Selecting and ranking interesting pairs of cyclic temporal granularities

Contents

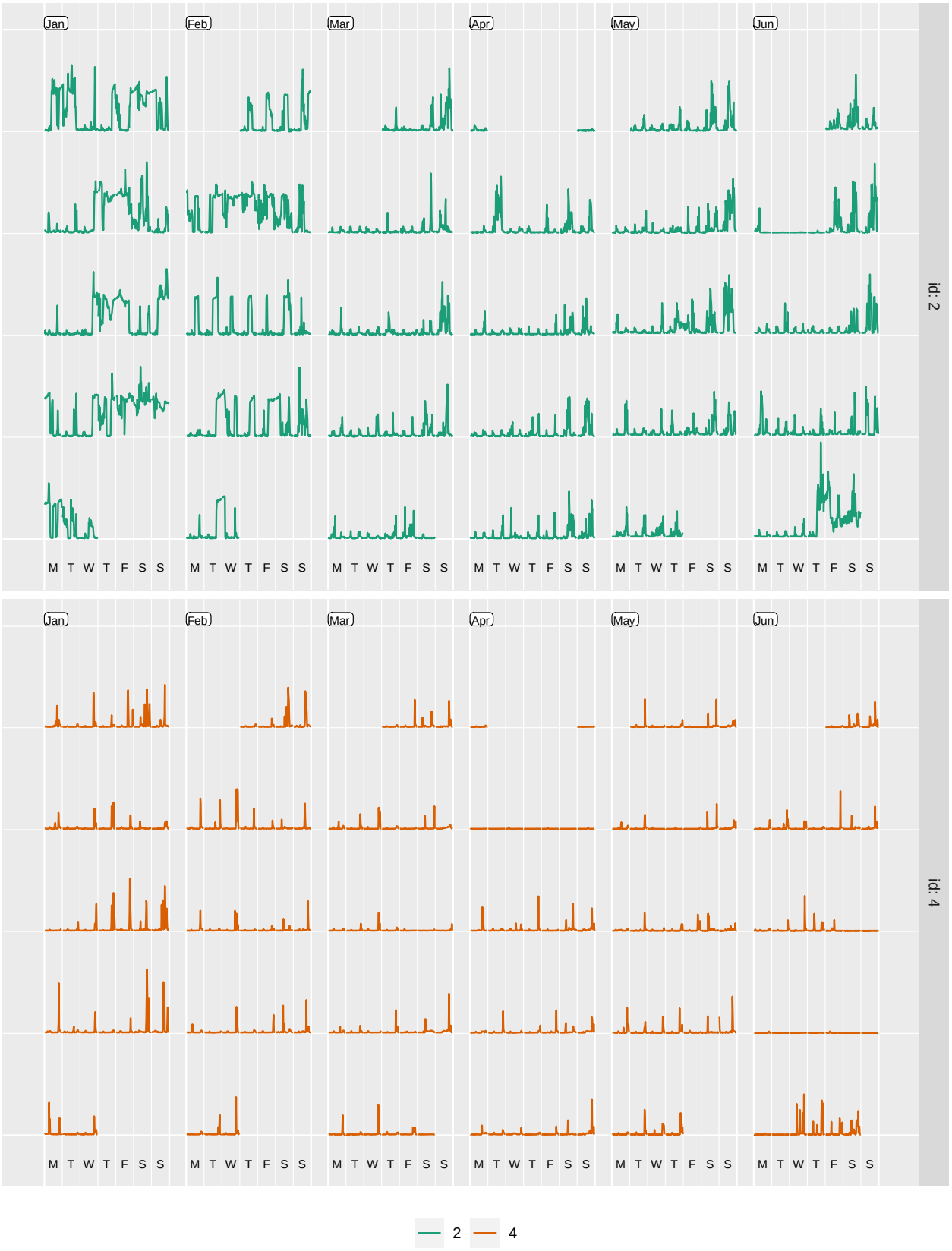


Figure 1: Calendar display.

1 Introduction

Exploratory data analysis, as coined by John W. Tukey (Tukey 1965) involves many iterations of finding structures and patterns that allows the data to be informative. With temporal data available at finer scales, exploring periodicity and their relationships can become overwhelming with so many possible cyclic temporal granularities (Gupta et al. 2020) to explore.

Take the example of the calendar display of electricity smart meter data (1) used in Wang, Cook, and Hyndman (2020) for four households in Melbourne, Australia. The authors show how hour-of-the-day interact with weekday and weekends and then move on to use calendar display to show daily schedules. The calendar display has several components in it, which helps us look at energy consumption across hour-of-the-day, day-of-the-week, week-of-the-month, and month-of-the-year at once. Some interaction of these cyclic granularities could also be interpreted from this display. This is a great start to have an overview of the energy consumption. However, if one wants to understand the periodicities in energy behavior and how the periodicities interact in greater details, it is not easy to comprehend the interactions of some periodicities' from this display, due to the combination of linear and cyclic representation of time. For example, this display might not be the best to understand how hour-of-the-day varies and month-of-year varies across week-of-the-month. Further, it is not clear what all interactions of cyclic granularities should be read from this display as there could be many combinations that one can look at. Moreover, calendar effects are not restricted to conventional day-of-week or month-of-year deconstructions (Gupta et al. (2020)) and could include other cyclic granularities like hour-of-week or day-of-fortnight, which could potentially become useful depending on the context.

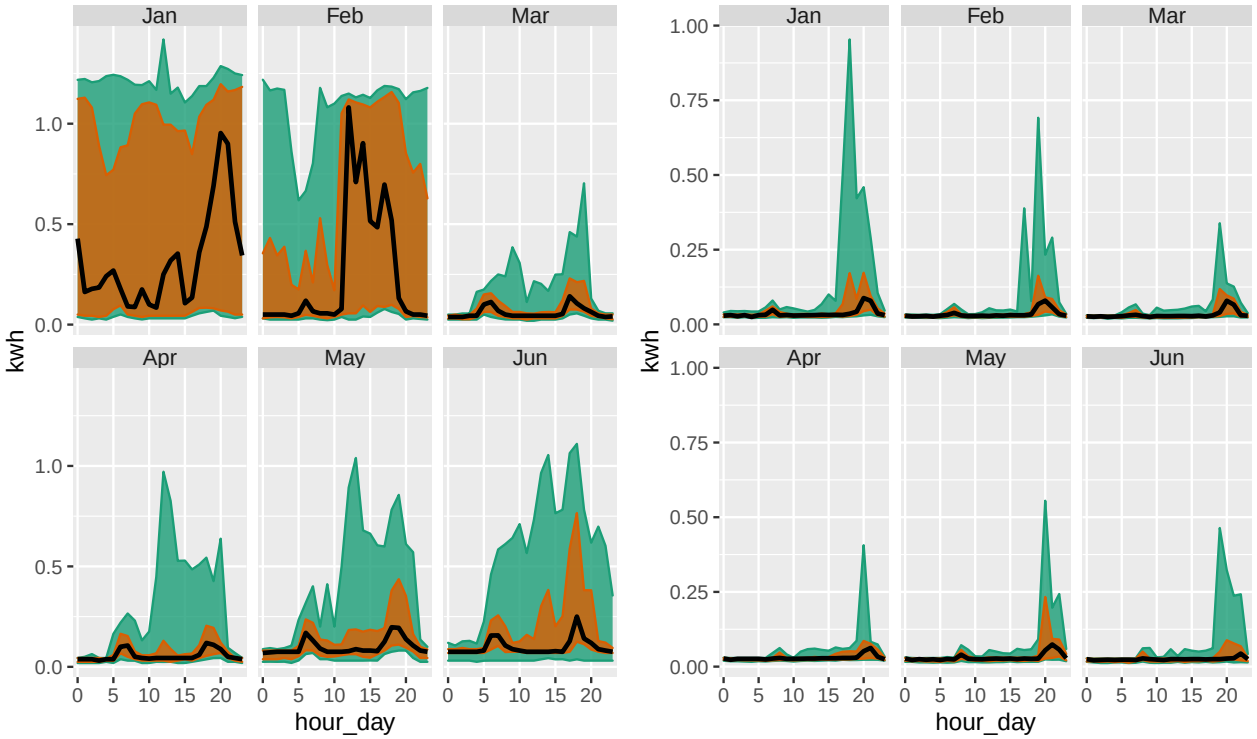


Figure 2: something

Moreover, there might be specific interactions that are interesting and others that are not and that too will vary with different households. For example, area distribution quantiles are plotted for household 2 and 4 in Figure 2a and b respectively. For the first household, the 75th and 90th percentile for Jan, Feb and July are very close, implying that energy usage for these months are generally on a much higher side due to the usage of air conditioners (in Jan and Feb) and heaters (in July). The energy consumption for household 2 is also

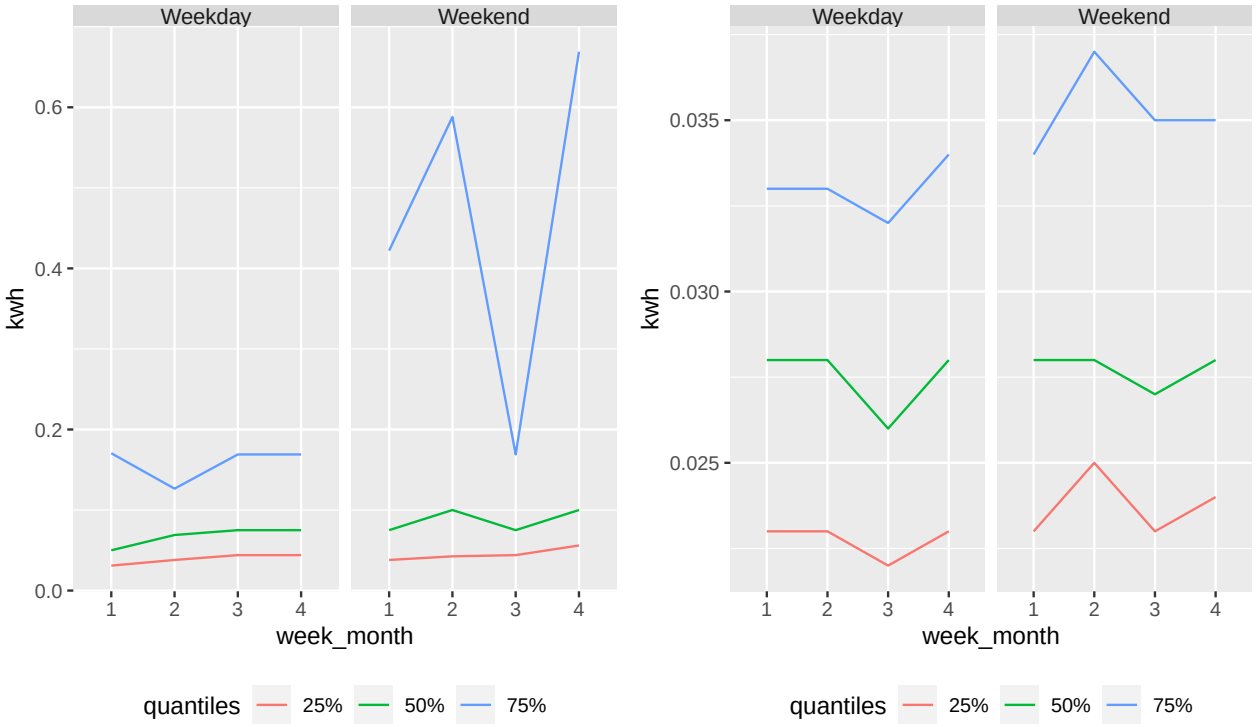


Figure 3: something2

higher relative to its own consumption for Jan, Feb and March but the 75th and 90th percentile are apart implying that contrary to the first household, the second household resorts to air conditioners and heaters much less regularly than the first one. Moreover, the 75th percentile distribution is not bimodal across hours of the day for the first household in those months, but the distribution looks similar for all months for the second household. Difference in the energy consumption seem to be varying both across month-of-year (facets) and hour-of-day (x-axis). And thus, both the cyclic granularities would deem important while studying the periodicities in the first household. However, it seems like energy consumption across hours of the day are not that different across different months for the second household. Differences seem to be more prominent across month-of-year (facets) than hour-of-day (x-axis). Again, look at **??** c and d, where energy consumption for these two households are plotted against (weekend/weekday, week-of-month). Here, for both households, the pattern of energy consumption vary across different weeks of the month irrespective of the fact it is a weekday or weekend. In that respect, the harmony pair (month-of-year, hour-of-day) seems to be more informative than (weekend/weekday, week-of-month) for the first household. It could be immensely useful to make the transition from all possible ways to only ways that could potentially be informative given a household.

The paper Gupta et al. (2020) describes how we can compute all possible combinations of cyclic time granularities. If we have n periodic linear granularities in the hierarchy table, then $n(n-1)/2$ circular or quasi-circular cyclic granularities could be constructed. Let N_C be the total number of contextual circular, quasi-circular and aperiodic cyclic granularities that can originate from the underlying periodic and aperiodic linear granularities. The mapping of the graphical elements chosen in the paper implies that, for a numeric response variable, the graphics display distributions across combinations of cyclic granularities, one placed at x-axis and the other on the facet. That essentially implies there are ${}^{N_C}P_2$ possible pairwise plots exhaustively, where each plot would display a pair of cyclic granularities. This is large and overwhelming for human consumption.

This is similar to Scagnostics (Scatterplot Diagnostics) by Tukey and Tukey (1988), which is used to discern meaningful patterns in large collections of scatterplots. Given a set of v variables, there are $v(v-1)/2$ pairs of variables, and thus the same number of possible pairwise scatterplots. Therefore for even small v, the

number of scatterplots can be large, and scatterplot matrices (SPLOMs) could easily run out of pixels when presenting high-dimensional data. Dang and Wilkinson (2014) and Wilkinson, Anand, and Grossman (2005) provides potential solutions to this, where few characterizations help us to locate anomalies for defining several measures aimed to detect anomalies in density, shape, trend, and other features in the 2D point scatters.

The paper (Gupta et al. (2020)) narrows down the search from ${}^{N_C}P_2$ plots by identifying pairs of granularities that can be meaningfully examined together (a "harmony"), or when they cannot (a "clash"). However, even after excluding clashes, the list of harmonies left could be enormous for exhaustive exploration. Hence, there is a need to reduce the search even further by including only those harmonies which are informative enough. Also, ranking the remaining harmony pairs based on how well they capture the variation in the measured variable could be potentially useful.

In this paper, we aim to build a new measure to follow through these two main objectives:

- To choose harmonies for which distributions of categories are significantly different
- To rank the selected harmonies from highest to lowest variation in the distribution of their categories.

2 The proposed distance measure

We are interested in assessing structure in probability distributions of the measured variable across bivariate cyclic granularities. We propose a measure called Weighted Maximum Pairwise Distances (wpd) to evaluate structure in such a design.

2.1 Idea

The principle employed for building a new metric is explained through a simple example explained in Figure 4. Each of these figures have the same panel design with 2 x-axis categories and 4 facet levels. Figure 4a has all x categories drawn from N(5, 10) distribution for each facet. It is not an interesting display particularly, as distributions do not vary across x-axis or facet categories. Figure 4b has x categories drawn from the same distribution within a facet and different for different facet categories. Figure 4b exhibits an exact opposite situation where distribution between the x-axis categories within each facet is different but they are same across facets. Figure 4d takes a step further by varying the distribution across both facet and x-axis categories. If we are asked to rank the displays in order of importance from minimum to maximum, we might order it as a, b, c and then d. It might be argued that it is not clear if b should precede or succeed c. Gestalt theory suggests that when items are placed in close proximity, people assume that they are in the same group because they are close to one another and apart from other groups. Hence, displays that capture more variation within different categories in the same group would be important to bring out different patterns of the data. With this principle, display b could be considered less informative as compared to display c.

With reference to the graphical design in **??**, therefore the idea would be to rate a harmony pair higher if the variation between different levels of the x-axis variable is higher on an average across all levels of the facet variables. Thus the metric could be obtained by computing maximum pairwise distances between distributions of the continuous random variable across x-axis categories for all facets and then taking the median of those maximum pairwise distances across facets. This would help capture the average maximum difference in distribution of the measurement variable explained by the two cyclic granularities together. We call this metric wpd which stands for Median Maximum Pairwise Distances. In the next section we shall see how we go about computing this measure.

2.2 Characterising distributions

Each of the data subsets in the data structure have multiple observations and may vary widely across different subsets due to the structure of the calendar, missing observations or uneven locations of events in

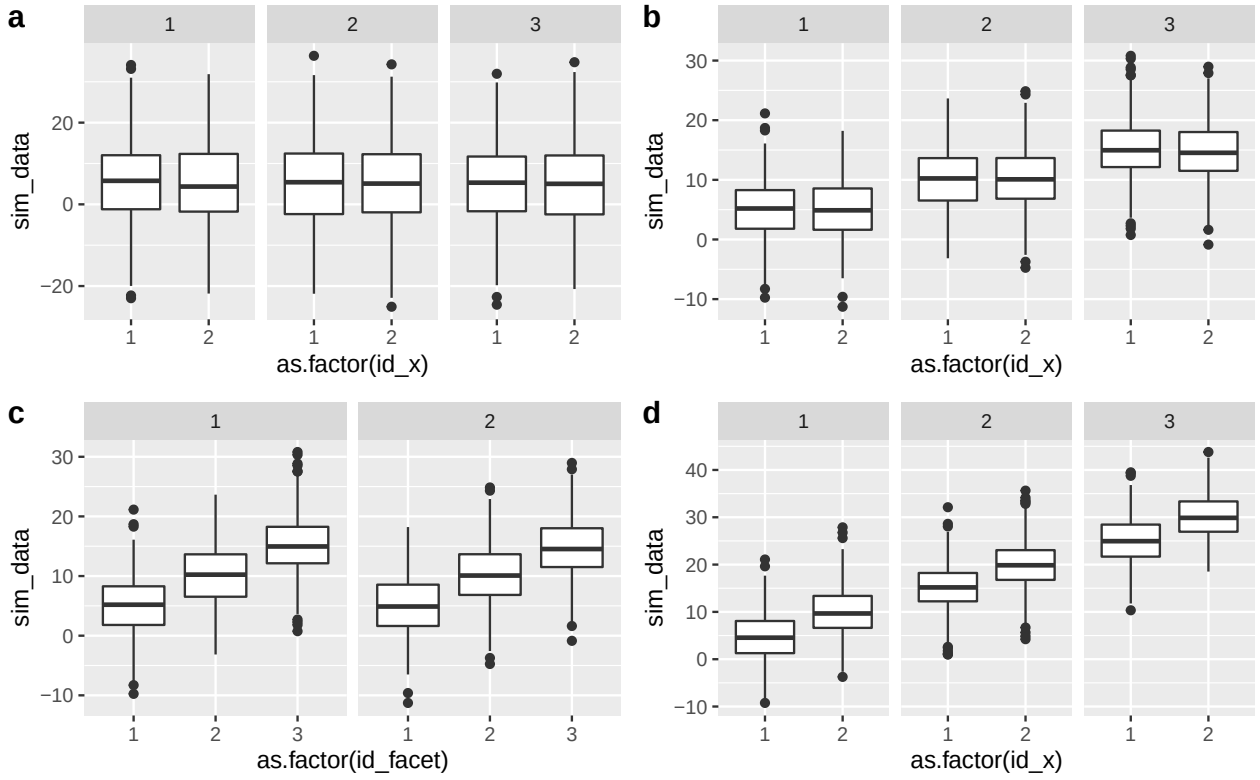


Figure 4: A graphical display with two categories mapped to x-axis and 4 categories mapped to facets with the distribution of a continuous random variable plotted on the y-axis. Display a is not interesting as the distribution of the continuous rv does not depend across x-axis or facet categories. Display b and c are more interesting than a since there is a change in distribution either across facets(b) or x-axis(a). Display d is most interesting as distribution of the rv changes across both facet and x-axis variable.

the time domain. The set of observations corresponding to each combination is assumed to be a sample from an unknown probability density function. While the whole population of observations has certain characteristics, we can typically never measure all of them. Often shape, central tendency, and variability are the common characteristics used to describe the distribution. Another way to describe the probability distribution is through quantiles. (Define quantiles here) Sample quantiles could be thought to estimate the population quantiles. But there are a large number of different definitions used for sample quantiles. The median-unbiased estimator is recommended (Rob's paper) because of its desirable properties of a quantile estimator and can be defined independently of the underlying distribution.

2.3 Distance between distributions

The most common divergence measure between distributions is the Kullback-Leibler (KL) divergence(Kullback and Leibler 1951) introduced by Solomon Kullback and Richard Leibler in 1951. The KL divergence, denoted $D(p(x), q(x))$ is a non-symmetric measure of the difference between two probability distributions $p(x)$ and $q(x)$ and is interpreted as the amount of information lost when $q(x)$ is used to approximate $p(x)$. Although the KL divergence measures the "distance" between two distributions, it is not a distance measure since it is not symmetric and does not satisfy the triangle inequality. The Jensen-Shannon divergence (Menéndez et al. 1997) based on the Kullback-Leibler divergence is symmetric and it always has a finite value. The square root of the Jensen-Shannon divergence is a metric, often referred to as Jensen-Shannon distance. Other common measures of distance are Hellinger distance, total variation distance and Fisher information metric.

In the context of this paper, the pairwise distances between the distributions of the measured variable are computed through Jensen-Shannon distance (JSD) which is based on Kullback-Leibler divergence and is defined by,

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

where $M = \frac{P+Q}{2}$ and $D(P||Q) := \int_{-\infty}^{\infty} p(x) f(\frac{p(x)}{q(x)})$ is the KL divergence between distributions $p(x)$ and $q(x)$. Probability distributions are estimated through quantiles instead of kernel density so that there is minimal dependency on selecting kernel or bandwidth.

2.4 Definition of the proposed distance measure

Consider two cyclic granularities A and B, such that $A = \{a_j : j = 1, 2, \ldots, J\}$ and $B = \{b_k : k = 1, 2, \ldots, K\}$ with A placed across x-axis and B across facets. Let the pairwise distances between pairs $(a_j b_k, a_{j'} b_{k'})$ be denoted as $d_{(jk,j'k')} = JSD(a_j b_k, a_{j'} b_{k'})$. Pairwise distances could be within-facets or between-facets. Figure 5 illustrates how the within-facet or between-facet distances are defined. Pairwise distances are within-facets (d_w) when $b_k = b_{k'}$, that is, between pairs of the form $(a_j b_k, a_{j'} b_k)$ as shown in panel (3) of Figure 5. If categories are ordered (like all temporal cyclic granularities), then only distances between pairs where $a_{j'} = (a_{j+1})$ are considered (panel (4)). Pairwise distances are between-facets (d_b) when they are considered between pairs of the form $(a_j b_k, a_j b_{k'})$.

From Section 2.1, the idea is to put more weights on within-facet distances than between-facet distances. Hence, for a suitable tuning parameter $\lambda > 1$, the pairwise distances $d_{(jk,j'k')}$ are transformed based on the distance type as follows:

$$d*_{(j,k),(j'k')} = \begin{cases} \lambda d_{(jk),(j'k')}, & \text{if } d = d_w \\ (1-\lambda) d_{(jk),(j'k')}, & \text{if } d = d_b \end{cases} \tag{1}$$

The maximum weighted pairwise distances are defined as:

$$WPD = max_{j,j',k,k'}(d*_{(jk),(j'k')}) \forall j, j' \in \{1, 2, \ldots, J\}, k, k' \in \{1, 2, \ldots, K\}$$

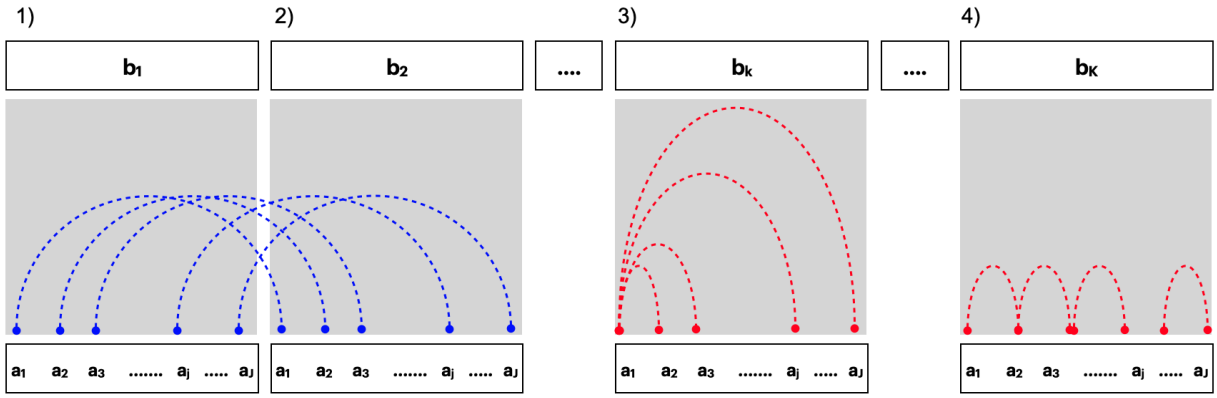


Figure 5: Within and between-facet distances shown for two cyclic granularities A and B, where A is mapped to x-axis and B is mapped to facets. The dotted lines represent the distances between different categories. Panel 1) and 2) show the between-facet distances. Panel 3) and 4) are used to illustrate within-facet distances when categories are un-ordered or ordered respectively. When categories are ordered, distances should only be considered for consecutive x-axis categories. Between-facet distances are distances between different facet levels for the same x-axis category, for example, distances between (a_1,b_1) and (a_1, b_2) or (a_1,b_1) and (a_1, b_3).

.

3 Normalisation

The distribution of wpd is different for different levels of facets and x-axis levels. This is because the statistic maximum which is used to define wpd is affected by the number of categories. The measure would have higher values if C_i or C_j has higher levels. However, we would ideally want a higher value of the measure only if there is significant difference between distributions across facet or x-axis categories, and not because the number of categories are higher. Therefore, in order to compare wpd across different combinations of facet and x-axis levels, we need to eliminate the impact of different levels of the facets and x-axis first and get a normalized measure. Henceforth we call the measure already discussed as wpd_{raw} and the normalised measure as wpd_{norm}. The measure wpd_{norm} could potentially lead to comparison of the measure across different panels and also identifying only the interesting panels from a data set. In the coming section, we have discussed two approaches to normalisation both through simulations.

3.1 Simulation study

Most of the behavior of the measure wpd was studied via simulation. The simulations explore how wpd performs under various designs and parameters and its limitations. To study the behavior of wpd, simulations were carried out for four different designs and the following factors that could potentially have an impact on the values of wpd:

- nx (number of levels of x-axis)
- $nfacet$ (number of levels of facets)
- λ (tuning parameter)
- ω (increment in each panel design)
- $dist$ (normal/non-normal distributions with different location and scale)
- n (sample size for each combination of categories)

- $nsim$ (number of simulations)a
- $nperm$ (number of permutations of data)

- *designs*
  D_{null} (No difference in distribution)
  D_{var_f} (Difference in distribution only across facets)
  D_{var_x} (Difference in distribution only across x-axis)
  $D_{var_{all}}$ (Difference in distribution in both facets and x-axis)

3.2 Simulation environment

R version 4.0.1 (2020-06-06) is used with platform: x86_64-apple-darwin17.0 (64-bit) running under: macOS Mojave 10.14.6 and MonaRCH, which is a next-generation HPC/HTC Cluster, designed from the ground up to address the computing needs of the Monash HPC community. The nodes and the storage used are as follows:

3.3 Methodology

3.3.1 Notations

Let $\{nx_i, i = 1, 2, \ldots, nx\}$, $\{nfacet_j, j = 1, 2, \ldots, nfacet\}$ be the set of x-axis and facet categories respectively. Each combination of nx_i and $nfacet_j$is being referred to as a *panel*. Then the total number of panel is $nx * nfacet$. Let the total number of pairwise distances that could result in each panel be $\{z_k, k = 1, 2, \ldots, nx * nfacet\}$. Here, $\{z_1 = nx_1 * nfacet_1\}$, $\{z_2 = nx_2 * nfacet_2\}$ and $\{z_k = nx * nfacet\}$. Now, let $\{x_{k,l}, k = 1, 2, \ldots, nx * nfacet, l = 1, 2, \ldots, nsim\}$ denote the values of wpd_{raw} obtained from the simulation study for k^{th} panel in the i^{th} simulation. Hence, for each of those k panel, we have $nsim$ values of wpd_{raw}.

3.3.2 Data generation

Observations are generated from a N(0,1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = nfacet = \{2, 3, 5, 7, 14, 20, 31, 50\}$ to cover a wide range of levels from very low to moderately high. That is, data is being generated for each of the panels $\{nx = 2, nfacet = 2\}, \{nx = 2, nfacet = 3\}, \{nx = 2, nfacet = 5\}, \ldots, \{nx = 50, nfacet = 31\}, \{nx = 50, nfacet = 50\}$. For each of the 64 panels, $ntimes = 500$ observations are drawn for each combination of the categories. That is, if we consider the panel $\{nx = 2, nfacet = 2\}$, 500 observations are generated for each of the combination of categories from the panel, namely, $\{(1, 1), (1, 2), (2, 1), (2, 2)\}$. The value of λ is set to 0.67. $nsim = 200$, which means each of these scenarios are run 200 times by permuting the data in each of the panels. And hence 200 values of wpd_{raw} are obtained for each of those 64 panels.

3.3.3 Permutation approach

The permutation approach ensures that the distribution of the normalised weighted pairwise distance measure has the same mean and standard deviation across all combinations of nx_i and $nfacet_j$. Thus, the normalised distances is computed as follows: $x_k^{norm} = (x_k - mean_l(x_{k,l}))/sd_l(x_{k,l})$, x_k^{norm} is the value of the wpd_{norm} for the k^{th} panel.

While this works successfully to make the mean and standard deviation across different nx and nfacet (as seen in Figure **??**), it is computationally heavy and time consuming, and hence less user friendly when being actually used in practice. Hence, we propose another approach to normalisation which is more approximate than exact but still has the same accuracy when compared to the permutation approach.

3.3.4 Modelling approach

A log-linear model is fitted to see how the values of wpd_{raw} changes with the values of nx and $nfacet$. The model is of the form

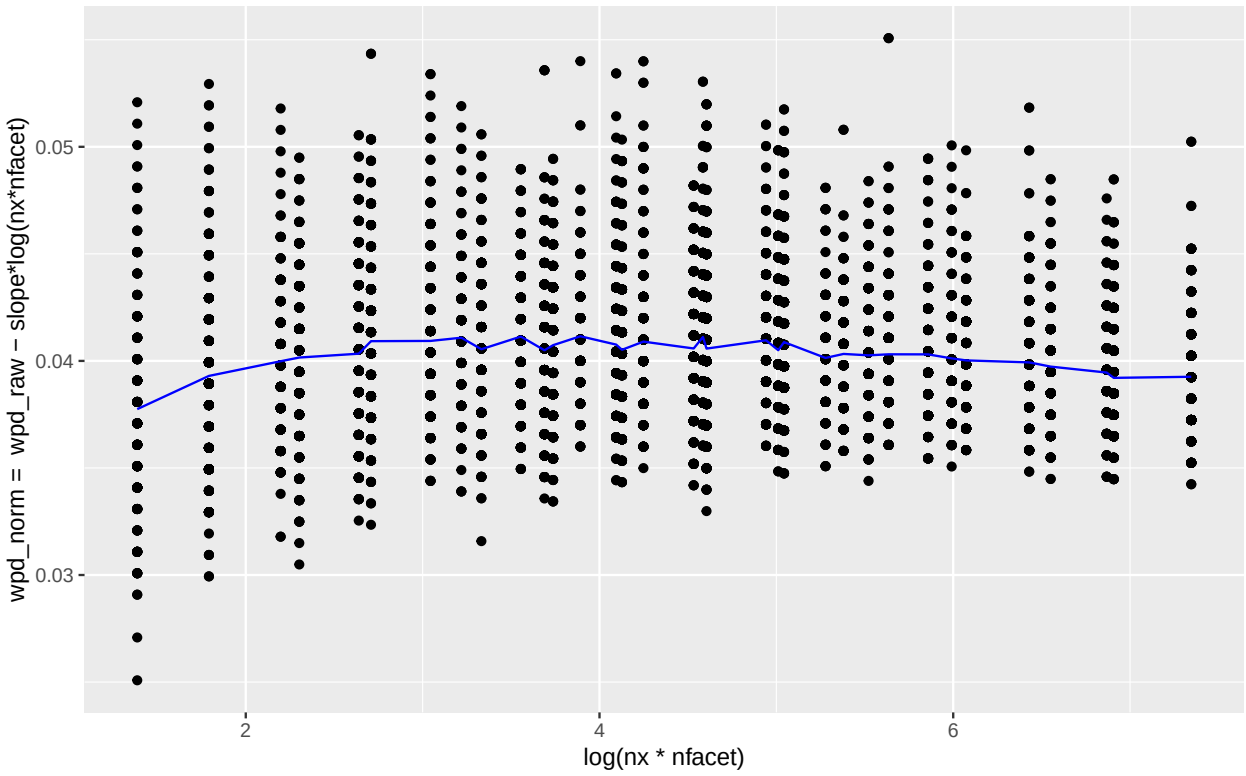
$$y_k = a + b * log(z_k) + e_k$$

, where, $y_k = median_l(x_{k,l})$ and e_k are idiosyncratic errors. We have gone with the approach of fitting a linear regression model to estimate the parameters a and b. The estimates and other model summary is given in **??**.

```
#>
#> Call:
#> lm(formula = actual ~ poly(log(`nx * nfacet`), 1, raw = TRUE),
#>     data = G21_median)
#>
#> Residuals:
#>        Min         1Q     Median         3Q        Max
#> -2.946e-03 -2.240e-04  5.135e-05  4.147e-04  1.014e-03
#>
#> Coefficients:
#>                                      Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                         4.003e-02  4.171e-04   95.97   <2e-16
#> poly(log(`nx * nfacet`), 1, raw = TRUE) 2.826e-03  8.796e-05   32.13   <2e-16
#>
#> (Intercept)                         ***
#> poly(log(`nx * nfacet`), 1, raw = TRUE) ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.0007881 on 32 degrees of freedom
#> Multiple R-squared:  0.9699, Adjusted R-squared:  0.969
#> F-statistic:  1032 on 1 and 32 DF,  p-value: < 2.2e-16
```

The final idea is to find a transformation on wpd_{raw} which would remove the effect of $nx * nfacet$ on wpd_{raw} and thus is defined as follows: $y^* = y - \hat{b} * log(z)$, where y^* is the $median(wpd_{norm})$, y is the $median(wpd_{raw})$, $\hat{b}$ is the estimated value of the parameter b, and $z = nx * nfacet$.

The above takes care of the mean and the heterogeneity of the median transformed measure. But, the original distribution will still have some dissimilarities in shape and location specially for small values of nx and $nfacet$ as could be seen in **??**



3.3.5 Linear model

A linear model is fitted to see how the values of wpd_{raw} changes with the values of nx and $nfacet$. The model is of the form

$$y = a + b * log(x) + e$$

, where $y = median(wpd_{raw})$ and $x = nx * nfacet$. wpd_{lm} is a transformation on wpd_{raw} which is designed to remove the impact of $nx * nfacet$ on wpd_{raw} and thus is defined as follows: $wpd_{lm} = wpd_{raw} - \hat{a} - \hat{b} * log(nx * nfacet)$ $wpd - lm - horizontal$ seems to have no relationship with $nx * nfacet$ as could be seen in Figure 6.

3.3.6 Generalised linear model

In the linear model approach, $wpd_{raw} \in R$ was assumed, whereas, wpd_{raw}, Jensen-Shannon Distance (JSD) lies between 0 and 1. Furthermore, JSD follows a Chi-square distribution, which is a special case of Gamma distribution and hence belongs to exponential family of distributions. Therefore, we can fit a generalized linear model instead of a linear model to allow for the response variable to follow a Gamma distribution. The inverse link is used when we know that the mean response is bounded, which is applicable in our case since $0 \leq wpd_{raw} \leq 1$.

We fit a Gamma generalized linear model with the inverse link which is of the form:

$$y = a + b * log(x) + e$$

, where $y = median(wpd_{raw})$, $x = nx * nfacet$. Let $E(y) = \mu$ and $a + b * log(x) = g(\mu)$ where g is the link function. Then $g(\mu) = 1/mu$ and $\hat{\mu} = 1/(\hat{a} + \hat{b}log(x))$. The residuals from this model $(y - \hat{y}) = (y - 1/(\hat{a} + \hat{b}log(x)))$ would be expected to have no dependency on x. Thus, wpd_{glm} is chosen as the residuals from this model and is defined as: $wpd_{glm} = wpd_{raw} - 1/(\hat{a} + \hat{b} * log(nx * nfacet))$.

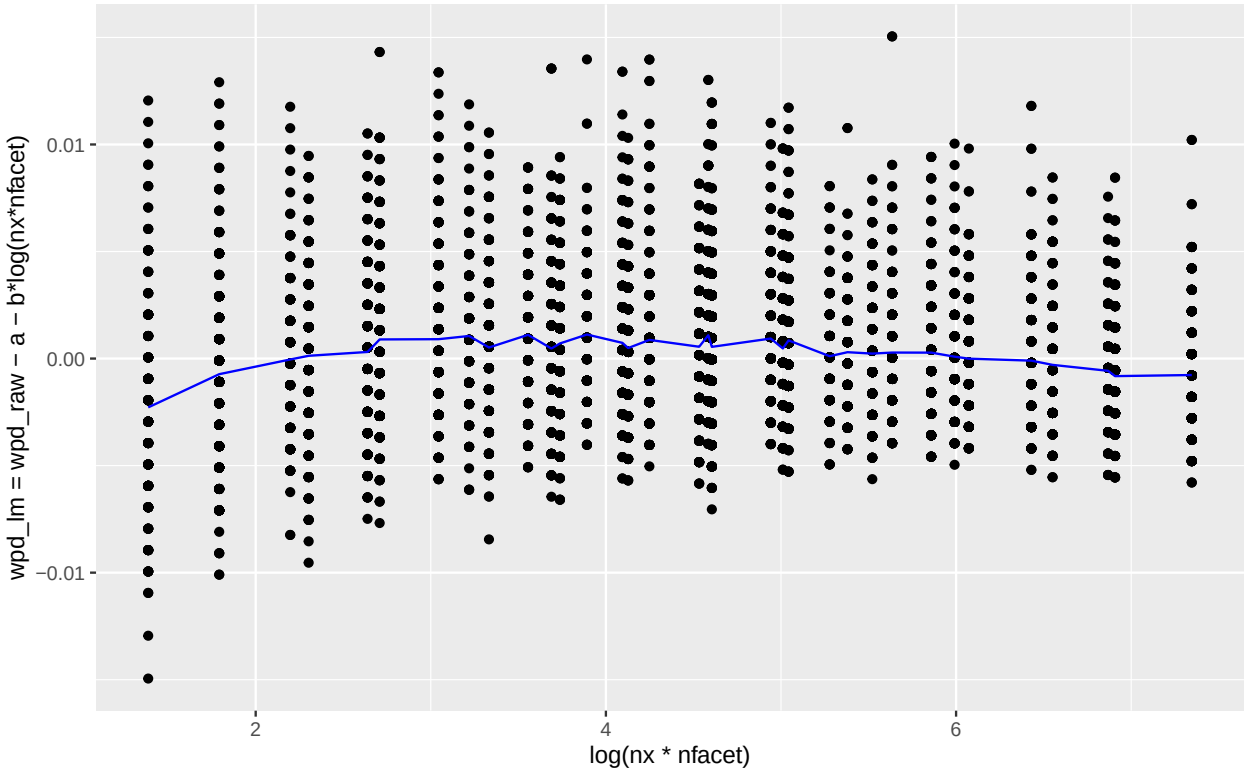


Figure 6: wpd_{lm} is plotted against log(nx*nfacet) and this transformation leads to median(wpd_{lm}) to having almost no relationship with log(nx*nfacet)

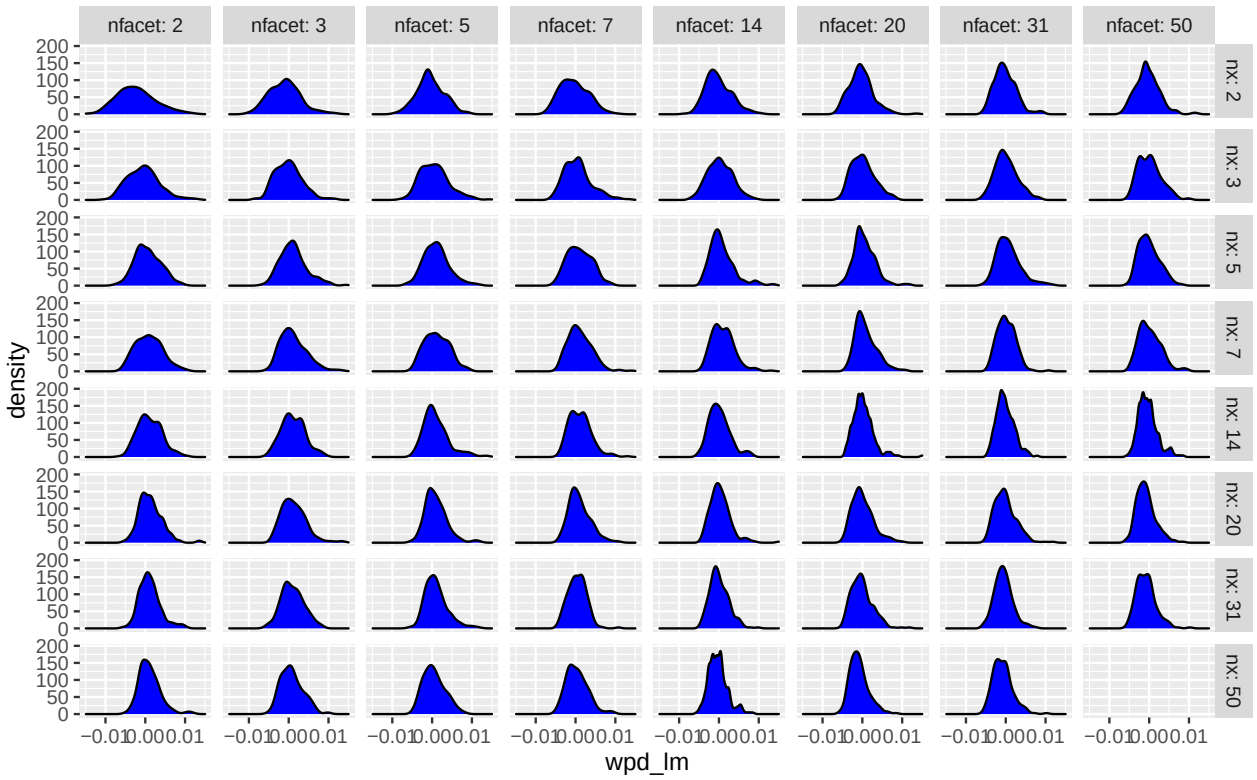
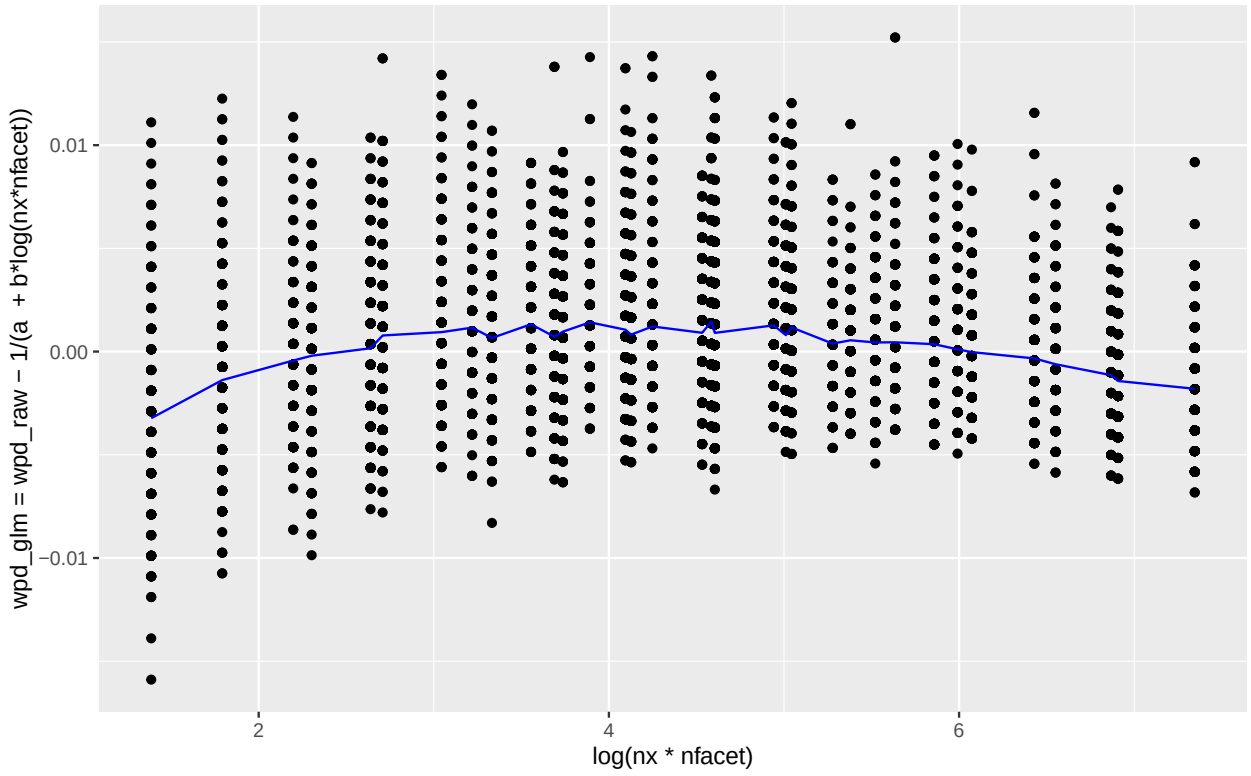


Figure 7: The distribution of wpd_{lm} is plotted. The distributions are more similar across higher nx and nfacet and are different for smaller nx and nfacet.



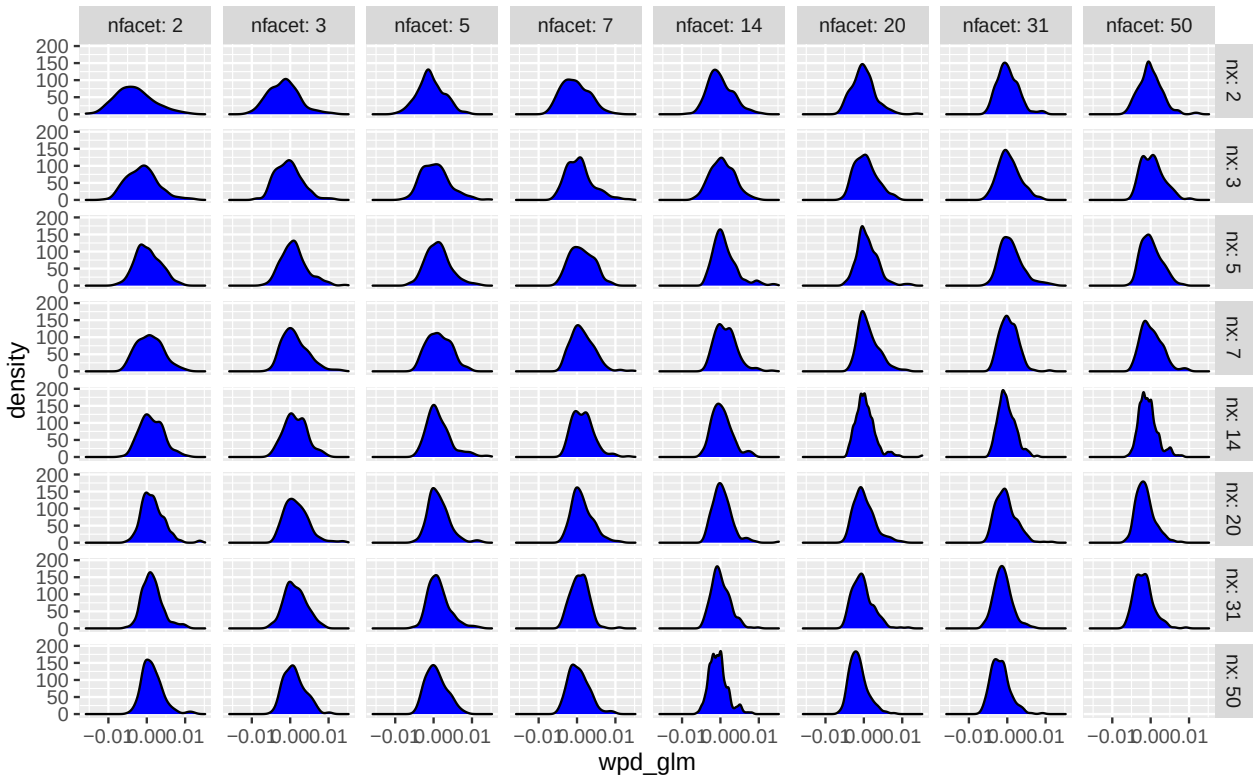


Figure 8: The distribution of wpd_{glm} is plotted. The distributions are more similar across higher nx and nfacet and dissimilar for fewer nc and nfacets.

3.4 Permutation approach to normalisation

The simulated data for each of the panels is permuted/shuffled $nperm = 200$ times and for each of those permutations wpd_{norm} is computed as follows: $wpd_{perm} = (wpd_{raw} - mean(wpd_{raw}))/sd(wpd_{raw})$. This is done so that the distribution of the normalised measure wpd_{norm} has the same mean and standard deviation across different nx and nfacet.

Please note that standardizing the variable wpd_{perm} in this approach leads to $location = 0$ and $scale = 1$ for this variable.

3.5 Bringing them both to the same scale

We see that the transformation through the modeling approach leads to very similar distribution across high nx and $nfacet$ (higher than 7) and not so much for lower nx and $nfacet$. Hence, the computational load of permutation approach could be alleviated by using the modeling approach for the higher nx and $nfacet$, however, it is important that we use the permutation approach for lower nx and $nfacet$. However, it is difficult to compare the transformed wpd from both of these approaches, since each of the variables is measured on a different scale (each of them have location 0). The transformed variables from the two approaches could be brought to the same scale so that for smaller categories, permutation approach is used and for larger categories, we can stick to modeling approach. These could be done through the following:

- Making the range of both the variables same by using min-max scaling method. In practice, however, we would only have one value of wpd_{raw} which we need to transform using the modeling approach. Hence, min-max scaling approach could not be used here.

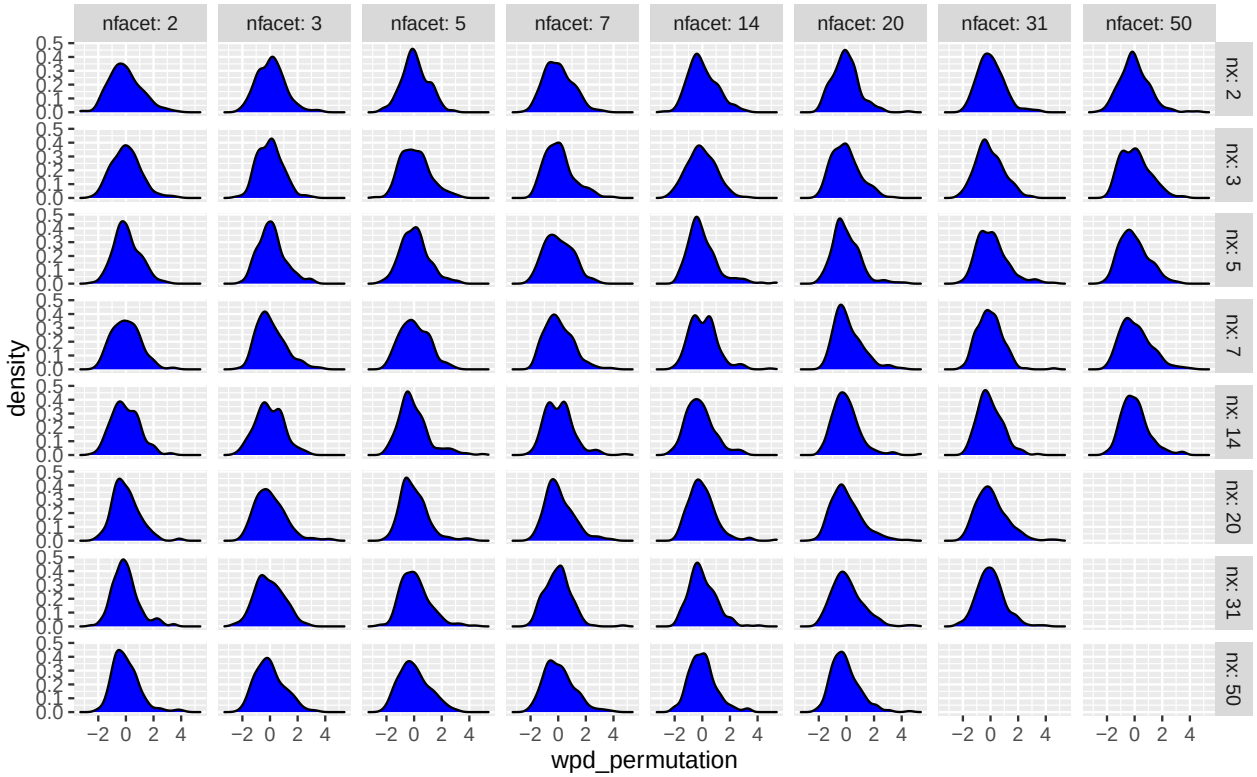


Figure 9: The distribution of $wpd_{permutation}$ is plotted. The distributions are more similar across different nx and nfacet (specially for small nx and nfacet) but this approach has the downside of more computational time.

- Standardizing the variables and expressing scores at standard deviation units. Again in practice, however, we would only have one value of wpd_{raw} which we need to transform using the modeling approach. Hence, standardizing scores could not be used here as we do not have the mean and standard deviation of a series while using transformation using modeling.
- Make the location and scale of both the approaches similar so that they could be compared. Please note that the range of values could be different in this case, however location and scale are brought to same levels.)

The measure wpd_{glm} has location 0 and standard deviation ~ 0.003, whereas the measure $wpd_{permutation}$ which is a z-score, has a normal distribution with location 0 and standard deviation 1. To bring them to the same scale, we have defined $wpd_{glm-scaled} = wpd_{glm} * 300$, which brings the standard deviation of $wpd_{glm-scaled}$ to almost 1, without changing the location.

The measure $wpd_{glm-scaled}$ seems to roughly follow a normal distribution except in the tails as could be seen in Figure 10 and the very method of permutation approach ensures that $wpd_permutation$ is also normally distributed. Further, they are brought to the similar scale and location and hence could be compared.

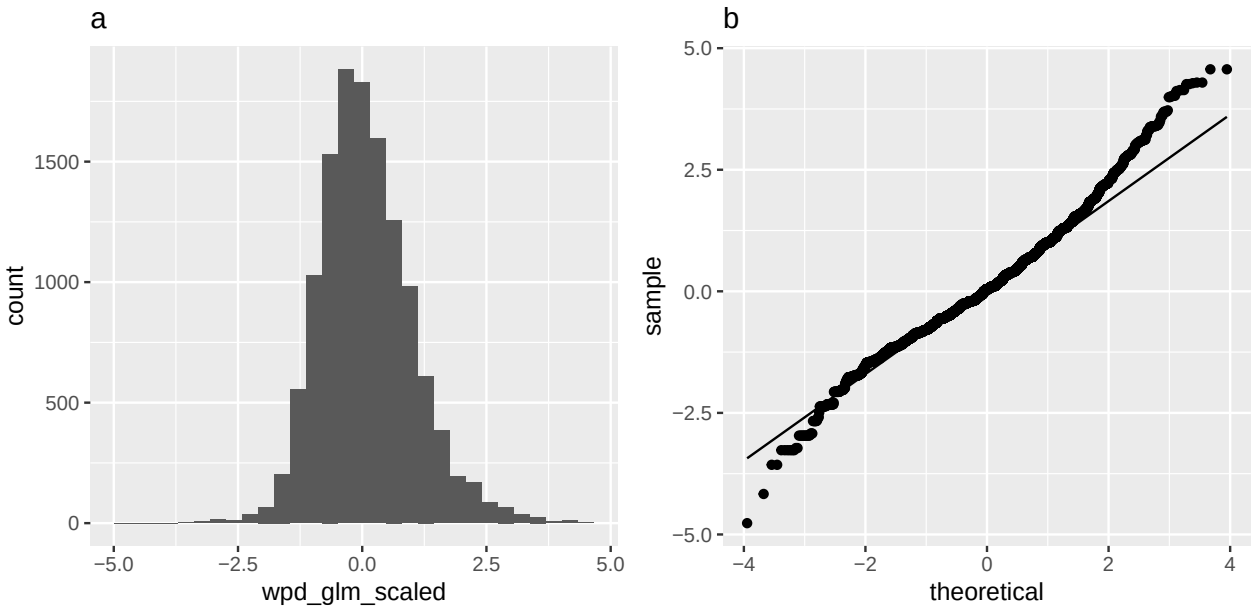


Figure 10: In panel a, the histogram of $wpd_{glm-scaled}$ is plotted. In parl b, the QQ plot is shown with the theoretical quantiles on the x-axis and $wpd_{glm-scaled}$ quantiles on the y-axis. The distribution looks symmetric and looks like normal except in the tails.

3.6 Results

This section reports the results of a simulation study that was carried out to evaluate the behavior of our distance measure across all potential factors. The results are reported in two parts: for the a) raw measure and then b) normalised one to show how the loopholes for the raw measures were removed using the normalized ones.

First, the behavior of wpd_{raw} and wpd_{norm} is explored in designs where there is no difference in distribution between x and facet categories. We have considered different initial distributions to study the impact of initial distribution under the null setup. Using two types of distributions, viz. normal and gamma (non-normal), we generated observations for each combination of nx and $nfacet$ from the following sets: $nx = \{2, 3, 5, 7, 14, 20, 31, 50\}$ and $nfacet = \{2, 3, 5, 7, 14, 20, 31, 50\}$ to cover a wide range of levels from very low to

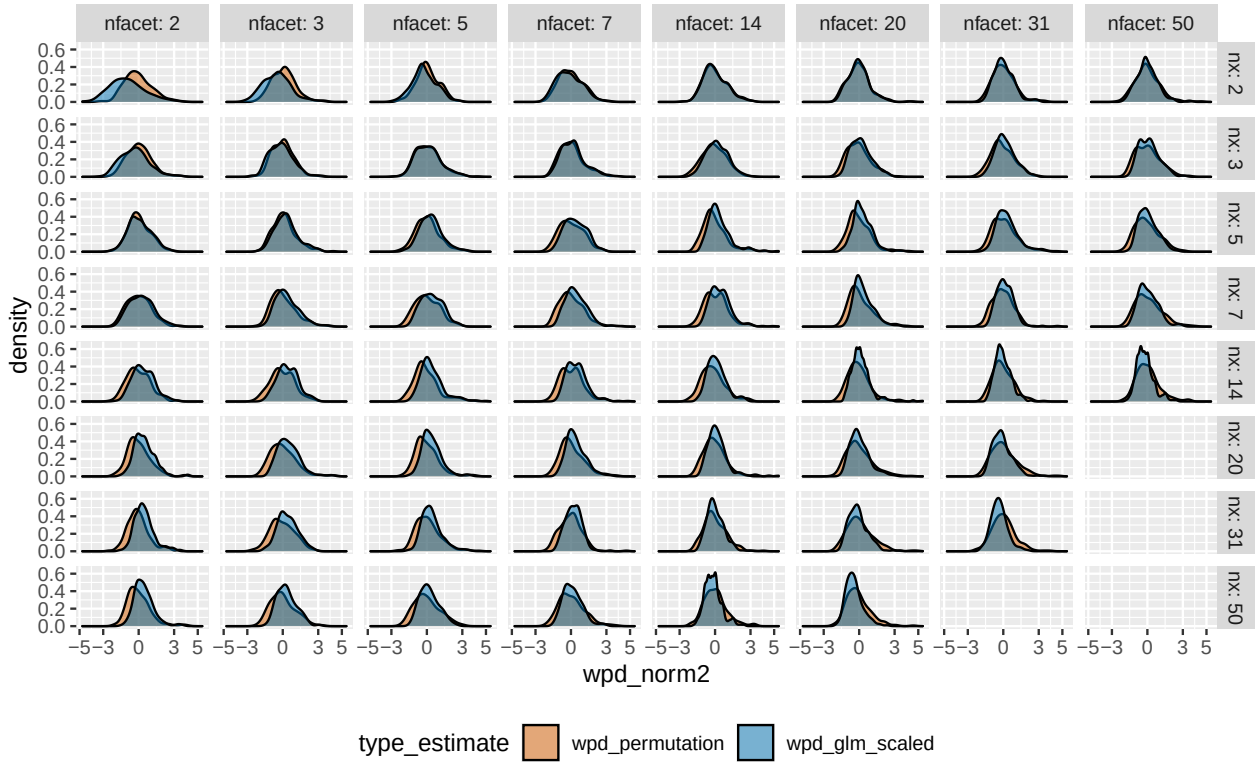


Figure 11: w$pd_permutation$ and $wpd_{glm-scaled}$ are plotted together on the same scale. They also have the same location and hence the values from these two approaches could be compared across panels. $wpd_{glm-scaled}$ would be used to normalise wpd_{raw} for higher nx and $nfacet$ and $wpd_{glm-scaled}$ would be used for smaller levels to alleviate the problem of computational time.

moderately high. Each combination is being referred to as a *panel*. That is, data is being generated for each of the panels $\{nx = 2, nfacet = 2\}, \{nx = 2, nfacet = 3\}, \{nx = 2, nfacet = 5\}, \ldots, \{nx = 50, nfacet = 31\}, \{nx = 50, nfacet = 50\}$. For each of the 64 panels, $ntimes = 500$ observations are drawn for each combination of the categories. That is, if we consider the panel $\{nx = 2, nfacet = 2\}$, 500 observations are generated for each of the combination of categories from the panel, namely, $\{(1,1),(1,2),(2,1),(2,2)\}$. The values of λ is set to 0.67, since we want to up-weigh the within-facet distances and that of ω is set to 0, since there is no significant differences between distributions in the null case. Observations were generated for each type of distribution changing the shape and scale to study the effect of shape, scale and type of distribution on wpd. The set of distributions considered for this purpose is $N(0,1), N(5,1), N(0,5), \Gamma(0.5,1), \Gamma(2,1)$. Each of the scenario is run $nsim = 200$ times to see the distribution of wpd values for each scenario.

Secondly, the behavior of raw and normalized wpd is explored in designs where there is in fact difference in distribution between facet categories (D_{var_f}) or across x-categories (D_{var_x}) or both ($D_{var_{all}}$). Using $\omega = \{1, 2, \ldots, 10\}$ and $\lambda = seq(from = 0.1, to = 0.9, by = 0.05)$, observations are drawn from a N(0,1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = nfacet = \{2, 3, 5, 7, 14, 20, 31, 50\}$. $ntimes = 500$ is assumed for this setup as well. Furthermore, to generate different distributions across different combination of facet and x levels, the following method is deployed - suppose the distribution of the combination of first levels of x and facet category is $N(\mu, \sigma)$ and μ_{jk} denotes the mean of the combination $(a_j b_k)$, then $\mu_{j.} = \mu + j\omega$ (for design D_{var_x}) and $\mu_{.k} = \mu + k\omega$ (for design D_{var_f}).

The tabulated values and graphical representations of the simulation results are provided in Appendix. The learning from the simulations are as follows: The values of the measure wpd_{raw} is least for D_{null}, followed by D_{var_f}, D_{var_x}) and $D_{var_{all}}$). This is a desirable result since the measure wpd_{raw} was designed such that this relationship holds. Furthermore, the distribution of the measure wpd_{raw} changes for different facet and x categories. The location of the distribution shifts to the right and it also becomes more skewed for more higher facet and x-axis categories. Now this is not desirable, as it would mean that we can't compare the wpd_{raw} values across different panels. The distribution of wpd_{norm} looks more similar with at least the mean and standard of the distributions being uniform across panels. This means, wpd_{norm} could be used to measure differences in distribution across panels. Also, note that since the data is processed using normal-quantile-transform, this measure is independent of the intial distribution of the underlying data and hence is also comparable across different data sets. This is valid for the case when sample size $ntimes$ for each combination of categories is at least 30 and $nperm$ used for computing wpd_{norm} is at least 100. More detailed results about the properties of wpd_{raw} and wpd_{norm} could be found in Appendix.

4 Choosing harmonies with significant wpd

Complete randomness in the measured variable indicates that the process follows a homogeneous underlying distribution over the whole time series, which essentially implies there is no interesting distinction across any different categories of the cyclic granularities. We can remove the harmonies for which no interesting patterns are observed through a randomization permutation method. Essentially, the assumption is that under the null hypothesis, there is no difference in categories between the pair of cyclic granularities in the chosen harmony. This method is based on the generation of randomly chosen reassignments (permutations) of the data across different cyclic granularities and the computation of wpd_{norm} for each of these reassignments. The percentages of times the theoretical distribution greater than or equal to the respective observed wpd_{norm} values are calculated and are used to obtain the P value. The procedure for the permutation test is:

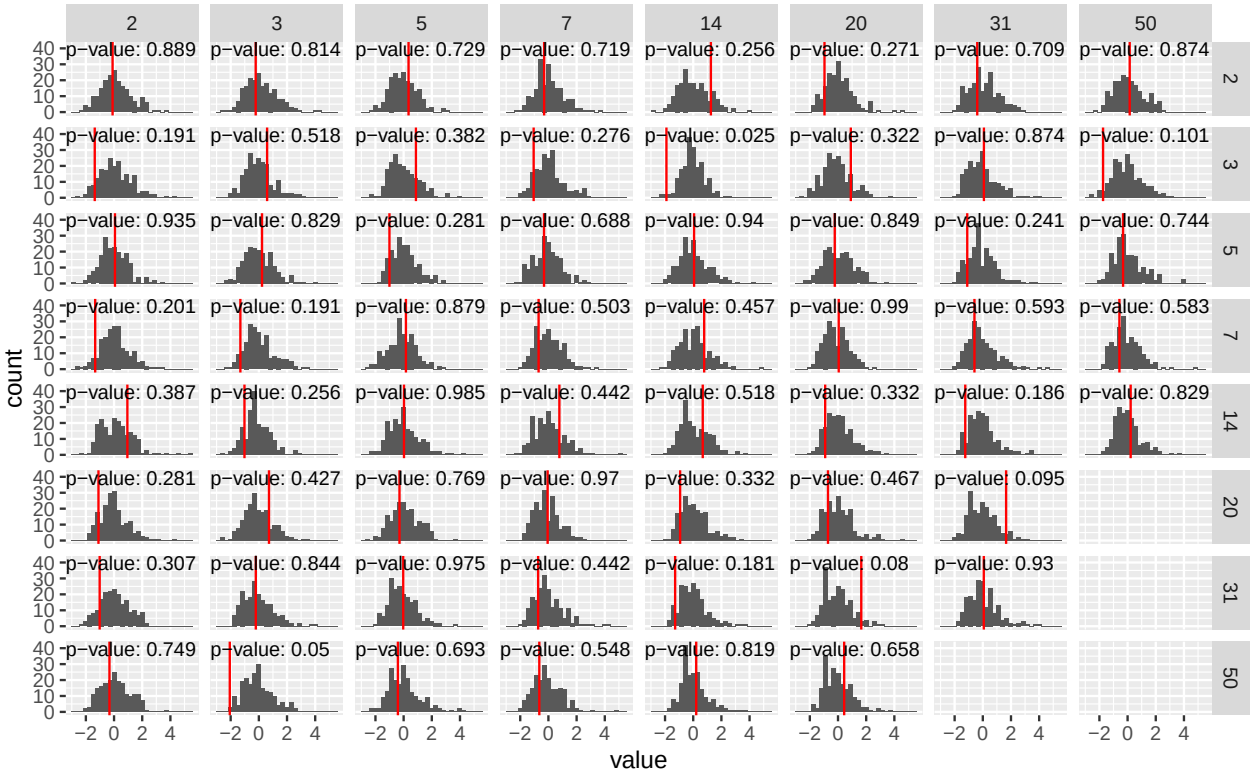
1. Given the data; $\{v_t : t = 0, 1, 2, \ldots, T-1\}$, the wpd_{norm} is computed and is represented by wpd_{obs}.
2. From the original sequence a random permutation is obtained: $\{v_t^* : t = 0, 1, 2, \ldots, T-1\}$.
3. wpd_{norm} is computed for the permuted sequence of the data and is represented by wpd_{perm_1}.
4. Steps (2) and (3) are repeated a large number of times M (M = 200).
5. For each permutation, one wpd_{perm_i} is obtained. Define $wpd_{sample} = \{wpd_{perm_1}, wpd_{perm_2}, \ldots, wpd_{perm_M}\}$.

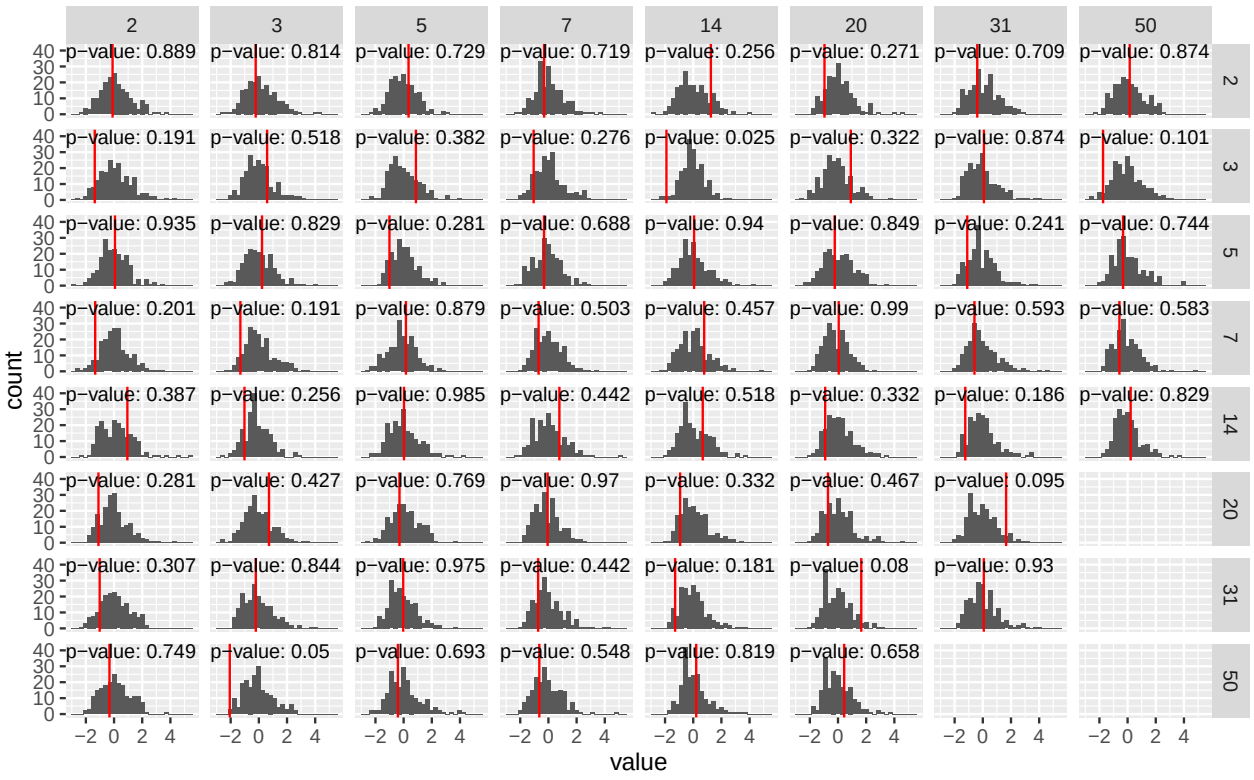
6. 95^{th} percentile of this wpd_{sample} distribution is computed and stored in $wpd_{threshold}$.

7. If $wpd_{obs} > wpd_{threshold}$, harmony pairs are accepted. Only one threshold for all harmony pairs.

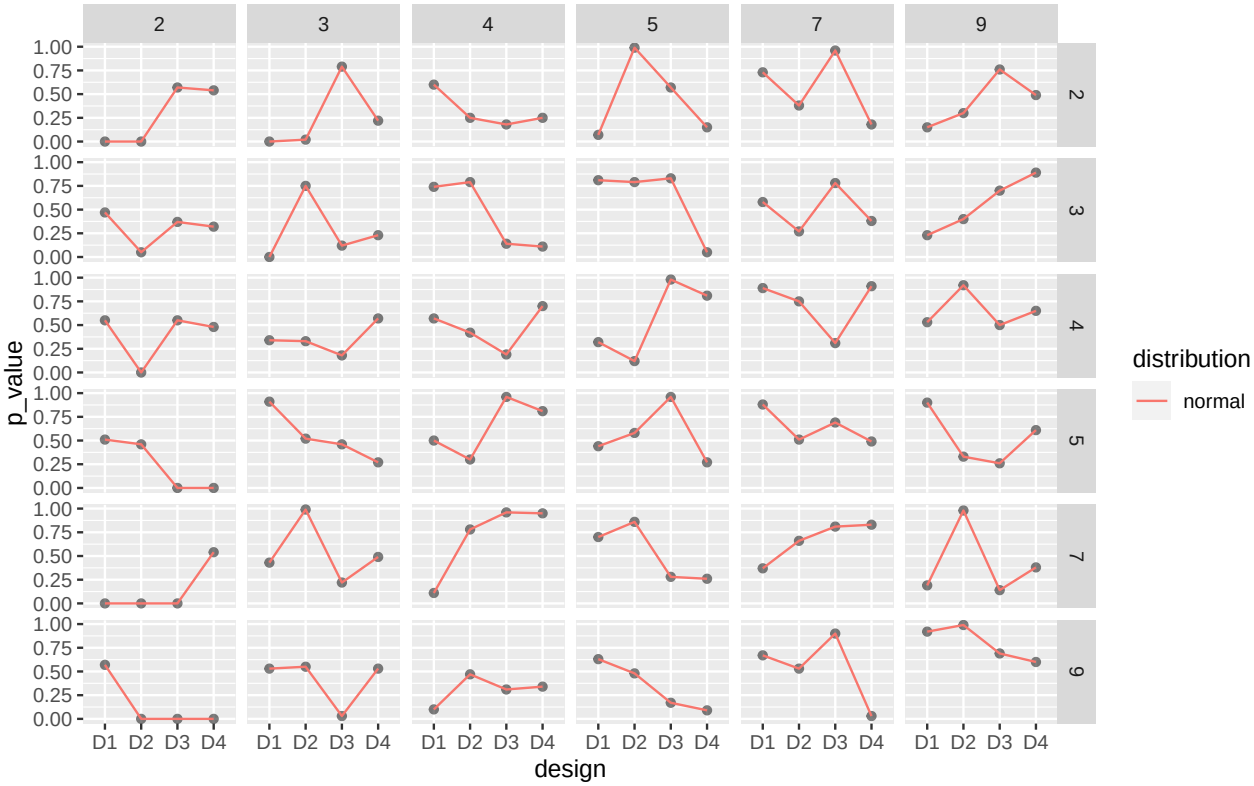
The p-value of the design D_{null} is size. The p-value of other designs is power. Confidence interval of the

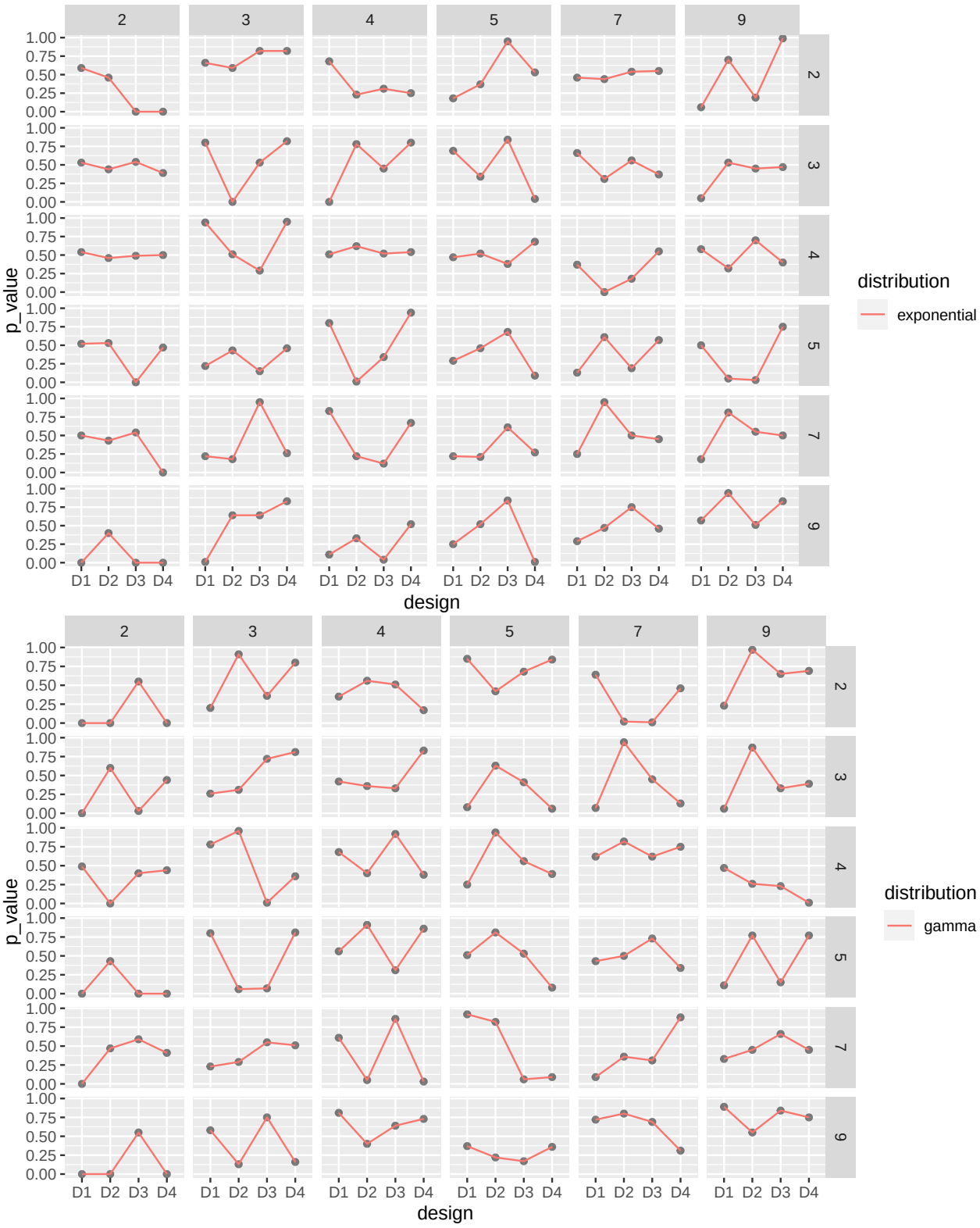
The p-value is almost always greater than 0.05, which means there is no significant differences between the different categories, which is true from the simulation design.

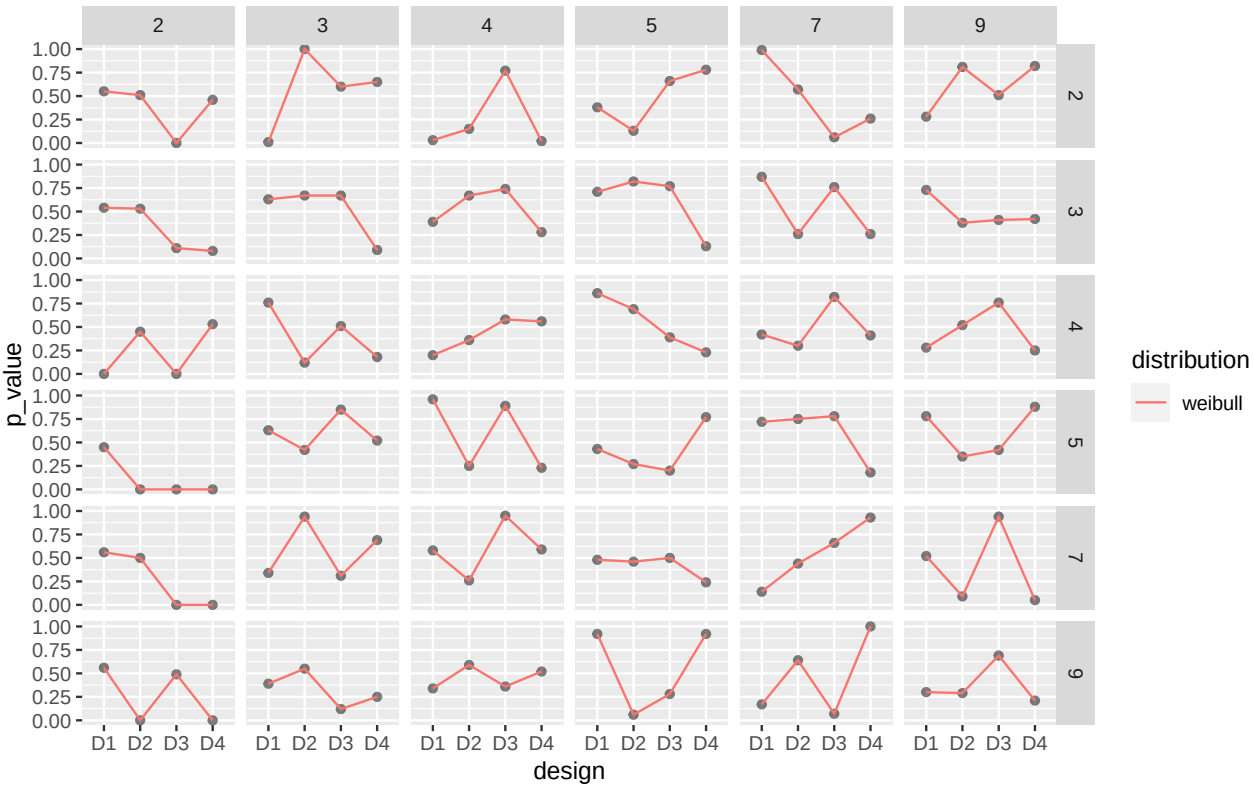




4.0.0.1 Characteristics under different simulation designs A set of simulation runs that are conducted and some outputs of which are reported.




2
3
4
5
7
9
p_value
1.00
0.75
0.50
0.25
0.00
D1 D2 D3 D4
design
distribution
exponential
gamma



5 Application

Clustering of electricity customers based on their consumption behavior facilitate effective market segmentation and hence better management and pricing. Electricity demand time series are the primary source of information of customers' consumption behavior. Due to technological developments, smart meters could provide large quantities of measurements on energy usage at more and more finer time intervals. Due to this extreme dimensionality along time of the demand data, it is a challenge to have clustering methods which could reduce the dimensionality to produce distinct clusters. One of the potential solutions is time series feature extraction. However, it is limited by the type of noisy, patchy and unequal time series that is very common in data sets of any residential customers. The idea is to choose similarity measures, which do not lose the core characteristic information of demand across different time deconstructions in aggregation processes and is robust against asynchronous or incomplete time series.

We use the data from one of the customer trials (Department of the Environment and Energy 2018) conducted as part of the Smart Grid Smart City project in Newcastle, New South Wales and some parts of Sydney. The data has customer wise data on energy consumption for every half hour from February 2012 to March 2014. The idea here is to use our similarity measure and show improved cluster distinction as compared to baseline feature based approaches. A data set of 100 homes is used to evaluate both approaches. This is different to deterministic segmentation which attempts to characterise load as a function of a fixed number of parameters, such as occupancy or types of appliances, and use those parameters for classification. We only use their load patterns across time deconstructions to define the similarity measure.

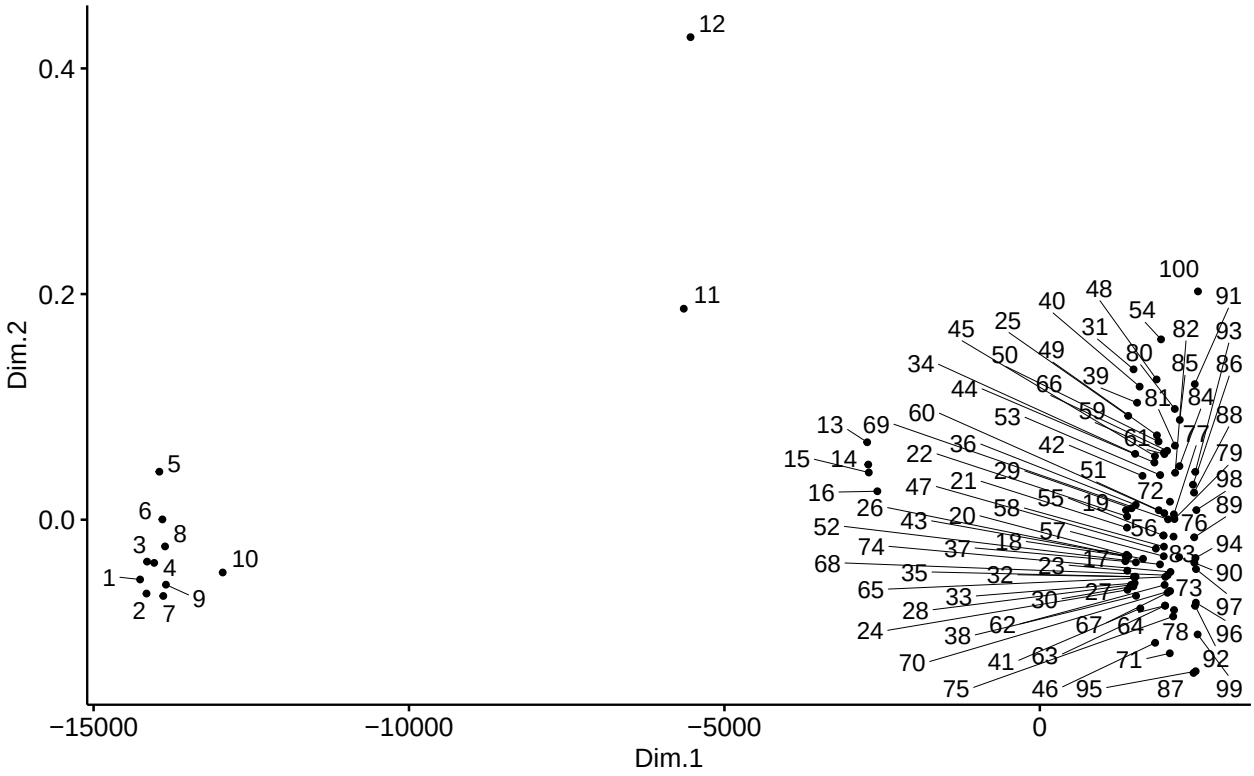
5.0.1 Multidimensional scaling and cluster analysis

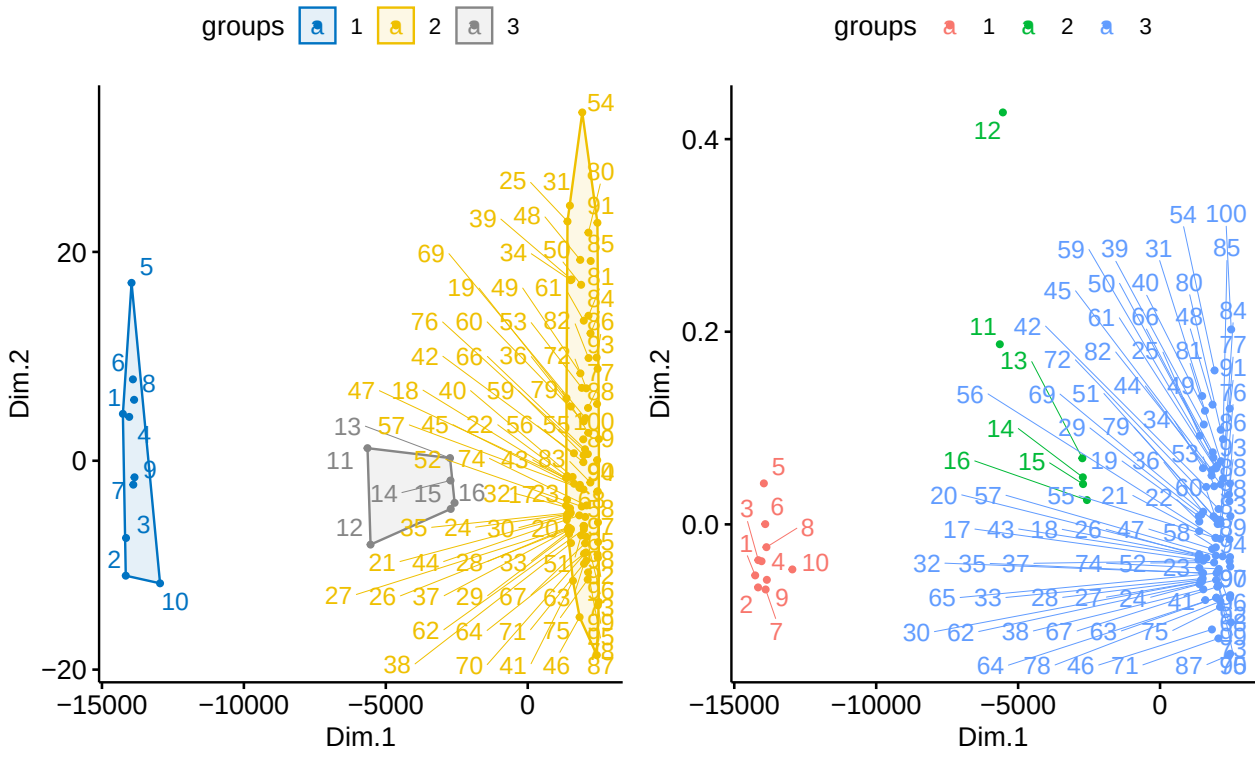
Cluster analysis and multidimensional scaling (MDS) methods are used to explore the temporal structure of various households, their interrelations and determine similar behaving clusters of households in this study. MDS methods are used to produce a lower dimensional display space where each household is represented by

a point and the distance between these points give the measure of similarity between these households. On the other hand, clustering techniques are used to provide information about the clustering structure of the households. Households that are in the same cluster can be considered to be more similar, while households are more dissimilar in their energy behavior across one or few temporal deconstructions.

wpd_norm is considered as the measure of similarity or dissimilarity while using MDS and clustering techniques. wpd_norm could be constructed for each pair of harmony for each household and then similarity could be measured across single or all harmonies together.

- check transformation and scalar approaches are same





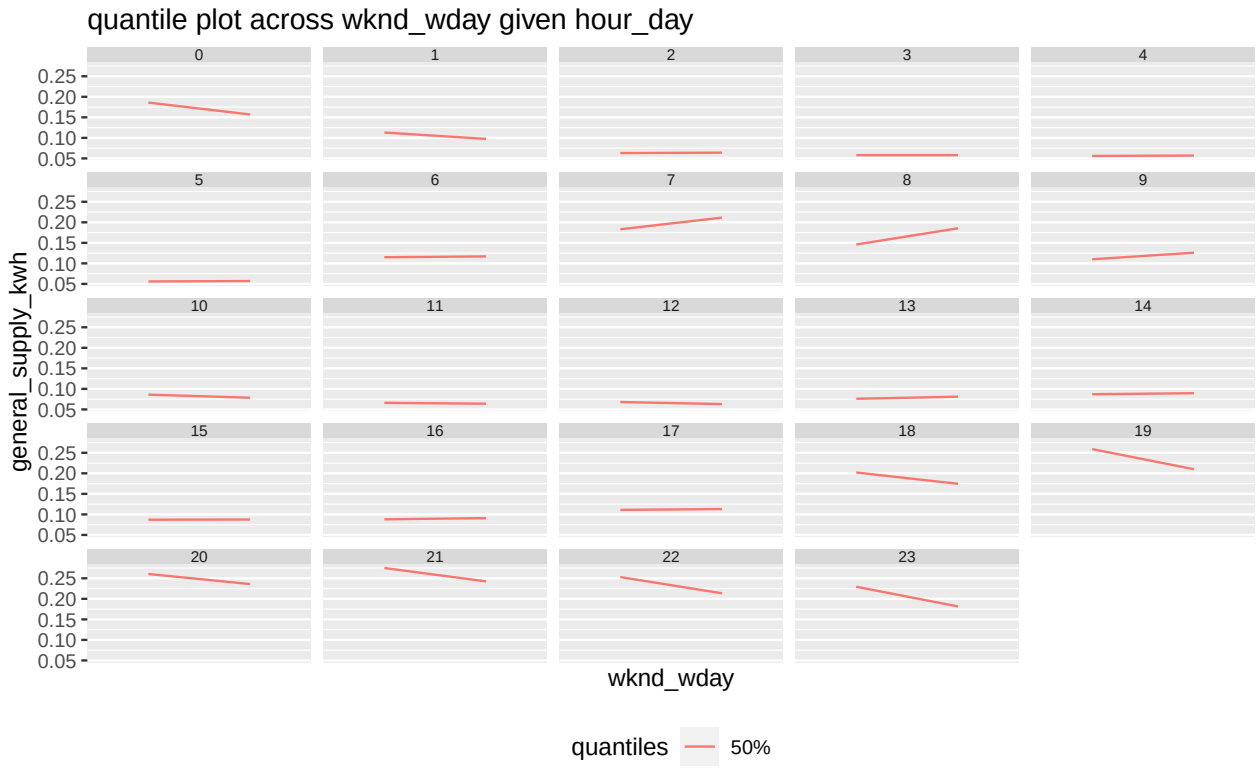
10018272 is a customer from cluster 3, for which hour_day/wknd-wday comes after wknd-wday/day-month as opposed to 10006414 (cluster 1) for whom the direction is reversed.

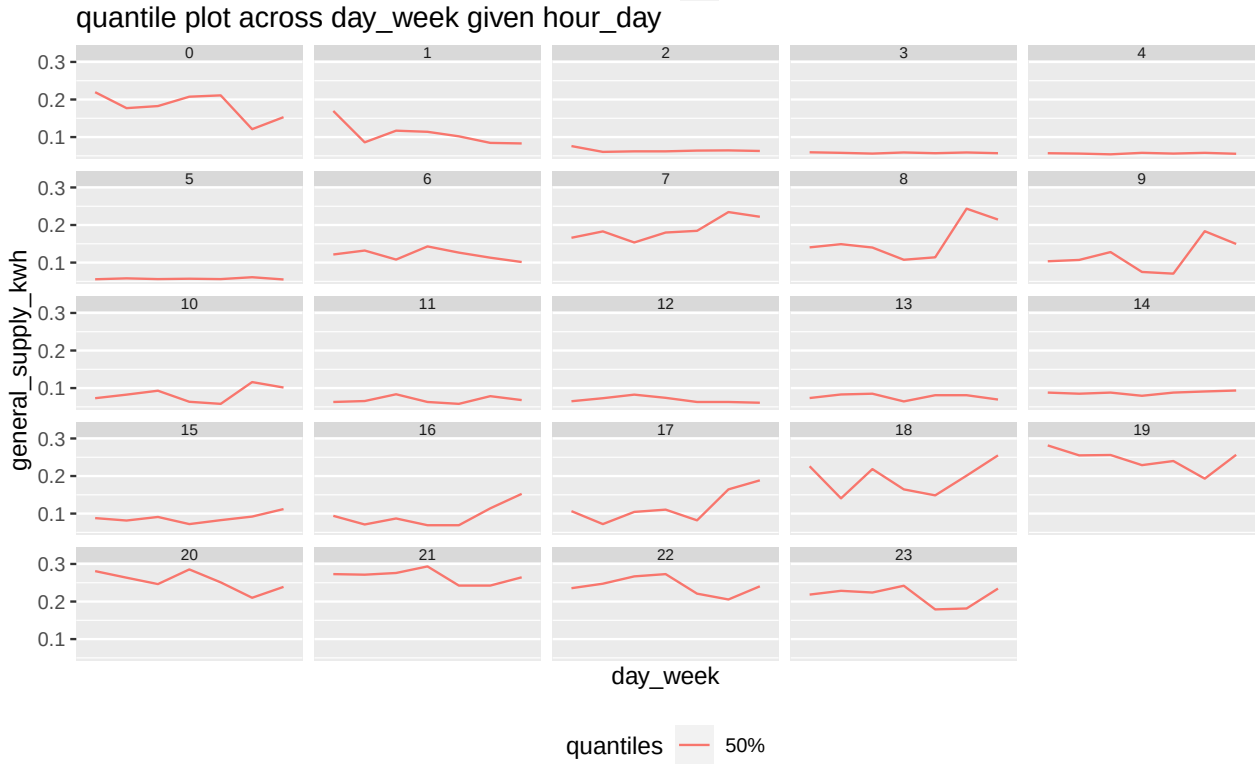
```
#> [1] 10006414
```

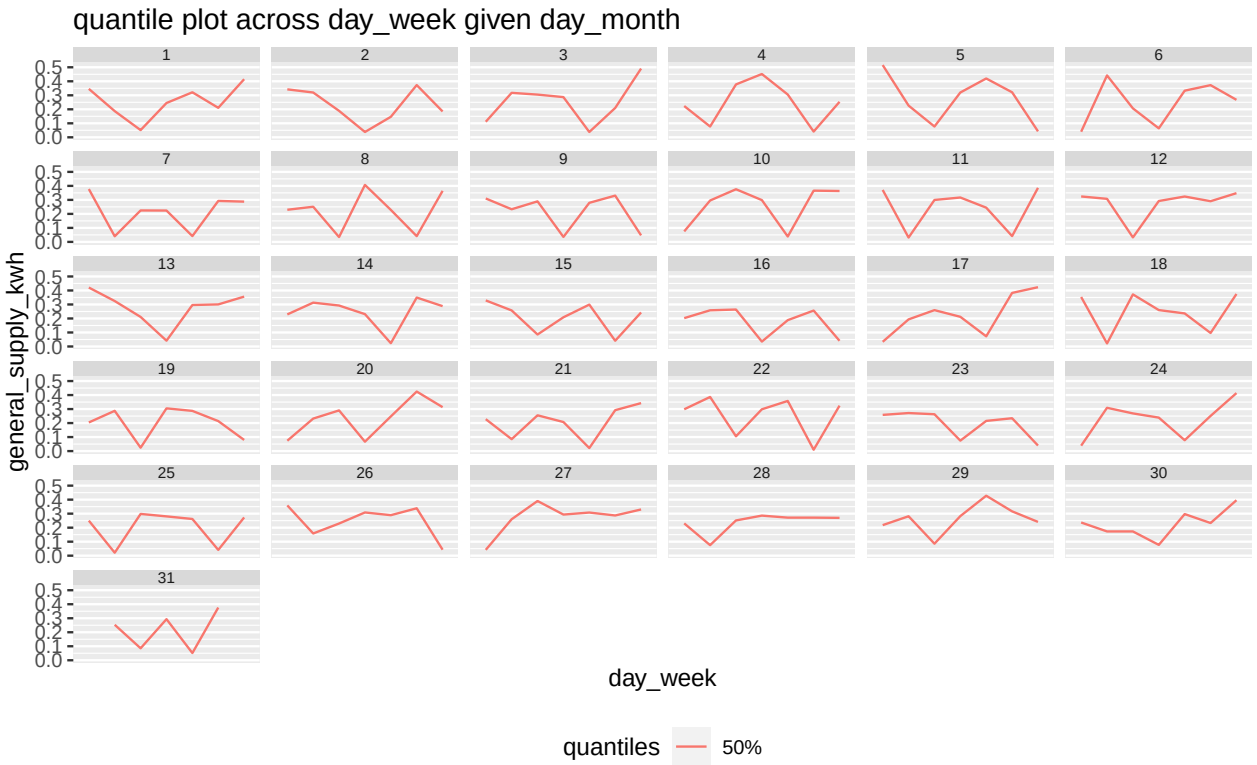
```
#> # A tibble: 16 x 6
#>    customer_id facet_variable x_variable facet_levels x_levels wpd_norm
#>          <dbl> <chr>          <chr>             <dbl>    <dbl>    <dbl>
#>  1    10006414 hour_day       wknd_wday            24        2    45.5
#>  2    10006414 hour_day       day_week             24        7    33.3
#>  3    10006414 hour_day       week_month           24        5    14.7
#>  4    10006414 wknd_wday      day_month             2       31    10.7
#>  5    10006414 day_month      wknd_wday            31        2     9.80
#>  6    10006414 wknd_wday      hour_day              2       24     9.07
#>  7    10006414 day_week       hour_day              7       24     8.71
#>  8    10006414 day_week       week_month            7        5     7.54
#>  9    10006414 day_week       day_month             7       31     5.45
#> 10    10006414 week_month     day_week              5        7     4.96
#> 11    10006414 wknd_wday      week_month            2        5     4.30
#> 12    10006414 hour_day       day_month            24       31     3.93
#> 13    10006414 week_month     hour_day              5       24     3.81
#> 14    10006414 day_month      day_week             31        7     3.44
#> 15    10006414 week_month     wknd_wday             5        2     1.63
#> 16    10006414 day_month      hour_day             31       24     1.62
```

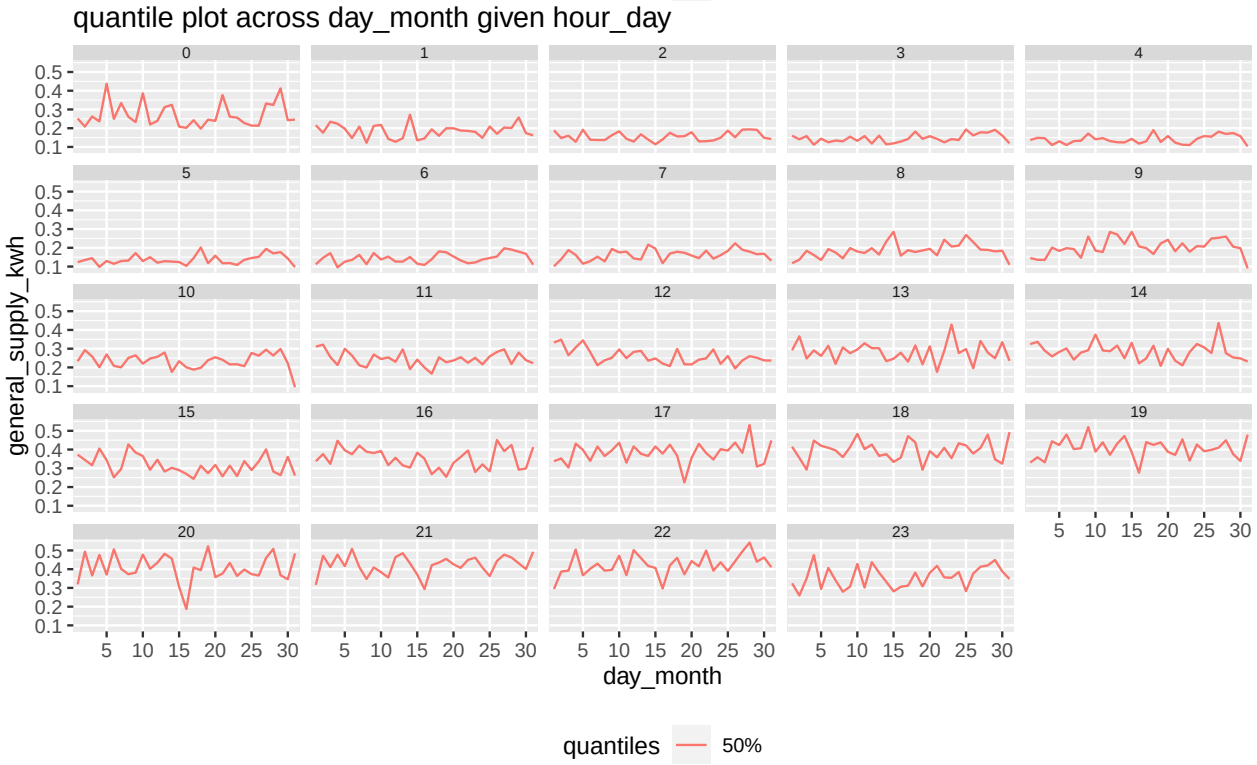
```
#> # A tibble: 16 x 6
#>    customer_id facet_variable x_variable facet_levels x_levels wpd_norm
#>          <dbl> <chr>          <chr>             <dbl>    <dbl>    <dbl>
#>  1    10007340 day_month      day_week             31        7   24.7
```

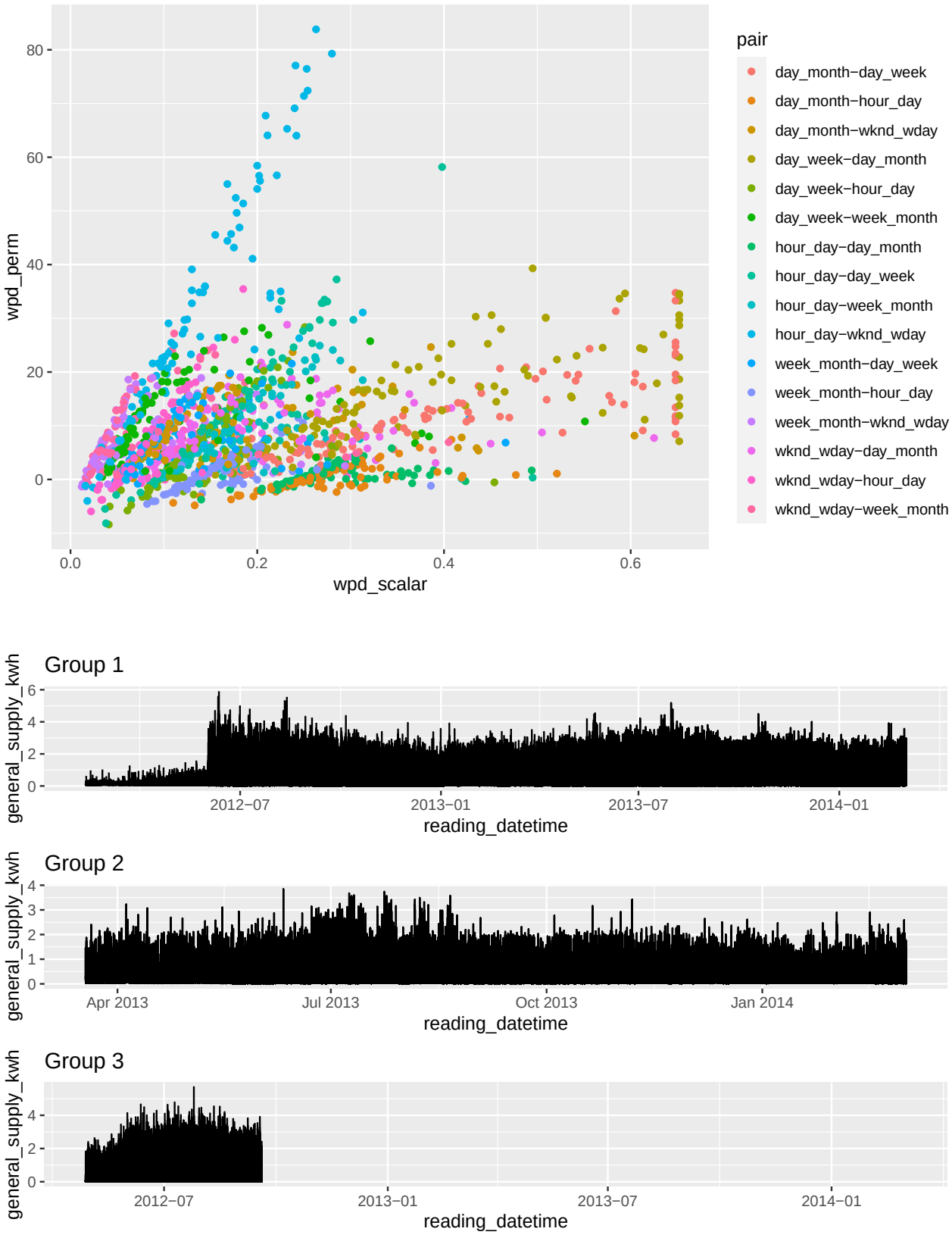
```
#>  2    10007340 day_month      wknd_wday           31        2   24.6
#>  3    10007340 hour_day       wknd_wday           24        2   24.2
#>  4    10007340 day_week       day_month            7       31   22.8
#>  5    10007340 wknd_wday      day_month            2       31   14.3
#>  6    10007340 day_week       week_month           7        5   10.4
#>  7    10007340 wknd_wday      week_month           2        5    7.51
#>  8    10007340 week_month     day_week             5        7    7.30
#>  9    10007340 hour_day       day_week            24        7    7.24
#> 10    10007340 week_month     wknd_wday            5        2    5.52
#> 11    10007340 hour_day       week_month          24        5    4.94
#> 12    10007340 wknd_wday      hour_day             2       24    0.808
#> 13    10007340 hour_day       day_month           24       31    0.728
#> 14    10007340 day_month      hour_day            31       24   -2.35
#> 15    10007340 week_month     hour_day             5       24   -2.61
#> 16    10007340 day_week       hour_day             7       24   -2.83

#> # A tibble: 16 x 6
#>    customer_id facet_variable x_variable facet_levels x_levels wpd_norm
#>          <dbl> <chr>          <chr>             <dbl>    <dbl>    <dbl>
#>  1    10018272 wknd_wday      day_month             2       31    7.10
#>  2    10018272 hour_day       wknd_wday            24        2    6.47
#>  3    10018272 day_week       hour_day              7       24    6.23
#>  4    10018272 day_month      day_week             31        7    5.55
#>  5    10018272 day_week       day_month             7       31    4.55
#>  6    10018272 day_month      wknd_wday            31        2    2.80
#>  7    10018272 week_month     day_week              5        7    1.96
#>  8    10018272 week_month     hour_day              5       24    1.82
#>  9    10018272 hour_day       day_month            24       31    1.67
#> 10    10018272 day_week       week_month            7        5    1.61
#> 11    10018272 hour_day       day_week             24        7    1.56
#> 12    10018272 hour_day       week_month           24        5    1.28
#> 13    10018272 wknd_wday      hour_day              2       24    1.26
#> 14    10018272 day_month      hour_day             31       24    1.08
#> 15    10018272 wknd_wday      week_month            2        5    0.714
#> 16    10018272 week_month     wknd_wday             5        2   -0.897
```

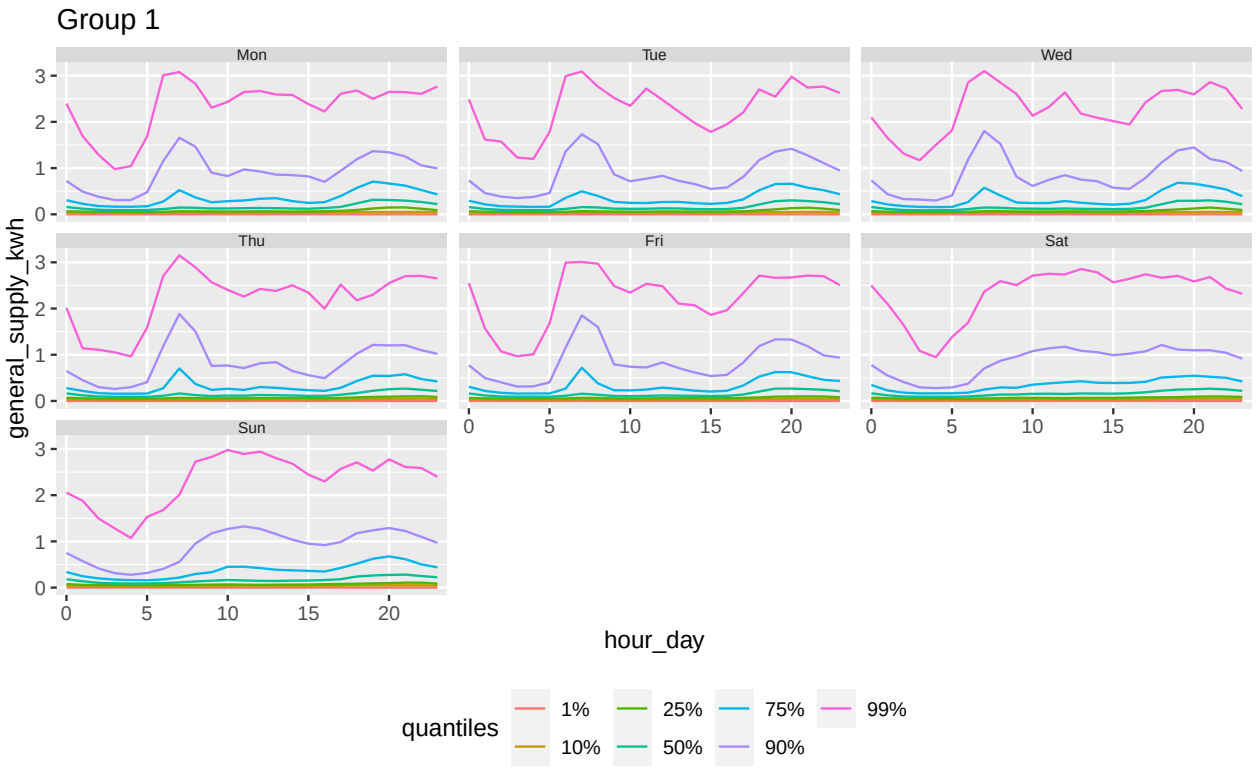

quantile plot across wknd_wday given hour_day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
0.25
0.20
0.15
0.10
0.05
general_supply_kwh
wknd_wday
quantiles
50%

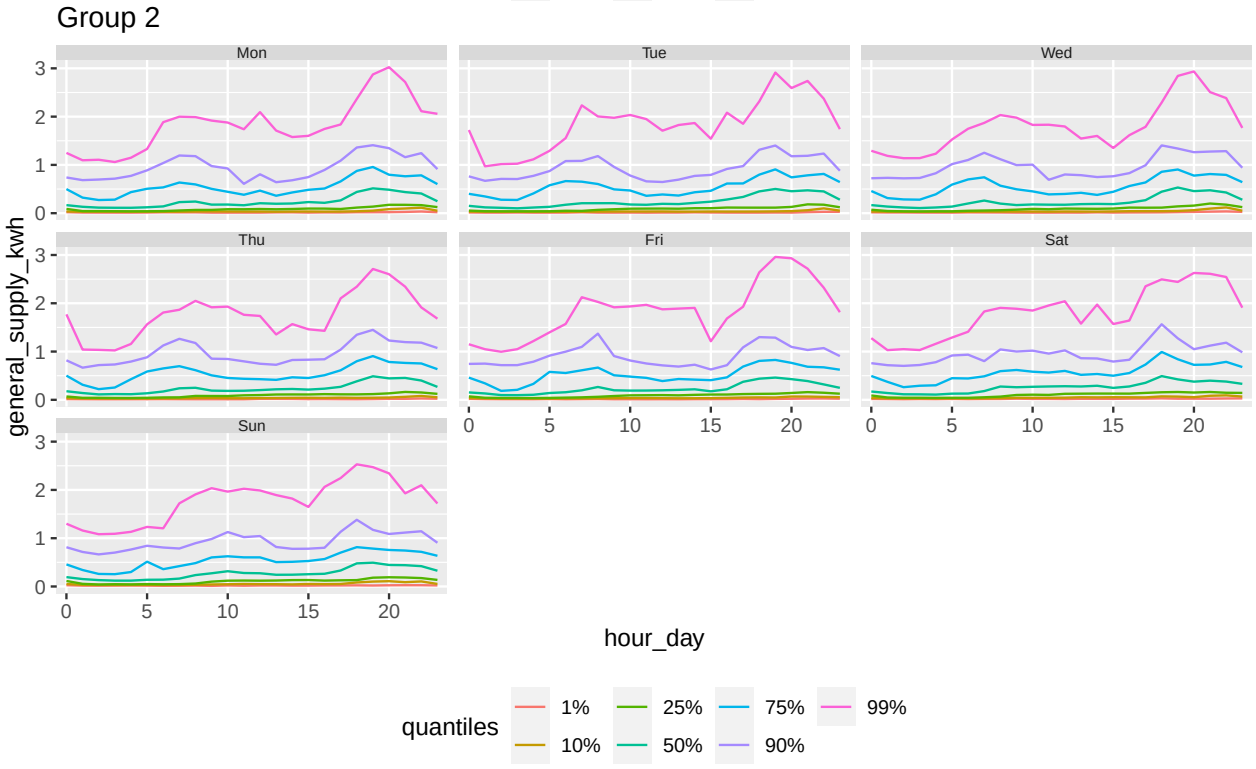

quantile plot across day_week given hour_day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
0.3
0.2
0.1
general_supply_kwh
day_week
quantiles
50%

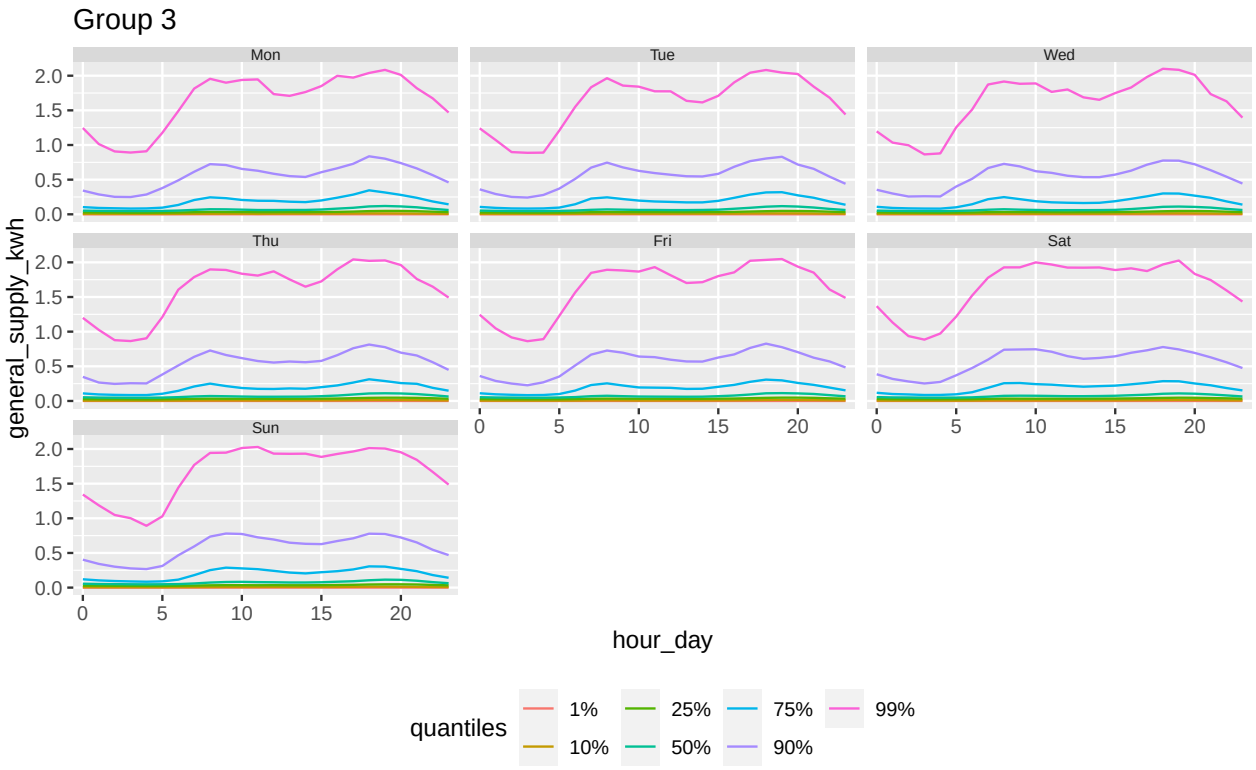

quantile plot across day_week given day_month
general_supply_kwh
day_week
quantiles
50%


quantile plot across day_month given hour_day
general_supply_kwh
day_month
quantiles
50%


wpd_perm
wpd_scalar
pair
day_month−day_week
day_month−hour_day
day_month−wknd_wday
day_week−day_month
day_week−hour_day
day_week−week_month
hour_day−day_month
hour_day−day_week
hour_day−week_month
hour_day−wknd_wday
week_month−day_week
week_month−hour_day
week_month−wknd_wday
wknd_wday−day_month
wknd_wday−hour_day
wknd_wday−week_month
Group 1
Group 2
Group 3
general_supply_kwh
reading_datetime
2012−07
2013−01
2013−07
2014−01
Apr 2013
Jul 2013
Oct 2013
Jan 2014


Group 1
Mon
Tue
Wed
Thu
Fri
Sat
Sun
general_supply_kwh
hour_day
quantiles
1%
10%
25%
50%
75%
90%
99%


Group 2
Mon
Tue
Wed
Thu
Fri
Sat
Sun
general_supply_kwh
hour_day
quantiles
1%
10%
25%
50%
75%
90%
99%



5.0.2 Advantage over taditional feature based clustering methods

Since probability distribution of the load is considered across different cyclic granularity, there is no discarding of valuable data that might result from bringing all customers to the same time horizon. Moreover, since our metric is based on probability distribution, it will not be sensitive to outlier days and hence the clustering process will not be biased. Moreover, feature extraction applies to all days for all customers at once and therefore does not support parallel computing. Our metric is computed for each household separately and can even run for each each harmonies separately and thus fits well into the parallel computing framework.

5.0.3 Write the method in notations

- notation of time series for each customer
- wpd
- harmonies
- cluster

5.0.4 Distinction, repeatability, and robustness metrics

5.0.5 Putting similar households on linear scale

6 Discussion points and future work

Exploratory data analysis involve many iterations of finding and summarizing patterns. With temporal data available at ever finer scales, exploring periodicity has become overwhelming with so many possible granularities to explore. This work refines the selection of appropriate pairs of granularities by identifying

those for which the differences between the displayed distributions is greatest, and rating these selected harmony pairs in order of importance for exploration.

A future direction of work could be to look at more individuals/subjects and group them according to similar periodic behavior. Behaviors across different cyclic granularities would be different for different subjects and one way to find groups would be to actually locate clusters who have similar periodic behavior.

7 Appendix

7.1 Null distribution

7.1.1 Size: Simulated same distribution for all combinations of categories for all harmony pairs.

Failure to reject the null hypothesis when there is in fact no significant effect.

7.1.2 Normalised maximum distances follow standard Gumbel distribution

7.1.3 Limiting distribution of median of normalised maximum distances is normal

Let a continuous population be given with cdf F(x) (cumulative distribution function) and median ξ (assumed to exist uniquely). For a sample of size $2n+1$, let $\tilde{x}$ denote the sample median. The distribution of $\tilde{x}$,under certain conditions, to be asymptotically normal with mean ξ and variance $\sigma_n^2 = \frac{1}{4}[f(\xi)]^2(2n+1)$, where $f(x) = F'(x)$ is the pdf (probability density function).

7.2 Power

7.3 Confidence interval

Failure to reject the null hypothesis when there is in fact a significant effect.

To estimate the sampling distribution of the test statistic we need many samples generated under the null hypothesis. If the null hypothesis is true, changing the exposure would have no effect on the outcome. By randomly shuffling the exposures we can make up as many data sets as we like. If the null hypothesis is true the shuffled data sets should look like the real data, otherwise they should look different from the real data. The ranking of the real test statistic among the shuffled test statistics gives a p-value.

7.3.1 Varying distribution across facet

7.3.2 Varying distribution across x-axis

7.3.3 Varying distribution across both facets and x-axis

7.3.4 Repeat all with varying facet and x-axis levels

Conclusion: The test should reject the null hypothesis if distributions are different.

Dang, T N, and L Wilkinson. 2014. "ScagExplorer: Exploring Scatterplots by Their Scagnostics." In *2014 IEEE Pacific Visualization Symposium*, 73–80.

Department of the Environment and Energy. 2018. *Smart-Grid Smart-City Customer Trial Data*. Australian Government, Department of the Environment; Energy: Department of the Environment; Energy, Australia. https://data.gov.au/dataset/4e21dea3-9b87-4610-94c7-15a8a77907ef.

Gupta, Sayani, Rob J Hyndman, Dianne Cook, and Antony Unwin. 2020. "Visualizing Probability Distributions Across Bivariate Cyclic Temporal Granularities," October. http://arxiv.org/abs/2010.00794.

Kullback, S, and R A Leibler. 1951. "On Information and Sufficiency." *Ann. Math. Stat.* 22 (1): 79–86.

Menéndez, M L, J A Pardo, L Pardo, and M C Pardo. 1997. "The Jensen-Shannon Divergence." *J. Franklin Inst.* 334 (2): 307–18.

Tukey, John W, and Paul A Tukey. 1988. "Computer Graphics and Exploratory Data Analysis: An Introduction." *The Collected Works of John W. Tukey: Graphics: 1965-1985* 5: 419.

Wang, Earo, Dianne Cook, and Rob J Hyndman. 2020. "Calendar-Based Graphics for Visualizing People's Daily Schedules." *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2020.1715226.

Wilkinson, Leland, Anushka Anand, and Robert Grossman. 2005. "Graph-Theoretic Scagnostics." In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 157–64. IEEE.