaters than others despite the existence of comparable electrical equipment and being subjected to similar weather conditions. The variations in behavior may occur on a regular basis, with some consumers being night owls and others being morning larks. Day-off habits may vary depending on whether consumers choose to remain at home or engage in outdoor activities. These recorded time series are asynchronous, with varying time lengths for different houses and missing observations. This huge amount of noisy, patchy, and unequal time-series common in residential smart meter data combined with the different causes of variation in their energy behavior makes the problem of effectively segmenting consumers with comparable consumption a particularly intriguing one.

Segmenting unlabeled time series patterns has received little attention in the literaturee. This is a situation when there is no additional information on the socio-economic charactertistics, property types or family size of the customers. There are two reasons why any additional consumer level data would not be available. Customer profiles, like prop-

erty features, vary over time and are not always updated in a timely manner by energy suppliers. Also such data might not be available when the anonymity of the customers are of utmost importance. So this study does not explain why consumption differs. Instead, this work investigates how much energy usage heterogeneity can be found in smart meter data and what some of the most common electricity use patterns are.

An aggregation approach to reduce the data size is not exactly ideal when dealing with highly heterogeneous profiles, as (XXX big data to the rescue example) demonstrates. The foundation for this study is Tureczek2017-pb, which conducts a systematic review of the current state of the art in smart meter data analytics, which evaluates approximately 2100 peer-reviewed papers and summarizes the main findings. None of the 34 selected papers which focus on clustering consumption are based on Australian smart meter data. The clustering is frequently applied directly to the raw data without scrutinizing for auto correlation and periodicity. The algorithm most ubiquitously employed is K-Means. But the omission of the time series structure or correlation in the analysis while employing K-Means leads to inefficient clusters. Principal Component Analysis or Self-Organizing Maps removes correlation structures and transitions the data to a reduced feature space, but it comes at a cost of interpretability of the final results. **?** has shown that a transformation of data to incorporate autocorrelation before K-Means clustering can improve performance and enable K-Means to deliver smaller clusters with less within-cluster variance. However, it does not explain the cluster composition by combining it with external data. Some papers present pre-processing of the smart-meter data before clustering through principal component analysis or factor analysis for dimensionality reduction or self-organizing maps for 2-Dimensional representation of the data (**?**). Other algorithms used in the literature include k-means variations, hierarchical methods and k-medoids based on a greedy algorithm have been designed to select typical periods in the time series. As the methods are often situation specific, it makes sense to compare them on the performance rather than any standard performance metric. A type of clustering based on information theory such as Shannon or Renyi entropy and their variants are addressed in , which differs from typical methods adopted for electricity consumer classification, based on the Euclidean distance notion. **?** presents strategy to address the problems on patchy, and unequal time-series

common in residential data sets by converting load time series into map models. Most time-series clustering models are limited to handling time domain with same start and end date and time. Most of the solutions to handle this like longest common subsequence, dynamic time warping are prone to computational limit with increased length of the series. The massive amount of data generated in such projects could be overwhelming for analysis. Electricity utilities can utilize the consumption patterns of customers to develop targeted tariffs for individual groups and alleviate the problem of volatility in production by capitalizing on the flexibility of consumers. Beyea (2010) has pointed out, there has been little discussion or exploration of the full potential of these data bases and their benefits can reach beyond the original intentions for collecting these data. Thus, there is a scope to investigate and analyze these data in various ways for a greater understanding of consumption patterns and how they correlate with other economic, physical or geographical factors. In this work, we are interested to see how we can utilize this dataset to group different customers with similar periodic behavior. Towards this goal, this chapter aims to: (a) describe the contents of the data set in SGSC database that we can utilize, and (b) propose a clustering algorithm to group customers with similar periodic behaviors. The distance metric introduced in Chapter 2 will be the inputs for this cluster analysis. One of the advantages of using our approach is that the technique is based on probability distributions instead of raw data. Many clustering approaches are limited by the type of noisy, patchy, and unequal time-series common in residential data sets. Since the distance measure considered is based on differences in probability distribution of time series, it is likely to be less sensitive to missing or noisy data.

Themes

- Dimension reduction: If each $P_{i,j,k}$ be considered to be a point in the space, key i would have mp dimensions as opposed to n_i dimensions in case of considering raw data. Hence for a large number of observations ($n_i >> mp$), this approach benefits by transitioning to a lower dimension.

- Avoid loss of information due to aggregation: This approach ensures key characteristic information of the data is not lost due to averaging or aggregation measures in an attempt to transition to a lower dimension. Hence, this approach could be thought

to somehow balance the drawback of considering raw data or aggregated data.

- Robustness to outliers: This approach could be adapted to be robust to outliers and extreme behaviors by trimming the tails of the probability distributions.
- Non-synchronized observed time periods: Considering probability distribution would imply the clustering process can handle keys that are observed over periods of time that are overlapping but don't necessarily coincide.
- Similar periodic behavior: Since cyclic granularities are considered instead of linear granularities, clustering would group keys that have similar behavior across these cyclic granularities. This implies they will be grouped according to their periodic behavior and not on the linear stretch of time over which they are observed.

The following contributions are made through the following chapter:

- Present a cluster analysis of SGSC dataset to group households with similar periodic behavior
- Cluster validation by relating to external data

2 Clustering methodology

The data set solely contains readings from smart meters and no information about the consumers' specific physical, geographical, or behavioural attributes. As a result, no attempt is made to explain why consumption varies. Instead, this work investigates how much energy usage heterogeneity can be found in smart meter data and what some of the most common electricity use patterns are. It is worth noting that when studying these dynamics, a variety of objectives may be pursued. One objective could be to group consumers 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. Most clustering algorithms offer only daily energy profiles throughout the hours of the day, but we suggest a broader approach to the problem, aiming to group consumers with similar shapes 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.

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 time-series features. First, we study the underlying distributions that may have resulted in different patterns across temporal granularities in order to identify a mechanism to classify them based on the similarity of those distributions. 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*

? introduced the tidy "tsibble" data structure to support exploration and modeling of temporal data. A tsibble comprises an index, optional key(s), and measured variables. An index is a variable with inherent ordering from past to present and a key is a set of variables that define observational units over time. A linear granularity is a mapping of the index set to subsets of the time domain. For example, if the index of a tsibble is days, then a linear granularity might be weeks, months or years. 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. [XXX reference 2nd chapter] 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 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 each 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 observations 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:

Inter-category 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

$$S^A_{x,y} = \sum_j D_{x,y}(A)$$

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

$$S^{A*B}_{x,y} = \sum_j \sum_k D_{x,y}(A, B)$$

(sum of distances between each combination of categories (j, k) of the harmony (A, B). When combining distances from individual cyclic granularities A and B,

$$S_{x,y}^{A,B} = S_{x,y}^{A}/J + S_{x,y}^{B}/K$$

is used, which could also be shown to be a distance metric easily. This is shown for cyclic granularity A and B, but could be practically extended to more granularities.

Intra-category distances

Compute weighted pairwise distances (*wpd*) (XXX reference) 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 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 objects 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*

  Handling trend, seasonality, non-stationarity and auto-correlation: Trend and seasonality are fundamental characteristics of time series data, and it is reasonable to define a time series according to its degree of trend and seasonality. These characteristics of the time series are lost or handled independently by considering probability distributions (trend is lost) across categories of cyclic granularities (by independently modeling all seasonal fluctuations), and so there is no need to de-trend or de-seasonalize the data before conducting the clustering method. There is no need to omit holiday or weekend patterns for similar reasons.

  Data transformation: Robust scaling method is used before computing the Inter-category distances and NQT is built-in transformation used for computation of *wpd*,

which forms the basis of Intra-category distances.

- *Clustering algorithm*

In the analysis of energy smart metre 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). Individual entities with the highest similarity computed using the desired distance metrics are sequentially merged using agglomerative algorithms. We can possibly employ any clustering method that supports the given distance metric as input.

- *Characterization of clusters*

Depending on the distance measure utilized for the study, the cluster characterization technique will differ. Clusters that utilise intra-category distances are characterised using multi-dimensional scaling and parallel coordinate displays. For inter-category 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.
Multidimensional scaling (MDS) (XXX reference) refers to a family of methods that analyse a matrix of distances or dissimilarities to provide a representation of the data points in a reduced-dimension space. There are many kinds of MDS, but they all solve the same fundamental issue: Given a $n*n$ matrix of dissimilarities and a distance measure, identify a configuration of n points $x_1, x_2, \ldots, x_n$ in the reduced dimension space R^q $(q < p)$ where the distance between the points is near to the dissimilarity between the points. All techniques must determine the coordinates of the points as well as the space dimension, q. Metric and nonmetric MDS are the two main kinds of MDS. Metric MDS methods presume a functional connection between the interpoint distances and the supplied dissimilarities and assume that the data are quantitative. We use metric MDS.
Parallel coordinate plots (XXX reference) Parallel coordinates have been extensively used to display high-dimensional and multivariate data, allowing for the detection of patterns

within the data via visual grouping.

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 features 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. A continuous measured variable y of length T indexed by $0, 1, \ldots T-1$. Three circular granularities $g1$, $g2$ and $g3$ are considered with 2, 3 and 5 levels respectively. Categories of $g1$, $g2$ and $g3$ are represented by $g10, g11$, $g20, g21, g22$ and $g30, g31, g32, g33, g34$. 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 $Mon, Tue, \ldots, Sun$. 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 method could preform differently with different jumps between consecutive categories. So a difference of $diff = \{1, 2, 5\}$ for corresponding categories are also considered. The code for creating these designs can be found in the Supplementary section (link to github repo). The results for $T = 300$ and $R = 25$ is shown, that means we have 25 time series each with length 300. The rest of the results could be found in the supplementary paper.

3.1 Data generating processes

Each of the categories $g10, g11$, $g20, g21, g22$, and $g30, g31, g32, g33, g34$ might be 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 ignores the linear progression of time, there is little value in adding time dependency in the data generating process. Time series data include basic properties such as trend, seasonality, and auto-correlation, and it is reasonable to construct

Table 1: Alternate distributions of different categories if they deviate from null.

granularity	alternate distribution
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)

a time series using these features. However, when examining distributions across categories of cyclic granularities, these time series features are lost or addressed independently by considering all seasonal fluctuations. Because the time span during which an entity is observed in order to ascertain its behavior is very short, it is assumed that the time series will remain stationary throughout the observation period. The data type is set to "continuous," and the setup is assumed to be gaussian. For example, in a null design, all categories are considered to be normal (0,1). The mean of the subsequent categories are incremented or decremented from this distribution based on the mean difference between them in non-null designs.

3.2 Data designs

3.2.1 Individual granularities

Three signifiant granularities

Consider the scenario when all three granularities $g1$, $g2$, and $g3$ are responsible for distinguishing the designs. That means that for at least one among the to-be-grouped designs, the pattern for each of $g1$, $g2$, and $g3$ will change. We consider different distributions across categories (as in Table 1) that will lead to different designs (as in Table 2). Figure **??** shows the linear and cyclic representation of the simulated variable under these five designs. For the consequent data designs, only graphical representations will be provided.

Two significant granularities

This is the case where one granularity will remain the same across all designs. We consider the case where the distribution of y 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

Table 2: 5 different designs resulting from considering different distributions across categories.

design	g1	g2	g3
design-1	null	null	null
design-2	alternate	null	null
design-3	null	alternate	null
design-4	null	null	alternate
design5	alternate	alternate	alternate

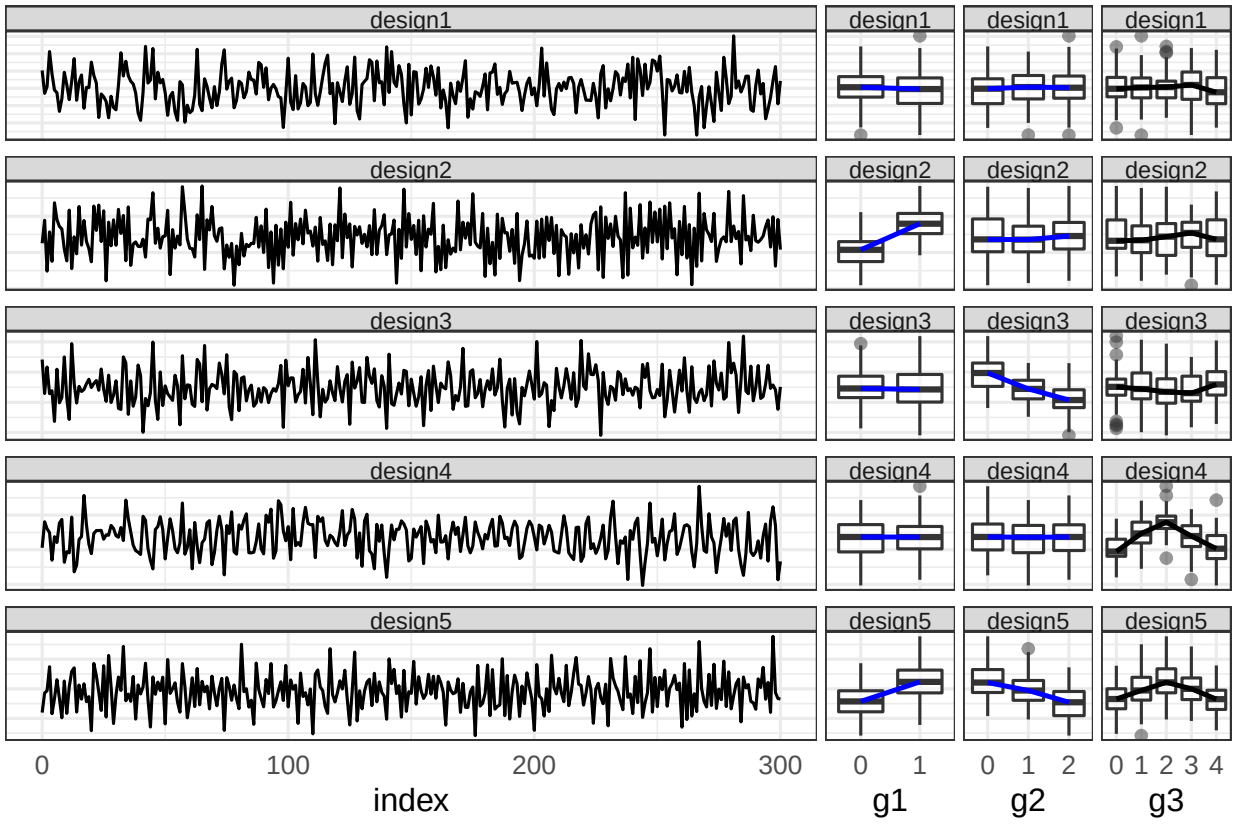


Figure 1: The linear (left) and cyclic (right) representation of the measured variable is shown. In this scenario, all of $g1$, $g2$ and $g3$ changes across at least one design. Also, it is not possible to comprehend these patterns across cyclic granularities or group similar series just by looking at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

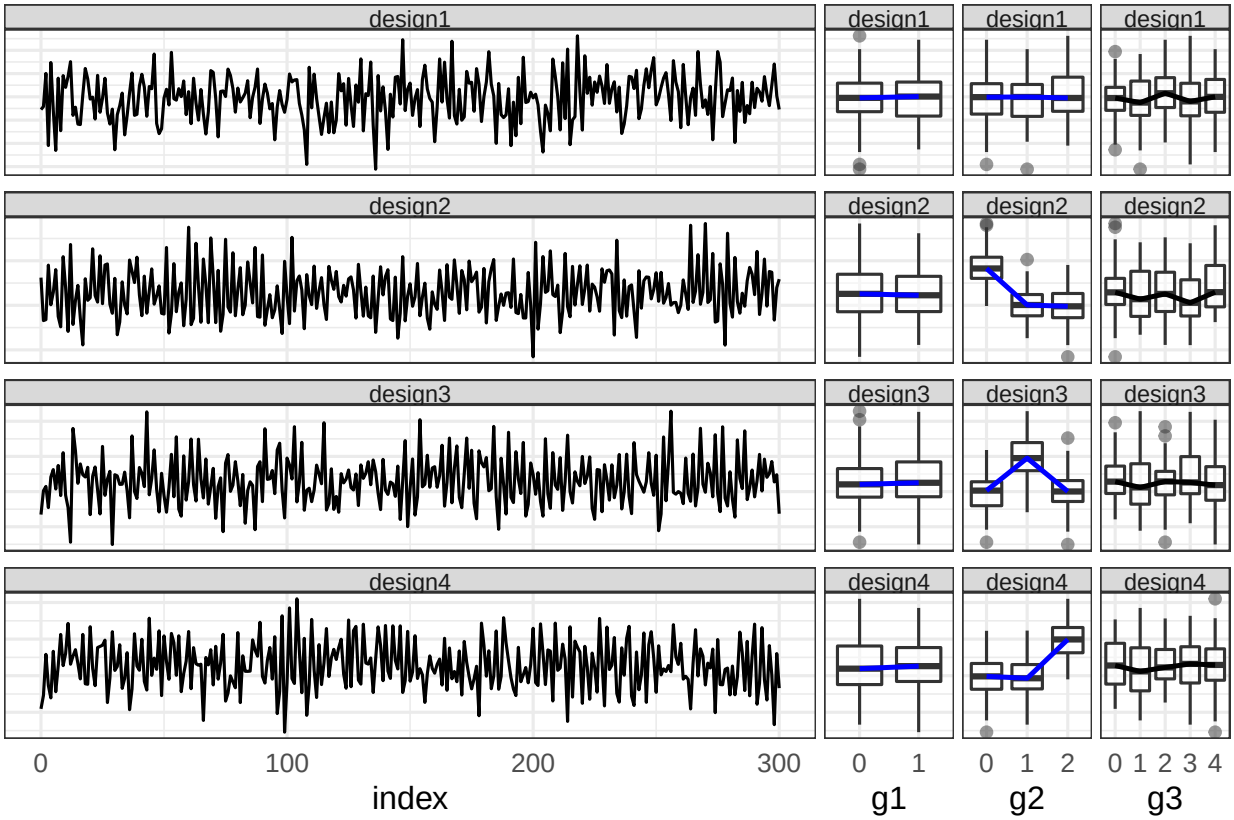
(#fig:plot-3gran)

responsible for distinguishing across designs. Figure **??** shows the linear and cyclic representation of y. The first panel shows raw plot of y in a linear scale and the second panel shows distribution of y across cyclic granularities namely $g1$, $g2$ and $g3$. As could be seen from the plots, it is impossible to decipher from the raw time plot that the time series variable shows such pattern across different granularities.

One signifiant granularity

Here only one granularity is responsible for distinguishing the designs. Designs change significantly only for the granularity $g2$. Figure **??** shows this.

```
## # A tibble: 1,204 x 6
##    index    g1    g2    g3 design  sim_data
##    <dbl> <dbl> <dbl> <dbl> <chr>      <dbl>
##  1     0     0     0     0 design1  -0.547
##  2     0     0     0     0 design2   1.55
##  3     0     0     0     0 design3  -1.31
##  4     0     0     0     0 design4  -1.81
##  5     1     1     1     1 design1  -0.419
##  6     1     1     1     1 design2  -0.856
##  7     1     1     1     1 design3   0.0464
##  8     1     1     1     1 design4  -1.08
##  9     2     0     2     2 design1   1.01
## 10     2     0     2     2 design2  -0.804
## # ... with 1,194 more rows
```



A subset of many possible designs are shown in Figure **??**. For the parameter space (XXX unique combinations shown in table YYY), 100, 500 independent replications of all possible combination of simulation parameters were generated. The clustering methodologies were run all these unique combinations and subsets of these to verify if the methodologies work as expected.

3.2.2 Interaction of granularities

When two granularities of interest interact, the connection between a granularity and the measured variable is determined by the value of the other interacting granularity. This happens when the effects of the two granularities on the measured variable are not additive. Consider a case with just two interacting granularities $g1$ and $g3$ of interest. As opposed to the last case, where we could play with the distribution of $(2 + 5)$ categories, with interaction we can play with the distribution of $(2 * 5)$ combination of categories.

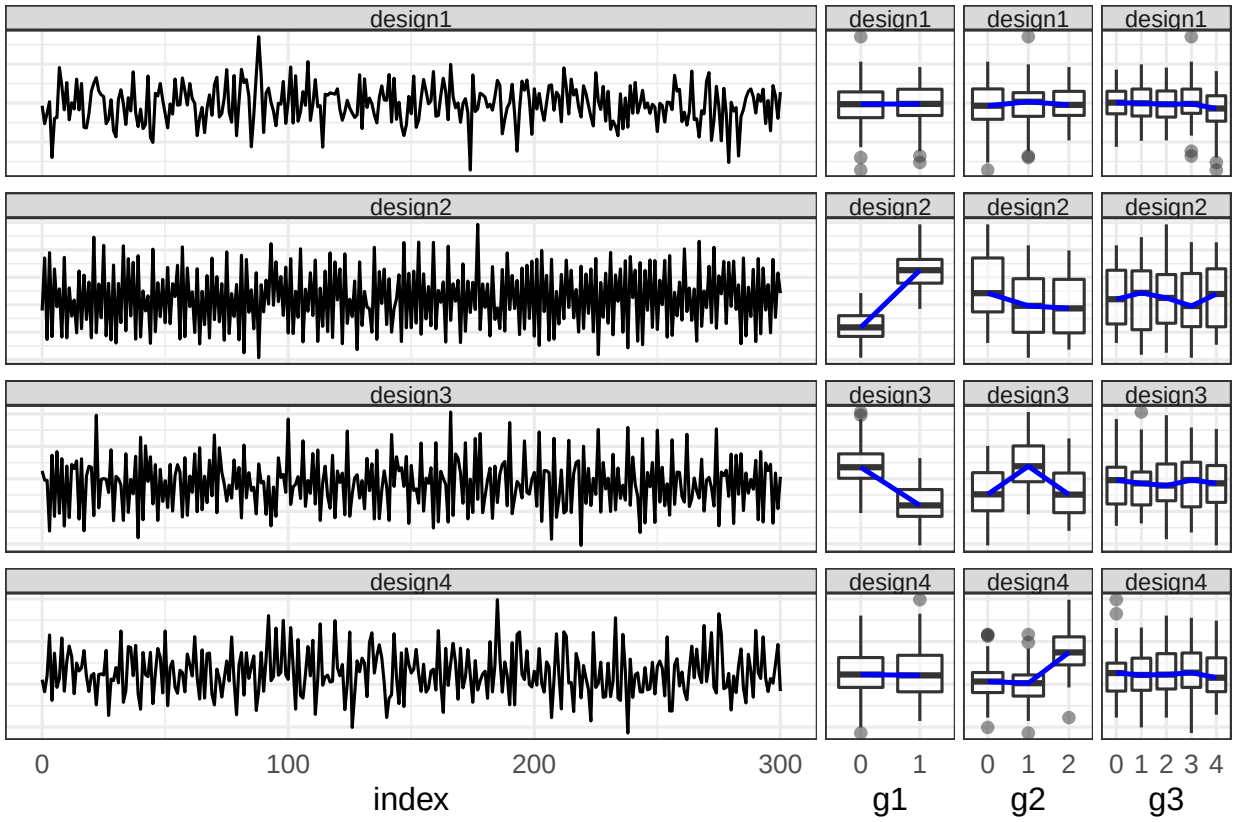


Figure 2: The linear (left) and cyclic (right) representation of the measured variable is shown. In this scenario, $g1$, $g2$ would change across atleast one design but $g3$ change remains same across all design. Thus $g3$ is not an important variable to distinguish these designs.

Granularity type	# Significant	# Replications
Individual # obs: 300, 500, 2000 # clusters: 6/7	1/2/3	25, 100, 200
Interaction # obs: 500, 2000 # clusters: 4	1/2	25, 100, 200

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 difference=1, the performances are pretty bad for js-robust methods and wpd method. The performance starts getting better with increasing difference and get worse with increasing number of replications. Length of series do not show to have any effect on the performance of the methods. It does not depend on if time series is ar or arma.

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 3, 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

The aim of this section is to illustrate that the proposed methodology can be used to understand repetitive behavior for several customers together.

First, we select the customers which do not have any implicit missing values and filtered their data for the year 2013. From this set, we randomly sample a set of 600 customers. Then we removed customers for which any of the hod (hour-of-day), moy (month-of-year) and wkndwday (weeknd/weekday) is a clash. We further removed 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 wanted to do instance selection. Since this is unlabeled data, there is no way to do external validation of our methodologies. Thus, we chose this way to see how well our methodology works in a cleaner data set as this one to classify customers who have similar behavior across all

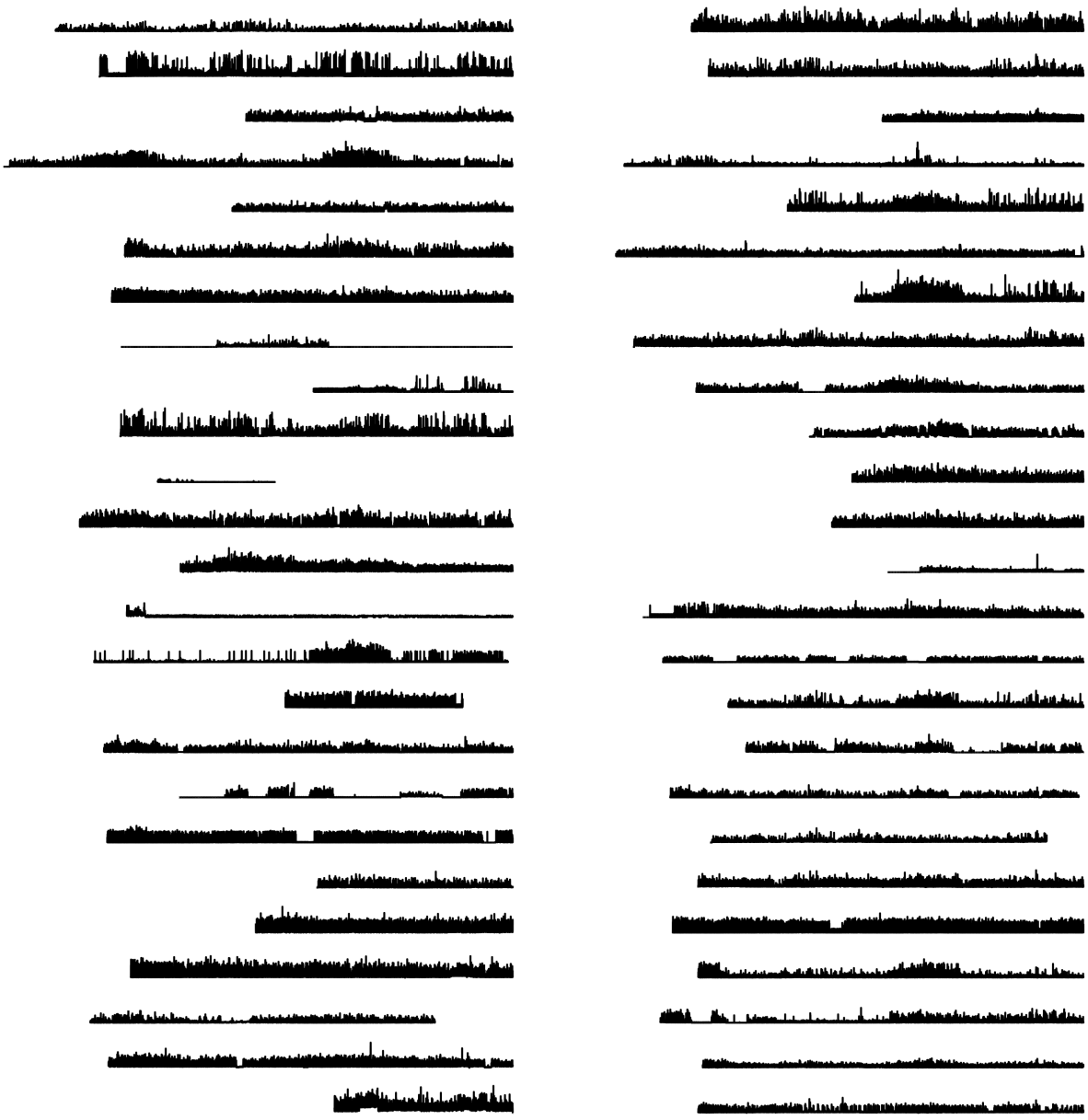


Figure 3: The raw data for 50 households are shown. It looks like there is a lot of missing values and unequal length of time series along with asynchronous periods for which data is observed. No insightful behavioral pattern could be discerned from this view other than when the customer is not at home.

significant granularities. Next we select 4 customers which are far apart from each other and 5 customers which are lying closest in distance to each of these 4 customers. These selection was done using the granularity *hod*. When we use our methodologies, it is based on all significant granularities and not only hod, but if *hod* is significant we can expect that the grouping would be similar to how we have initially chosen the set of customers. This procedure is analogous to instance-wise classfication methods that selects similar instances (neighbours) for each instance (anchor) and treats the anchor and its neighbours as the same class.

4.2 js-based clustering

The distribution of electricity demand for the selected 24 customers across hour-of-day is shown in 4. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. According to our prototype selection method, each row represents a distinct shape of daily load, and all customers in the same row should have similar daily profile.After clustering these consumers, all customers with the same colour represent the same group. Except for customer id 8269176, there is unanimity across design and groups. Even though their daily form resembles Group 2, our clustering approach places them in Group 3 (which is design 4 in our case). Because our method uses hod, moy, and wkndwday, there may be some mismatches depending on only one variable.
The distribution of electricity demand for the selected 24 customers across month-of-year is shown in Figure 5. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. Intriguingly, customer id 8269176 appears to be in the appropriate group (Group 3) because its monthly profile is more similar to Group 3 than Design 4. So, while our clustering approach failed to place the customer in the correct hour-of-day group, it did so for month-of-year. When contrasting the moy to hod profile, there are greater behavioural differences across customers within a group.

Characterization of clusters is the final stage of a cluster analysis. If we are convinced that we have identified a collection of clusters that can be distinguished from one another, we should intend to characterize them more formally, both statistically and qualitatively. We quantify them by producing statistics for each variable. We may look at the findings

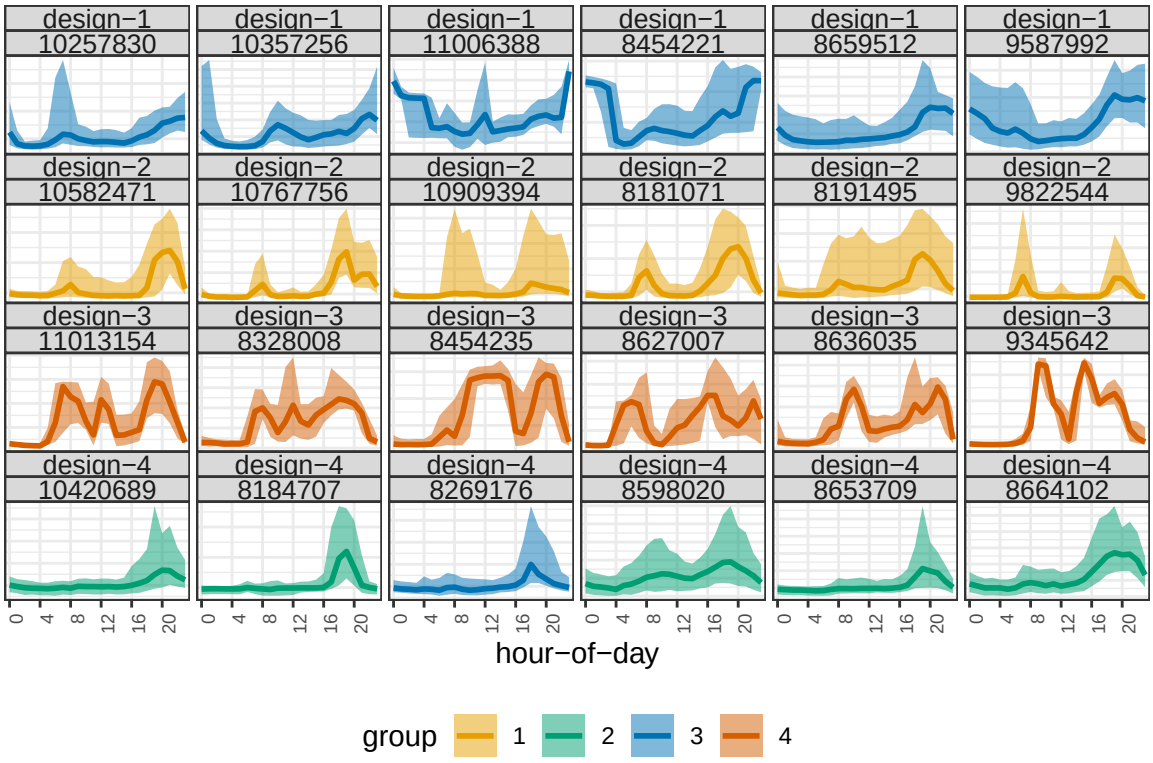


Figure 4: The distribution of electricity demand for the selected 24 customers across hour-of-day. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. According to our prototype selection method, each row represents a distinct shape of daily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

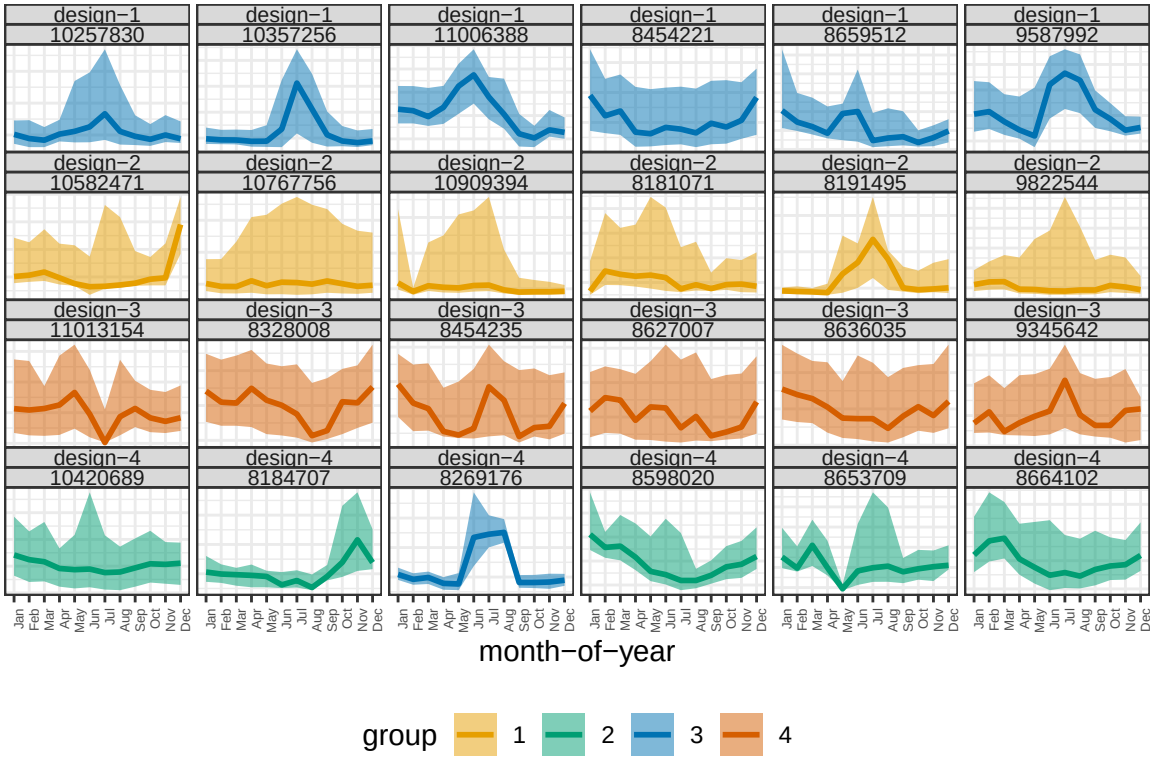


Figure 5: The distribution of electricity demand for the selected 24 customers across month-of-year. The median is represented by a line and shaded region represents the area between 25^{th} and 75^{th} percentile. Intriguingly, customer id 8269176 appears to be in the appropriate group (Group 3) because its monthly profile is more similar to Group 3 than Design 4. So, while our clustering approach failed to place the customer in the correct hour-of-day group, it did so for month-of-year. When contrasting the moy to hod profile, there are greater behavioural differences across customers within a group.

in graphs, and enhance our qualitative descriptions of the groupings.

We discover 4 qualitative clusters of varying shapes in the distribution of all consumers in the first panel of Figure 6. Group-1 includes consumers who work 9-5, get up and conduct morning activities from 7-10am, and then depart. Then they return home in the evening to cook supper and perform other activities, giving the evening a greater peak than the morning. Group 2 is the group that rushes out of the house in the morning to get to work. They only return at night and do all activities at night, so there is no morning peak. The third one has a strong early morning and late night hours. These consumers may be flexible students or elderly retirees who are night owls. Presence of children or stay-at-home parents is indicated by Group-4's almost equivalent morning, afternoon and evening profile. All of this may be validated with further information about the customer. The second panel of Figure 6 shows that month-of-year qualitative clusters are not as distinguishable as hour-of-day. Group 2 is the most distinct and uses the most power during the summer, possibly owing to the use of air conditioners. Group 4 has a flat profile, indicating no significant month-to-month changes. Groups 1 and 3 have heaters on in the winter but consume less energy in the summer. Since gas is not available in all of NSW LGAs, it is possible that customers' heater usage is recorded in electricity rather than gas.

The third panel of Figure 6 shows that the wknd-wday groups exhibit no significant changes across clusters, indicating that they may be a nuisance variable for these consumers.

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

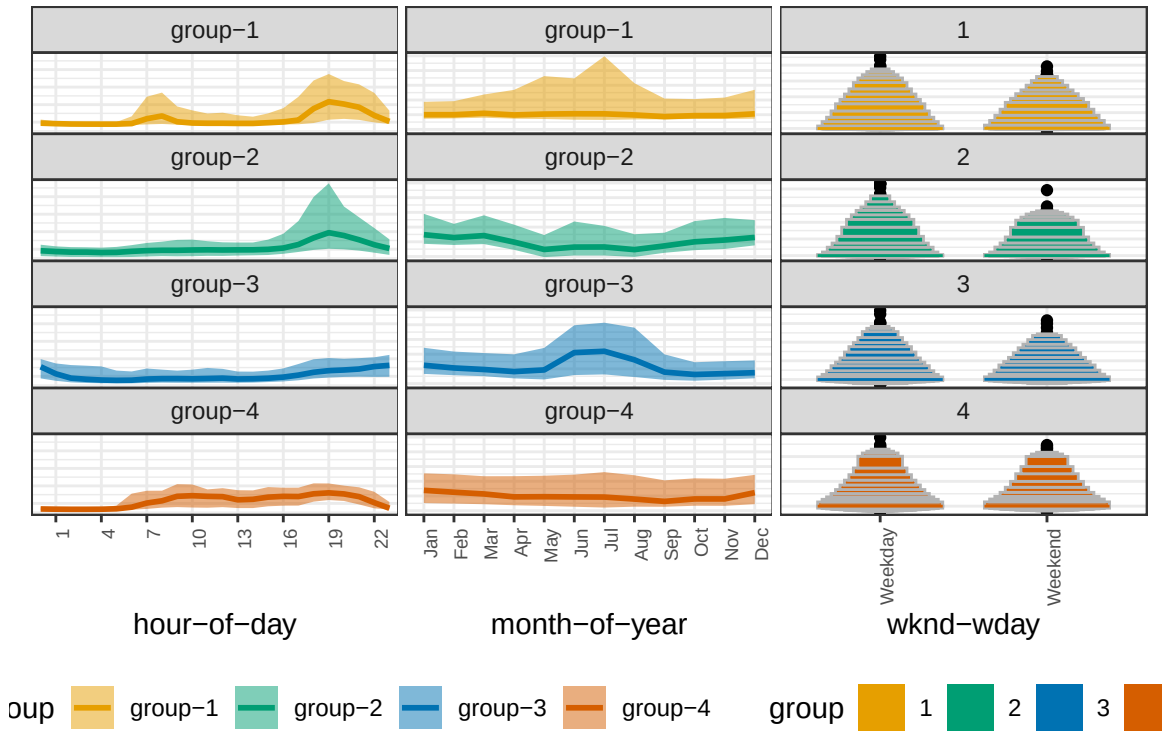


Figure 6: The distribution of electricity demand for the clusters across hour-of-day. The median is represented by a line and the shaded region represents the area between 25^{th} and 75^{th} percentile. Each cluster is characterised by unique shape across the granularity it is plotted against. For wknd-wday differences across different groups are not distinct suggesting that it might not be that important a variable to distinguish different clusters. This fact we will be re-established when we see the importance of each granularity throough the parallel coordinate plot.

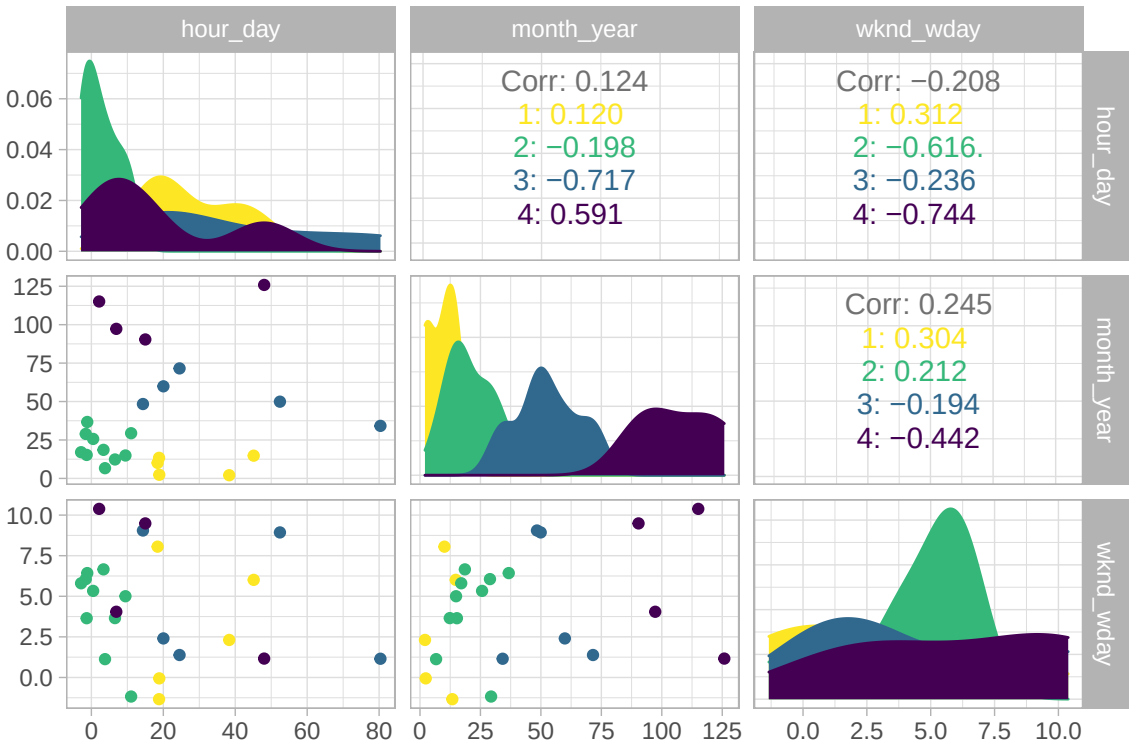
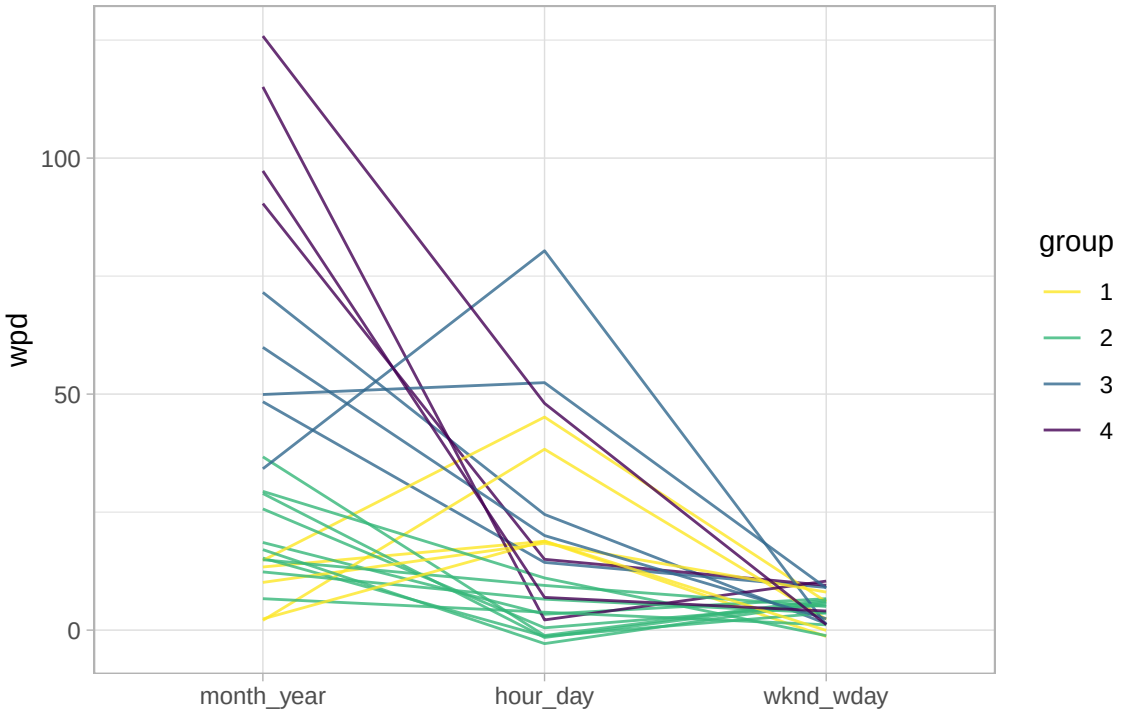


Figure 7: A ggpairsplot and parallel coordinate plot are used to depict each of the 24 customers. The ggpairs plot four distinct clusters across the month-of-year, which are less prominent across the hour-of-day and wknd-wday. The parallel coordinate plot ranks the variables in order of importance, indicating that the month-of-year is the most important in identifying clusters, whereas wkdn-wday is the least significant and has the least variability among the three variables.

4.3 wpd-based clustering




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

We discover 4 qualitative clusters of varying shapes in the distribution of all consumers in the first panel of Figure 6. Group-1 includes consumers who work 9-5, get up and conduct morning activities from 7-10am, and then depart. Then they return home in the evening to cook supper and perform other activities, giving the evening a greater peak than the morning. Group 2 is the group that rushes out of the house in the morning to get to work. They only return at night and do all activities at night, so there is no morning peak. The third one has a strong early morning and late night hours. These consumers may be flexible students or elderly retirees who are night owls. Presence of children or stay-at-home parents is indicated by Group-4's almost equivalent morning, afternoon and evening profile. All of this may be validated with further information about the customer.

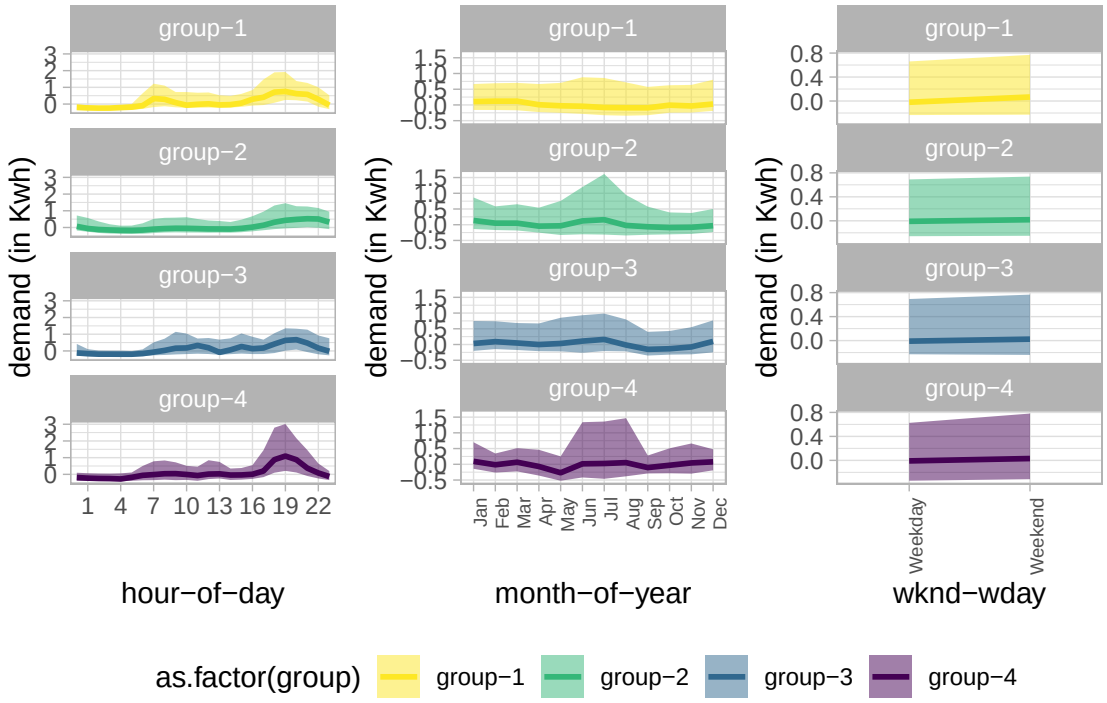


Figure 8: The distribution of electricity demand for the clusters across hour-of-day. The median is represented by a line and the shaded region represents the area between 25^{th} and 75^{th} percentile. Each cluster is characterised by unique shape across the granularity it is plotted against. For wknd-wday differences across different groups are not distinct suggesting that it might not be that important a variable to distinguish different clusters. This fact we will be re-established when we see the importance of each granularity throough the parallel coordinate plot.

The second panel of Figure 6 shows that month-of-year qualitative clusters are not as distinguishable as hour-of-day. Group 2 is the most distinct and uses the most power during the summer, possibly owing to the use of air conditioners. Group 4 has a flat profile, indicating no significant month-to-month changes. Groups 1 and 3 have heaters on in the winter but consume less energy in the summer. Since gas is not available in all of NSW LGAs, it is possible that customers' heater usage is recorded in electricity rather than gas.

The third panel of Figure 6 shows that the wknd-wday groups exhibit no significant changes across clusters, indicating that they may be a nuisance variable for these consumers.

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

5 Discussion

References