A new metric for automatic discovery of periodic patterns in time series

Contents

1 Introduction

Exploratory data analysis, as coined by John W. Tukey (Tukey 1965) involves many iterations of finding structures and patterns that allow 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 in Figure (1) used in Wang, Cook, and Hyndman (2020b) for four households in Melbourne, Australia. The authors show how hour-of-the-day interact with weekdays 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 to 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, for example, how day-of-week relates to month-of-year, could also be interpreted from this display. This is a great way for having an overview of energy consumption. However, if one wants to understand the periodicities in energy behavior and how the periodicities interact in greater detail, 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 or month-of-year varies across week-of-the-month as well as with each other. Furthermore, 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. Possible areas where it is useful are monitoring heart rates which could record number of heartbeats every minute or analyze web search data for which data is available for a temporal scale as fine as second.

Even when it is known which all interactions to look at, all of them would not be interesting and that too will vary across different households. For example, area distribution quantiles are plotted for household 2 and 4 in Figure 2a and 2b respectively across month-of-year on facets and hour-of-day on x-axis. For the first household, the 75th and 90th percentile for January and February are very close, implying that energy usage for these months are generally on a much higher side, possibly due to the usage of air conditioners (January and February are peak summer in Australia), however for other months (Autumn and winter), the difference between the percentiles are not high, implying this household do not use so much heater as compared to air conditioner. Also, a lot of households in Victoria use gas heating and hence the usage of heaters might not be reflected here. The energy consumption for household 2 is also higher in summer months relative to autumn or winter, but the 75th and 90th percentile are far apart in all months, implying that the second household resorts to air conditioners much less regularly than the first one. Moreover, the 75th percentile distribution does not follow the same pattern as 90th percentile for all months for the first household, whereas, the pattern looks pretty similar for all months for the second household. Difference in the energy consumption could vary both across month-of-year (facets) and hour-of-day (x-axis). For the first household, both these cyclic granularities would deem important. Although, it seems like energy consumption across hours of the day are not that different across different months for the second household or atleast differences seem to be more prominent across month-of-year (facets) than hour-of-day (x-axis). It could be immensely useful to make the transition from all possible ways to only ways that could potentially be informative.

The paper Gupta et al. (2020) describes how we can compute all possible combinations of cyclic time granularities. Let N_C be the total number of contextual circular, quasi-circular and aperiodic cyclic granularities that can originate from the underlying linear granularities. The graphical mapping is such that distributions of a numeric response variable is displayed 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 problem 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, even for 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 in density, shape, trend, and other features in the 2D point scatters.

This work is a natural extension of our work (Gupta et al. (2020)), which 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. Buja et al. (2009) and Majumder, Hofmann, and Cook (2013) present some methods to quantify the strength of pattern and noise through visual inference, similar to numerical testing. But this is an evolving field as human cognition (which act as the statistical tests in visual inference) might vary across humans even for the same plots. In this paper, we build a new distance

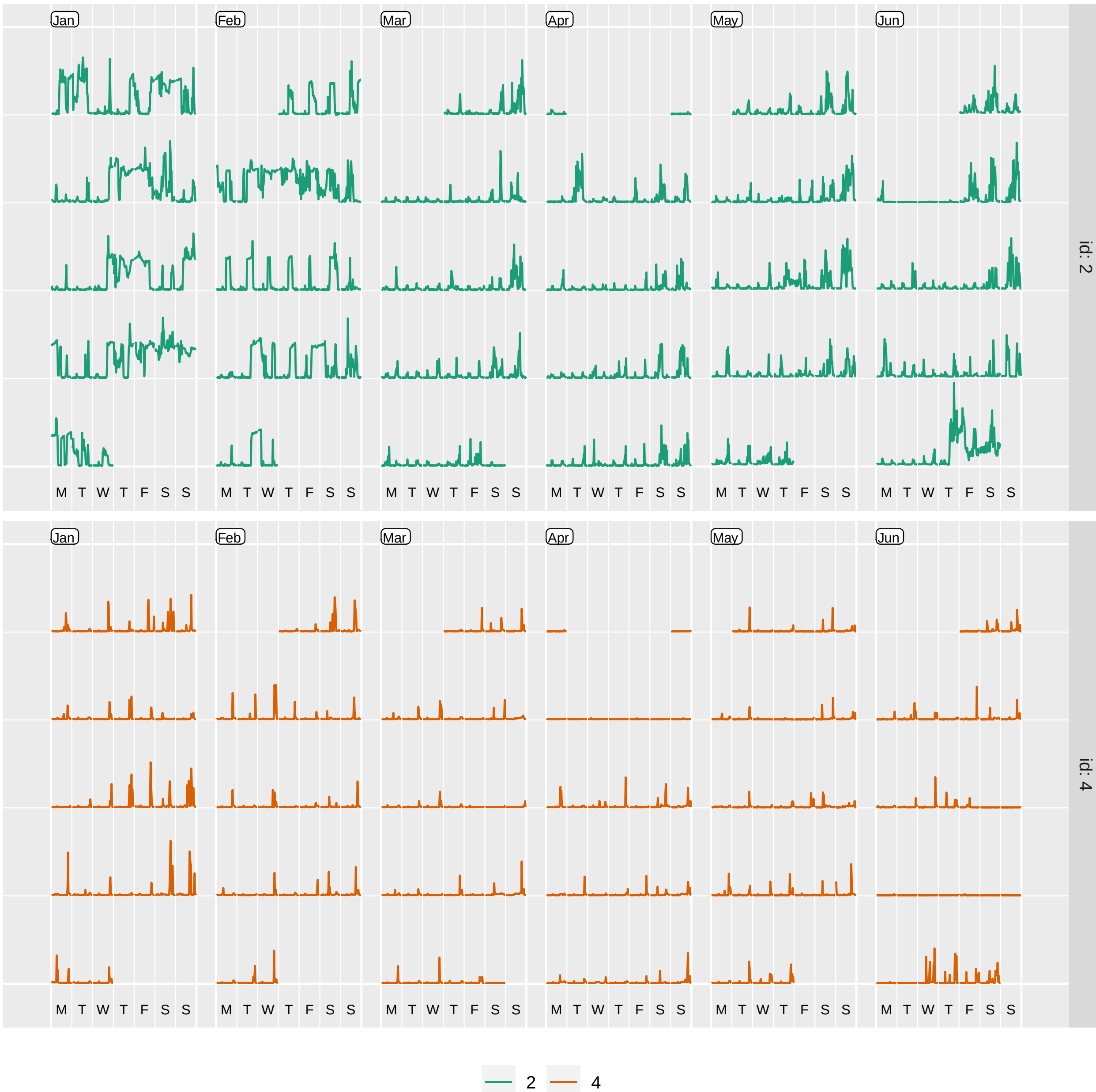


Figure 1: Calendar displays faceted by two households are shown. The calendar displays are rich in information and has several components. For example, it can be observed easily that the level of energy consumption by household 2 is higher than household 4. Also, it gives an overview on when these households are away for holidays and which months consumptions are highest (Jan, Feb). However,when it comes to discerning periodic patterns, analyzing all possible periodic patterns through this display is overwhelming due to combination of linear and cyclic representation of time. For example, there is more going on in terms of peaks and troughs in id 2 compared to id 4, but it is not clear which all periodic patterns are dominant and if it differs between households.

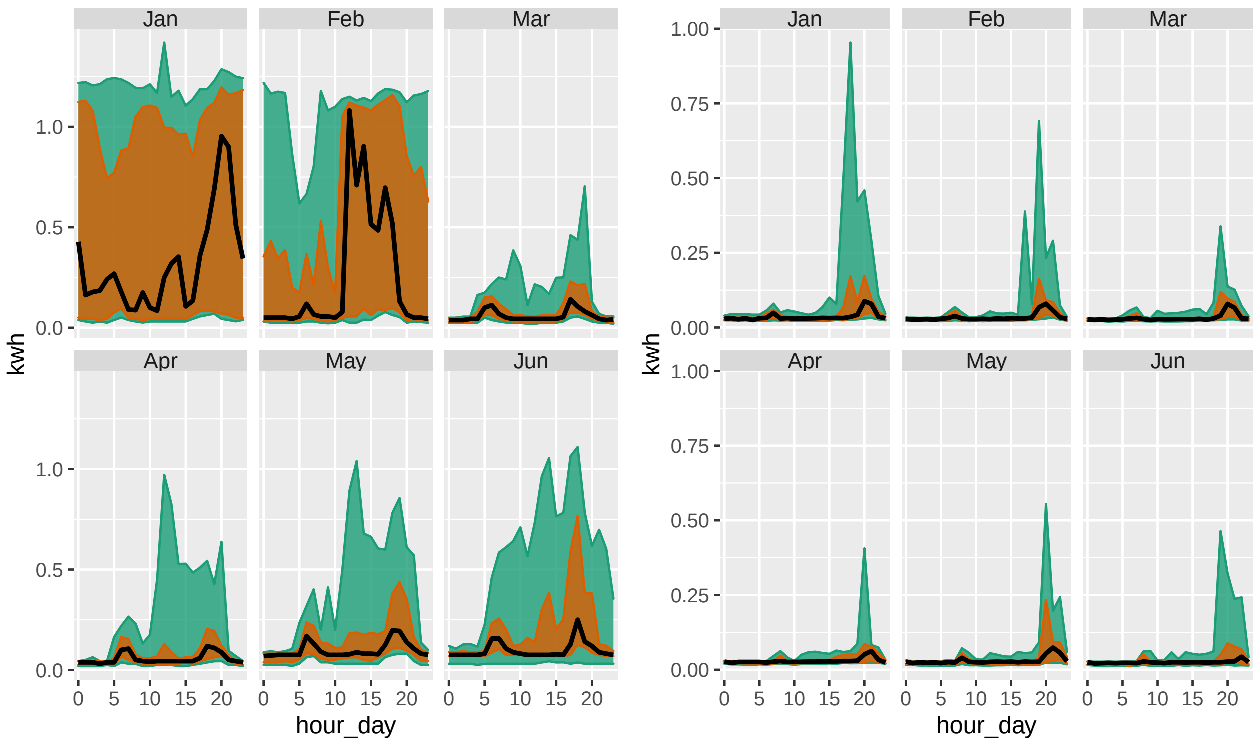


Figure 2: Distribution of energy consumption displayed through area quantile plots across two cyclic granularities month-of-year and hour-of-day for id2 (a) and id4 (b). The black line is the median, whereas the purple band covers the 25th to 75th percentile, the orange band covers the 10th to 90th percentile, and the green band covers the 1st to 99th percentile. Difference between 90th and 75th quantiles less for (Jan, Feb) in a) suggesting that it is a more frequent user of air conditioner than b). Energy consumption in a) changes across both granularities, whereas for b) daily pattern stays same irrespective of the months.

measure which could be used to detect automatically detect significant periodic patterns.

Our contributions in this paper are:

- We introduce a new distance measure for detecting periodic interactions. This induces data reduction which allows for identification of patterns, if any, in the time series data.
- We show that the distance metric could be used to rank the periodic patterns based on how well they capture the variation in the measured variable as they have been normalized for different number of comparisons.
- We device a framework for choosing a threshold, which will result in detection of only significantly interesting periodic patterns in the time series data.

The article is organized as follows. Section 2 introduces a new distance measure, discusses the reasoning behind choosing such a measure and presents some results to study the behavior of the measure. Section 3 describes a methodology to normalize the distance measure so that it can qualify as a measure that can be compared across different comparisons and datasets. Section 4 discusses how to choose a threshold to select only significant harmonies. Section 5 presents an application to a residential smart meter data in Melbourne to show how this distance measure acts as a way to automatically detect periodic patterns in time series.

2 A distance measure for quantifying patterns in harmonies

We are interested in assessing the structure of the measured variable across bivariate cyclic granularities. We propose a measure called Weighted Maximum Pairwise Distances (wpd) to quantify the structure in such a design. The principle employed towards this goal is explained through a simple example explained in Figure 3. Each of these figures have the same panel design with 2 x-axis categories and 3 facet levels. Figure 3a 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 3b has x categories drawn from the same distribution within a facet but the mean has been incremented by 5 units for every consecutive facets. Figure 3c exhibits an exact opposite situation where distribution between the x-axis categories within each facet is different but they are same across facets. For this situation, mean of only the x-axis categories are increased by 5 units for each consecutive category. Figure 3d takes a step further by varying the distribution across both facet and x-axis categories. If the displays are to be ranked in order of importance from minimum to maximum, then an obvious choice would be placing a followed by b, c and then d. It might be argued that it is not clear if b should precede or succeed c in the ranking. 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 in mind, display b is considered less informative as compared to display c. Hence, with reference to the graphical design in Gupta et al. (2020), 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.

Intuitively, while finding a structure or measuring the strength of patterns in Figure 3, it makes sense to look for within-group and between-group variations. Larger variation would imply stronger patterns, whereas small variation would imply the underlying structure is not changing within or between group. Thus, a distance measure aimed to capture this structure should ideally estimate these within-group and between-group variations. One of the potential ways to do this is to measure the distances between distributions of the continuous random variable measured within and between groups, weigh them basis if they are within or between groups and then take the maximum of those distances as an estimate of the maximum variation in the structure. This section starts with possible ways of characterizing distributions and computing distances between them and then describe in details how the measure *wpd* is defined. This is similar to Hyndman, Liu, and Pinson (2018) where the authors compute the Jensen Shannon distance between two density estimates by computing percentiles and stresses the advantages to working with percentiles rather than the raw data directly in case of missing observations. Working with quantiles also ensure that unsynchronized time series could be handled.

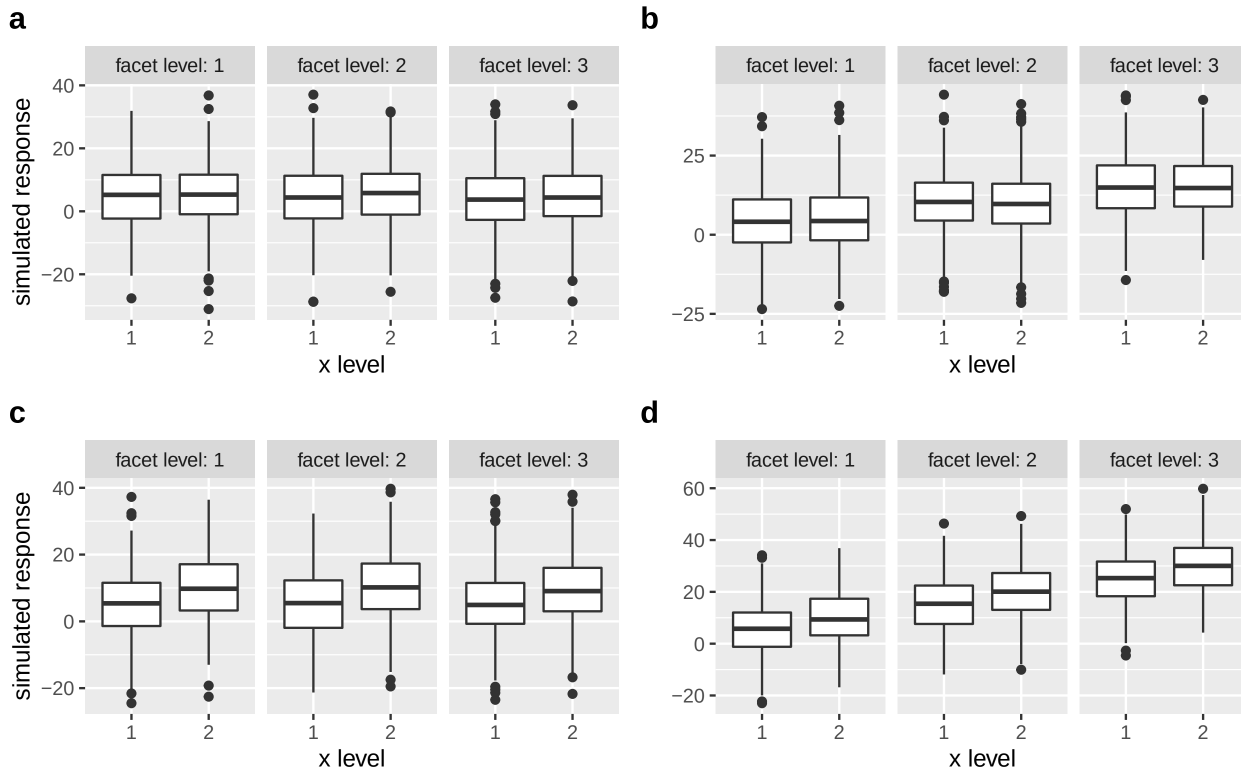


Figure 3: A graphical display with two categories mapped to x-axis and 3 categories mapped to facets with the distribution of a continuous random variable plotted on the y-axis. Display a is not interesting as the distribution of the variable does not depend on x or facet categories. Display b and c are more interesting than a since there is a change in distribution either across facets (b) or x-axis (c). Display d is most interesting in terms of displaying the strongest pattern as distribution of the variable changes across both facet and x-axis variable.

2.1 Notations

Consider two cyclic granularities A and B, such that $A = \{a_j : j = 1, 2, \dots, J\}$ and $B = \{b_k : k = 1, 2, \dots, K\}$ with A placed across x-axis and B across facets. Let $v = \{v_t : t = 0, 1, 2, \dots, T-1\}$ be a continuous variable observed across T time points. Let the four elementary designs as described in Figure 3 be D_{null} where there is no difference in distribution of v for A or B, D_{var_f} denotes the set of designs where there is difference in distribution of v for B and not for A. Similarly, D_{var_x} denotes the set of designs where difference is observed only across A. Finally, $D_{var_{all}}$ denotes those designs for which difference is observed across both A and B.

Table 1: Nomenclature table

variable	description
nx	number of x categories
nfacet	number of facet categories
N_C	number of cyclic granularities
H_{N_C}	set of harmonies
lambda	tuning parameter
omega	mean increment

variable	description
wpd	raw distance measure
$nperm$	number of permutations for threshold/normalization
$nsim$	number of simulations
wpd_{norm}	normalized distance measure
$wpd_{threshold}$	threshold for significance
D_{null}	null design
D_{var_f}	design with varying facets only
D_{var_x}	design with varying x only
$D_{var_{all}}$	design with varying both facets and x

2.2 Characterising distributions

Multiple observations of v correspond to the subset $v_{jk} = \{s : A(s) = j, B(s) = k\}$. The number of observations and the structure might vary widely across subsets due to the structure of the calendar, missing observations or uneven locations of events in the time domain. Each v_{jk}'s $\forall j \in \{1, 2, \ldots, J\}, k \in \{1, 2, \ldots, K\}\}$ are assumed to be drawn from a continuous probability distribution and have certain characteristics. Often shape, central tendency, and variability are the common characteristics used to describe a distribution. Mean, median or mode are generally used to describe the center of the distribution, while range, standard deviation, quantiles, standard errors and confidence intervals are used to describe variability. Quantiles are chosen as a way to characterize distributions in this paper.

The quantile of a distribution with probability p is defined as $Q(p) = F^{-1}(p) = inf\{x : F(x) > p\}$, $0 < p < 1$ where $F(x)$ is the distribution function. There are two broad approaches to quantile estimation, viz, parametric and non-parametric. The benefit of using a non-parametric estimator is that there are less rigid assumptions made about the nature of the underlying distribution of the data. Sample quantiles could be used for estimating population quantiles in a non-parametric setup. Hyndman and Fan (1996) describes the many ways of defining sample quantiles and recommends the use of median-unbiased estimator because of *desirable properties of a quantile estimator and can be defined independently of the underlying distribution.*. The `stats::quantile()` function in R Core Team (2019) could be used for practical implementation where type = 8 refers to the algorithm corresponding to the median-unbiased estimator. The default quantile chosen in this paper is percentiles computed for $p = 0.01, 0.02, \ldots, 0.99$, where for example, the 99^{th} percentile would be the value corresponding to $p = 0.99$ and hence 99% of the observations would lie below that.

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 by $D(q_1||q_2)$ is a non-symmetric measure of the difference between two probability distributions q_1 and q_2 and is interpreted as the amount of information lost when q_2 is used to approximate q_1. 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 between distributions are Hellinger distance, total variation distance and Fisher information metric.

In this paper, we propose to use the pairwise distances between the distributions of the measured variable through Jensen-Shannon distance (JSD), defined by,

$$JSD(q_1||q_2) = \frac{1}{2}D(q_1||M) + \frac{1}{2}D(q_2||M)$$

where $M = \frac{q_1+q_2}{2}$ and $D(q_1||q_2) := \int_{-\infty}^{\infty} q_1(x) f(\frac{q_1(x)}{q_2(x)})$ is the KL divergence between distributions q_1 and q_2.

Furthermore, these distances are distributed as chi-squared with m degrees of freedom (Menendez1997-in), if the continuous distribution is being discretized with m discrete values. Taking sample percentiles to approximate the integral would mean taking $m = 99$. As the degrees of freedom m get larger, the chi-square distribution approaches the normal distribution.

2.4 Computation

The distance measure *wpd* between two cyclic granularities A and B is aimed to capture the strength of the structure by estimating the maximum within-group and between-group variations. Furthermore, the intended aim of *wpd* is to capture differences in categories irrespective of the distribution from which the data is generated. Hence, as a pre-processing step,the raw data is normal quantile transformed (reference) so that the quantiles of the transformed data follows a standard normal distribution. The steps employed for computing the distance measure is summarized as follows:

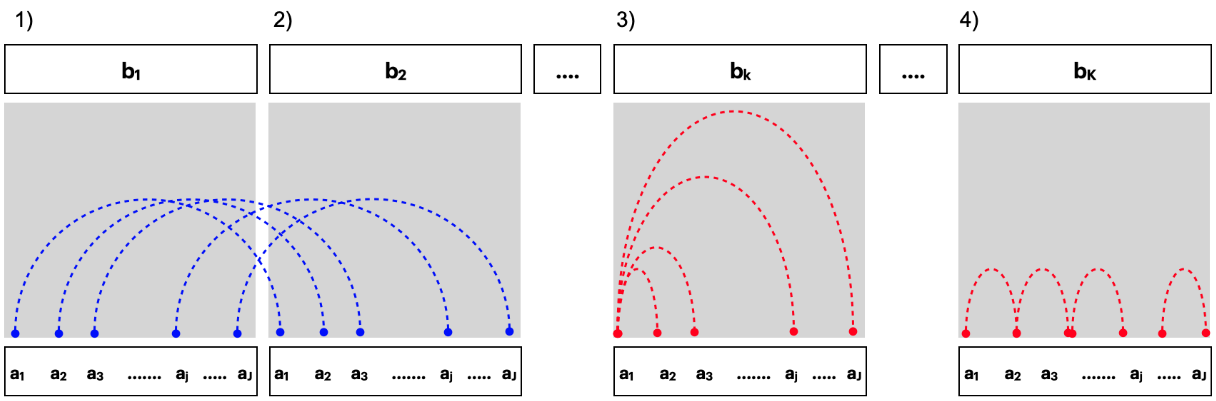


Figure 4: Within and between-facet distances shown for two cyclic granularities A and B, where A is mapped to x-axis and B is mapped to facets. The dotted lines represent the distances between different categories. Panel 1) and 2) show the between-facet distances. Panel 3) and 4) are used to illustrate 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).

1. Perform NQT on the measured variable v_t to obtain v^*_t.
2. Fix harmony pair (A, B).
3. Percentiles of v^*_{jk} are computed and stored in q_{jk}. Repeat for all pairs of categories of the form $(a_j b_k, a_{j'} b_{k'}) : \{a_j : j = 1, 2, \dots, J\}, B = \{b_k : k = 1, 2, \dots, K\}$.
4. The pairwise distances between pairs $(a_j b_k, a_{j'} b_{k'})$ denoted by $d_{(jk,j'k')} = JSD(q_{jk}, q_{j'k'})$ is computed.
5. The pairwise distances $d_{(jk,j'k')}$ is transformed using a suitable tuning parameter $(0 < \lambda < 1)$ depending on if they are within-facet(d_w) or between-facets(d_b) as follows:

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

5. The wpd is then computed as $wpd = max_{j,j',k,k'}(d*_{(jk),(j'k')}) \forall j, j' \in \{1, 2, \dots, J\}, k, k' \in \{1, 2, \dots, K\}$.
6. Repeat Steps 2-5 for all harmony pairs in H_{N_C}.

Pairwise distances could be within-facets or between-facets. Figure 4 illustrates how the within-facet or between-facet distances are defined. Pairwise distances are within-facets 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 4. 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 when they are considered between pairs of the form $(a_j b_k, a_j b_{k'})$. Number of between-facet distances would be ${}^K C_2 * J$ and number of within-facet distances are $K * (J - 1)$ (ordered) and ${}^J C_2 * K$ (un-ordered). If the measure is intended to put more importance in pointing towards distributional differences between x categories, a $\lambda > 0.5$ should be chosen.

2.5 Properties

Simulations were carried out to explore the behavior of *wpd* under the following factors that could potentially impact the values of *wpd*: nx, $nfacet$, λ, ω, *dist* (normal/non-normal distributions with different location and scale), *ntimes*, and *designs* and results are presented in two parts. The dependence of *wpd* on nx and $nfacet$ under D_{null} is presented here, which lays the foundation for the next section. The rest of the results that discuss the relationship of the *wpd* with other factors is presented in details in the Supplementary section of the paper. They show that the designs D_{var_f} and D_{var_x} intersect at $\lambda = 0.5$ and hence for up-weighing designs of the form D_{var_x}, $\lambda = 0.67$ has been considered for computation of *wpd* in the rest of the paper.

2.5.1 Simulation design

Observations are generated from a Gamma(2,1) 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. Each combination is being referred to as a *panel*. That is, data is being generated for each of the panels $\{nx = 2, nfacet = 2\}, \{nx = 2, nfacet = 3\}, \{nx = 2, nfacet = 5\}, \dots, \{nx = 50, nfacet = 31\}, \{nx = 50, nfacet = 50\}$. 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 *wpd* is obtained for each of the panels. This design corresponds to D_{null} as each combination of categories in a panel are drawn from the same distribution. Furthermore, the data is simulated for each of the panels $nsim = 200$ times, so that the distribution of *wpd* under D_{null} could be observed. $wpd_{l,s}$ denotes the value of *wpd* obtained for the l^{th} panel and s^{th} simulation.

2.5.2 Results

Figure 5 shows the distribution of *wpd* plotted across different nx and $nfacet$ categories. Since under D_{null}, there is no difference in distributions across different categories, we expect the distance measure *wpd* to reflect that as well and have the same distribution across categories. But Figure 5 shows that both the location and scale of the distributions change across panels. This is not desirable under D_{null} as it would mean comparisons of *wpd* values is not appropriate across different nx and $nfacet$. Figure 6 shows how the median of *wpd* varies with the total number of distances $nx * nfacet$ for each panel. The median increases abruptly for lower values of $nx * nfacet$ and slowly for higher $nx * nfacet$.

3 Normalization for the number of comparisons

The distribution of *wpd* is different for different levels of facets and x-axis levels. This is because the statistics maximum which is used to define *wpd* is affected by the number of comparisons (resulting pairwise distances).

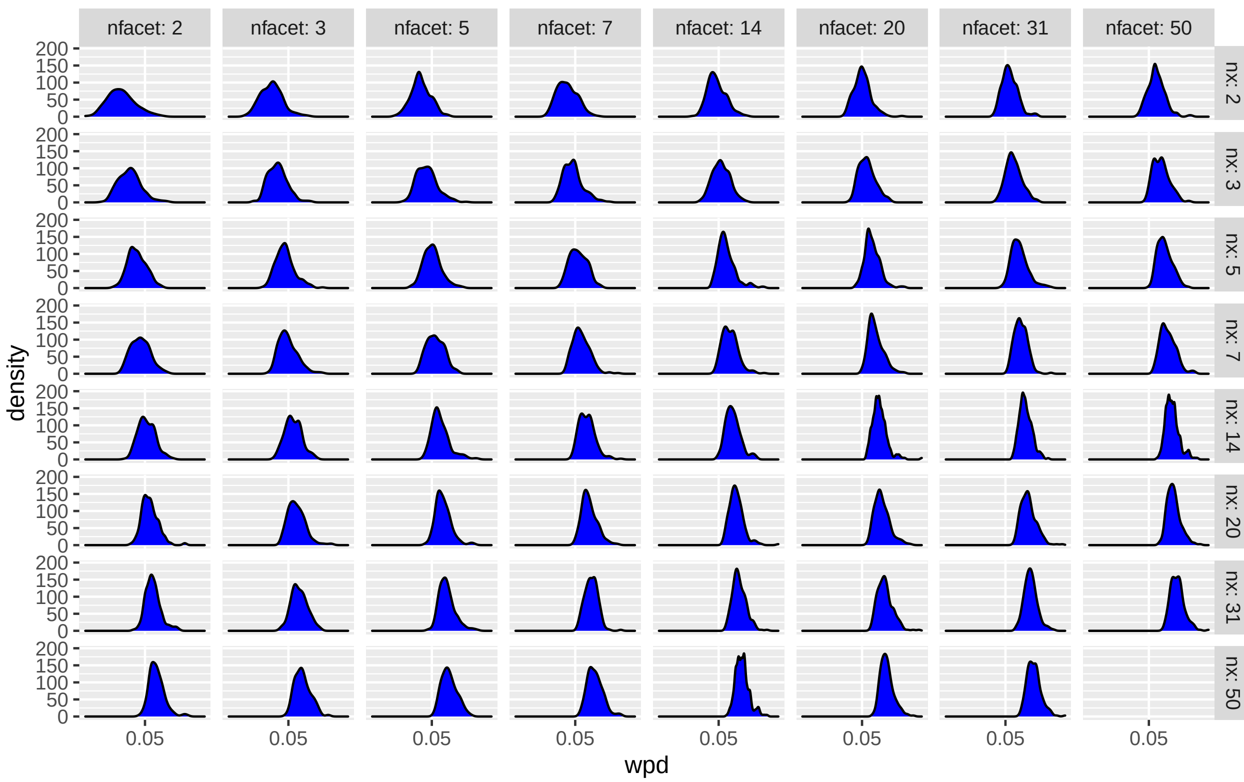


Figure 5: Distribution of wpd is plotted across different nx and $nfacet$ categories under D_{null}. Both shape and scale of the distribution changes for different comparisons. This is not desirable since under null design, the distribution of the distance measure is not expected to capture any differences.

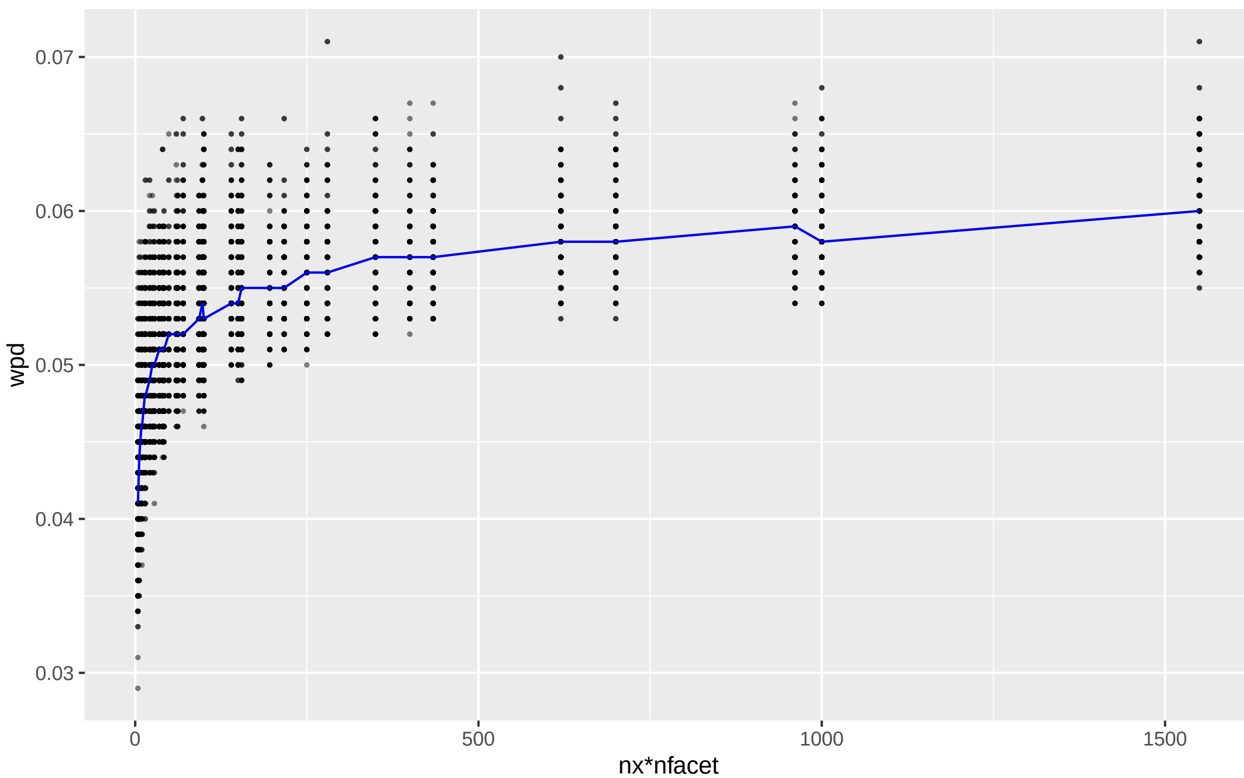


Figure 6: wpd is plotted against $nx * nfacet$ (the maximum number of pairwise comparisons) and the blue line represents the median of the multiple values for each $nx * nfacet$. The median increases abruptly for lower values of $nx * nfacet$ and slowly for higher $nx * nfacet$. Thus, the measure will have higher values for higher levels in nx or $nfacet$.

The measure would have higher values if A or B has higher levels. However, we would ideally want a higher value of the measure only if there is a significant difference between distributions across facet or x-axis categories, and not because the number of categories J or K is high. Therefore, in order to compare wpd across different combinations of facet and x-axis levels, we need to eliminate the impact of different number of comparisons and get a normalized measure. Henceforth, we call the normalized measure as wpd_{norm}. The measure wpd_{norm} could potentially lead to comparison of the measure across different panels and also help distinguishing the interesting panels from a data set. We discuss two approaches for normalization, both of which are substantiated using simulations.

3.1 Methodology

The transformed wpd which is normalized for the values of nx and $nfacet$ is denoted by wpd_{norm}. Two approaches have been employed - the first one involves a permutation method to make the distribution of the transformed wpd similar for different comparisons and the second one fits a model to represent the relationship between the two variables and defines wpd_{norm} as the residual of the model.

3.1.1 Permutation approach

This method is somewhat similar in spirit to bootstrap (reference) or permutation (reference) tests, where the goal is to test the hypothesis that the groups under study have identical distributions. This method, in essence, accomplishes a different goal of making the location and scale of different groups (panels) same under D_{null}. The steps are as follows:

1. Compute wpd for a harmony pair (A, B) for the original measured variable v_t and store it in wpd_{orig}.
2. Consider a permutation of the original measured variable v_{perm_1} and again compute wpd for the permuted data. Store it in wpd_{perm_1}.
3. Repeat Step 2 for a large number ($nperm = 200$) of random permutations of the data yielding $nperm$ values : $wpd_{perm_1}, wpd_{perm_2}, \ldots, wpd_{perm_{nperm}}$. Store the vector in wpd_{perm}.
4. Define $wpd_{perm} = \frac{(wpd_{orig} - \bar{wpd_{perm}})}{sd(wpd_{perm})}$, where $\bar{wpd_{perm}}$ and $sd(wpd_{perm})$ are the mean and standard deviation of wpd_{perm} respectively.

Standardizing wpd in the permutation approach ensures that the distribution of wpd_{perm} has the same $mean = 0$ and $sd = 1$ across all comparisons under D_{null}. While this works successfully to make the location and scale similar across different nx and $nfacet$ (as seen in Figure 7), it is computationally heavy and time consuming, and hence less user friendly when being actually used in practice. Hence, we propose another approach to normalization which is more approximate than exact but still has the similar accuracy when compared to the permutation approach.

3.1.2 Modelling approach

Linear model

A linear model is fitted to model the relationship between wpd and $nx * nfacet$. The model is of the form

$$y_l = a + b * log(z_l) + e_l$$

, where, $y_l = median_m(wpd_{l,m})$, z_l is the l^{th} panel and e_l are idiosyncratic errors, with parameters a and b being estimated from the data generated through the simulation study described in Section 2.5.1. The estimates and other model summary is given in Table 2.

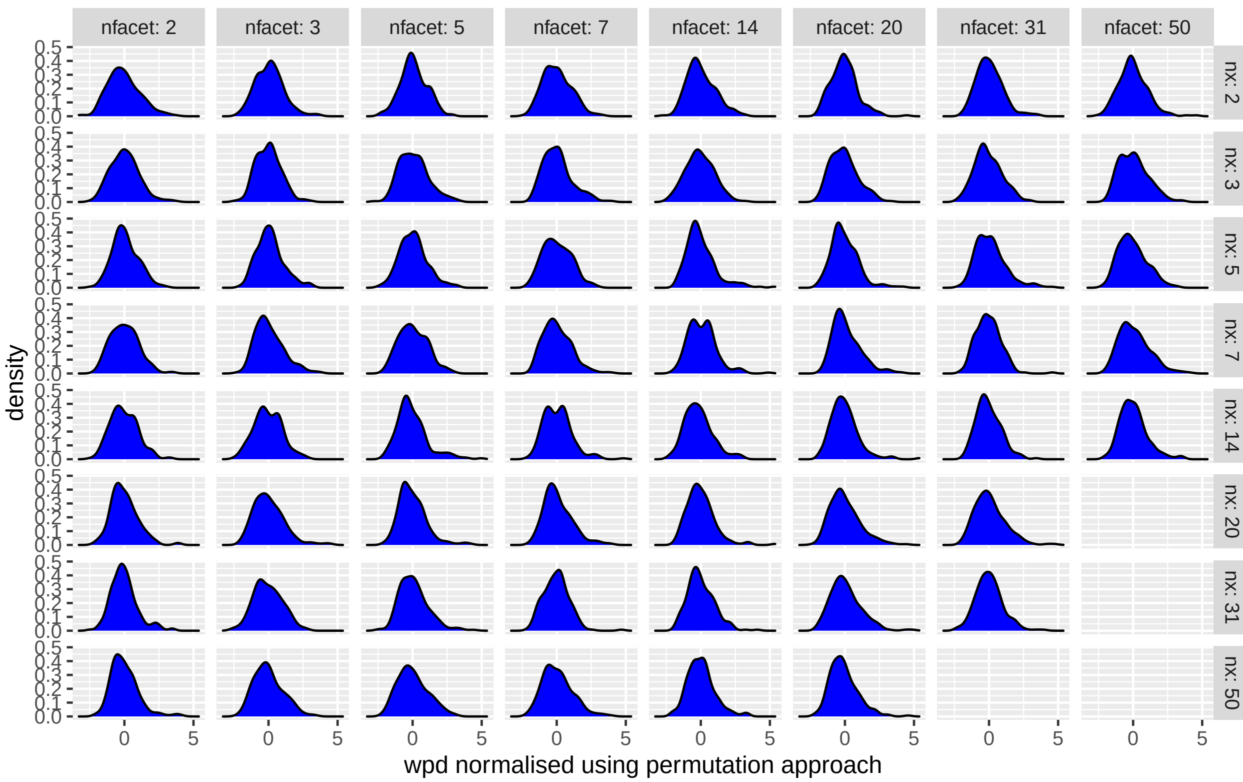


Figure 7: Distribution of wpd_{perm} is plotted across different nx and $nfacet$ categories. Both shape and scale of the distributions are now similar for different panels under the null design.

Table 2: Results of linear model to capture the relationship between wpd and number of comparisons.

term	estimate	std.error	statistic	p.value
(Intercept)	0.0403697	0.0003814	105.85081	0
log(‘nx * nfacet‘)	0.0027565	0.0000804	34.27252	0

The idea is to find a transformation on wpd which would remove the effect of $nx * nfacet$ on wpd and thus is defined as the residuals: $y^* = y - \hat{a} - \hat{b} * log(z)$, where y^* is the wpd_{linear}, $\hat{a}$ and $\hat{b}$ are the estimated values of the parameter a and b and $z = nx * nfacet$.

Defining wpd_{linear} in this way forces the mean to be zero and variability to be uniform across the median wpd values as could be seen in Figure 8.

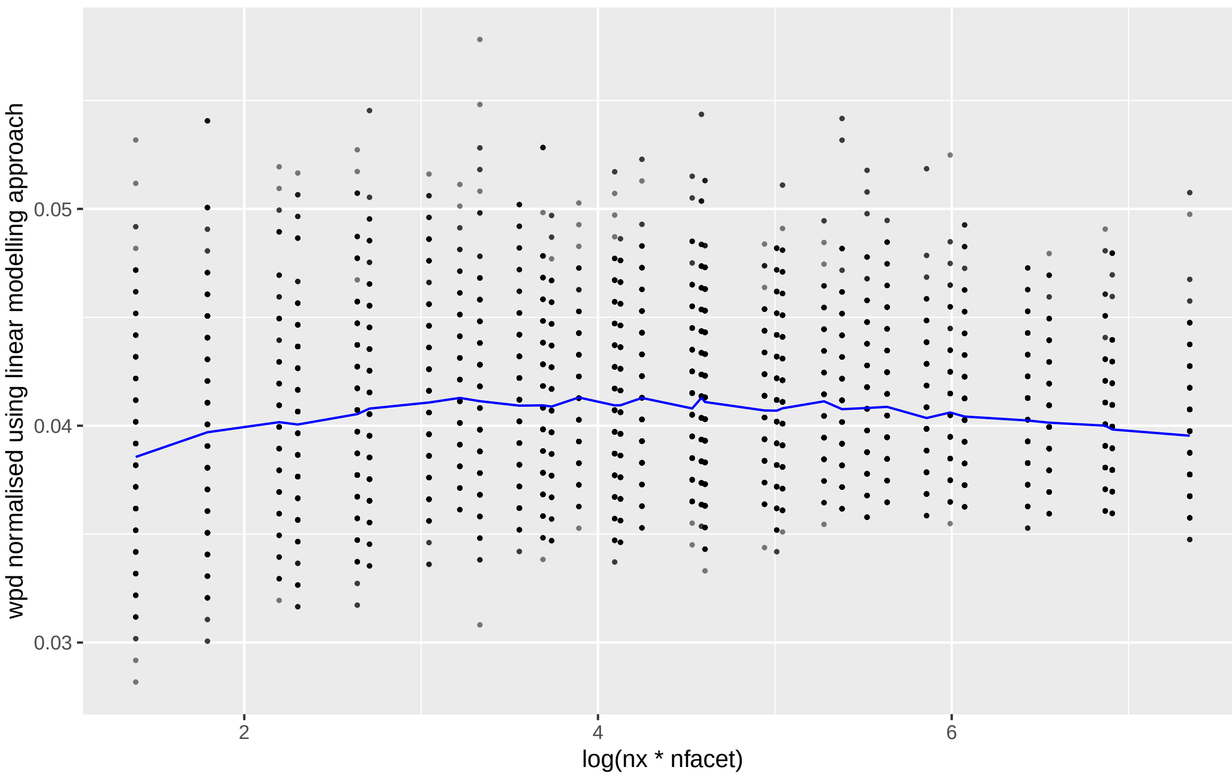


Figure 8: wpd_{norm}^{linear} which are defined as the residuals of the linear model have mean zero and are homogenous. Further, by design they are independent of the $nx * nfacet$ and hence could be a potential candidate for $wpd_n orm$.

Generalized linear model

In the linear model approach, $wpd \in R$ was assumed, whereas, wpd is a Jensen-Shannon Distance (JSD) and lies between 0 and 1 (reference). Furthermore, JSD follows a Chi-square distribution, which is a special case of Gamma distribution. Therefore, a generalized linear model could be fitted 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 $0 \leq wpd \leq 1$.

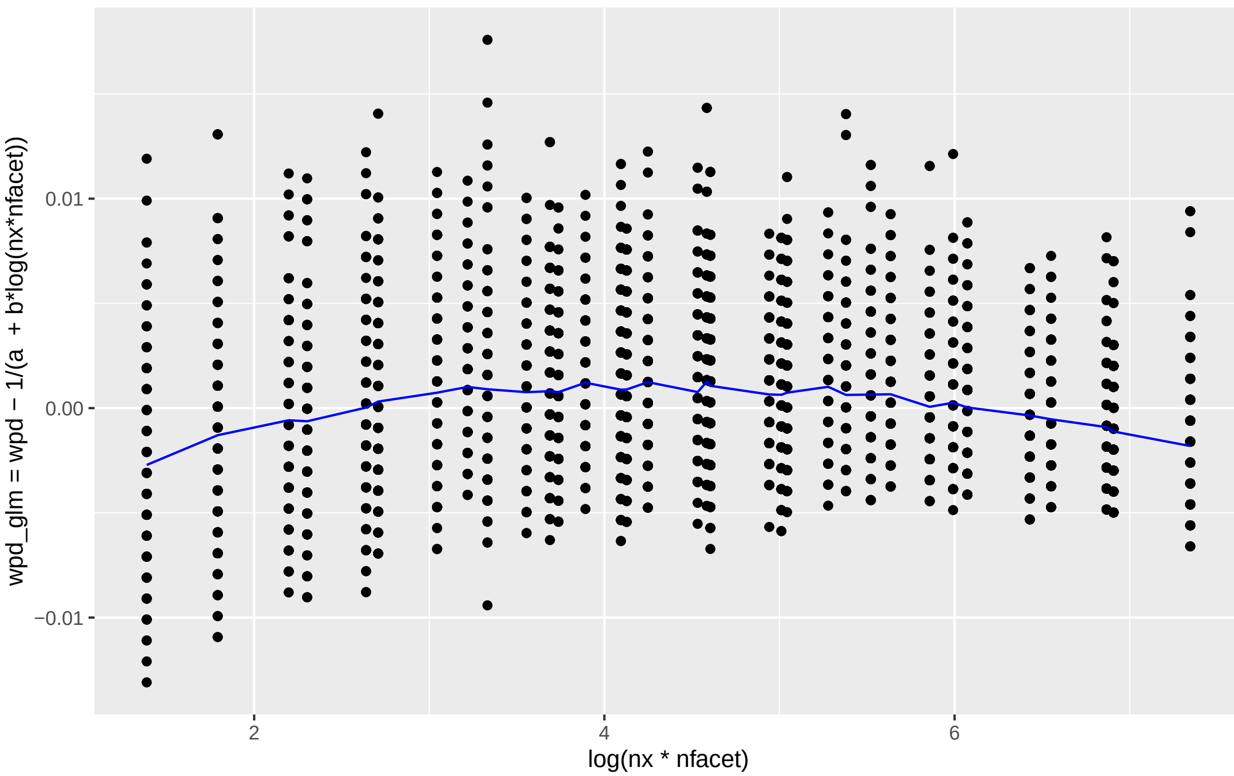
We fit a Gamma generalized linear model with the inverse link which is of the form:

$$y_l = a + b * log(z_l) + e_l$$

, where $y_l = median_m(wpd_{l,m})$, z_l is the l^{th} panel and e_l are idiosyncratic errors. Let $E(y) = \mu$ and $a + b * log(z) = g(\mu)$ where g is the link function. Then $g(\mu) = 1/\mu$ and $\hat{\mu} = 1/(\hat{a} + \hat{b}log(z))$. The residuals from this model $(y - \hat{y}) = (y - 1/(\hat{a} + \hat{b}log(z)))$ would be expected to have no dependency on z. Thus, wpd_{glm} is chosen as the residuals from this model and is defined as: $wpd_{glm} = wpd - 1/(\hat{a} + \hat{b} * log(nx * nfacet))$.

Table 3: Results of generalised linear model to capture the relationship between wpd and number of comparisons.

term	estimate	std.error	statistic	p.value
(Intercept)	23.5587633	0.2207781	106.7079	0
log(‘nx * nfacet‘)	-0.9971299	0.0444104	-22.4526	0



3.2 Combining normalizing approaches

We see that the transformation through the modeling approach leads to very similar distribution across high nx and $nfacet$ (higher than 5) 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 $nfacet$. It is difficult to compare and use the transformed measure from both of these approaches alternatively without bringing them to the same scale. The transformed variables from the two approaches have to be brought to the same scale so that for smaller categories, permutation approach is used and for larger categories, modeling approach is used.

The measure wpd_{glm} has a $\hat{\mu}_{glm} = 0$ and $\hat{sd}_{glm} = 0.003$ whereas the measure wpd_{perm} which is a z-score, has an expected normal distribution with $\hat{\mu}_{perm} = 0$ and $\hat{sd}_{perm} = 1$. To bring them to the same scale, we have defined $wpd_{glm-scaled} = wpd_{glm} * \frac{\hat{sd}_{perm}}{\hat{sd}_{glm}}$, which changes the scale of wpd_{glm} 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_{perm} is also normally distributed. Further, they are brought to similar scale and location and hence could be compared or used interchangeably for different comparisons based on their performance.

Thus, the wpd_{norm} is defined as follows:

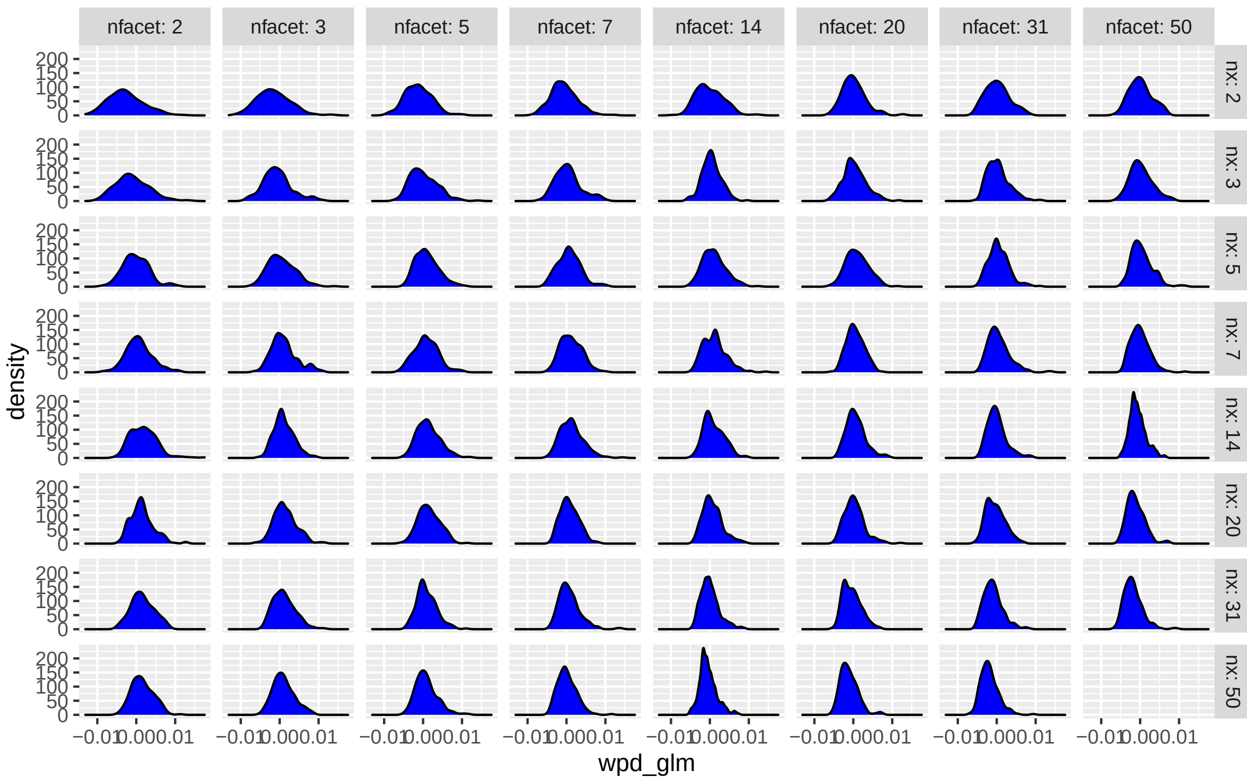


Figure 9: The distribution of wpd_{glm} is plotted. The distributions are more similar across higher nx and nfacet and dissimilar for fewer nc and nfacets.

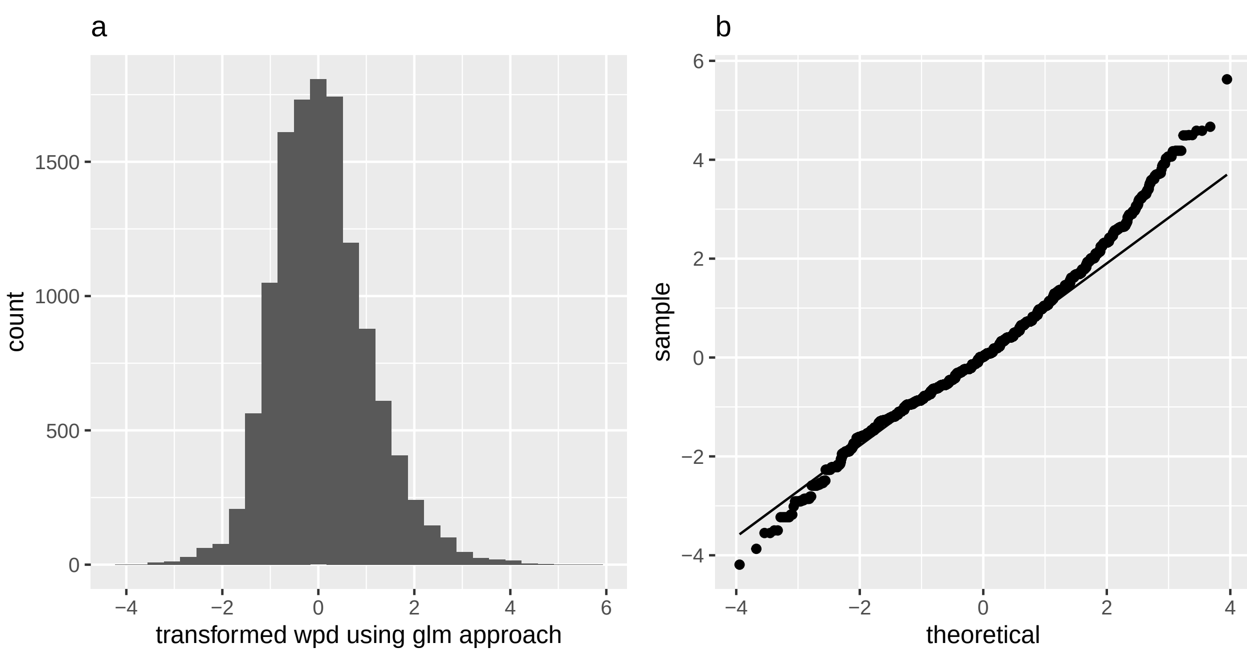


Figure 10: In panel a, the histogram of $wpd_{glm-scaled}$ is plotted. In part b, the QQ plot is shown with the theoretical quantiles on the x-axis and $wpd_{glm-scaled}$ quantiles on the y-axis. The distribution looks symmetric and looks like normal except in the tails.

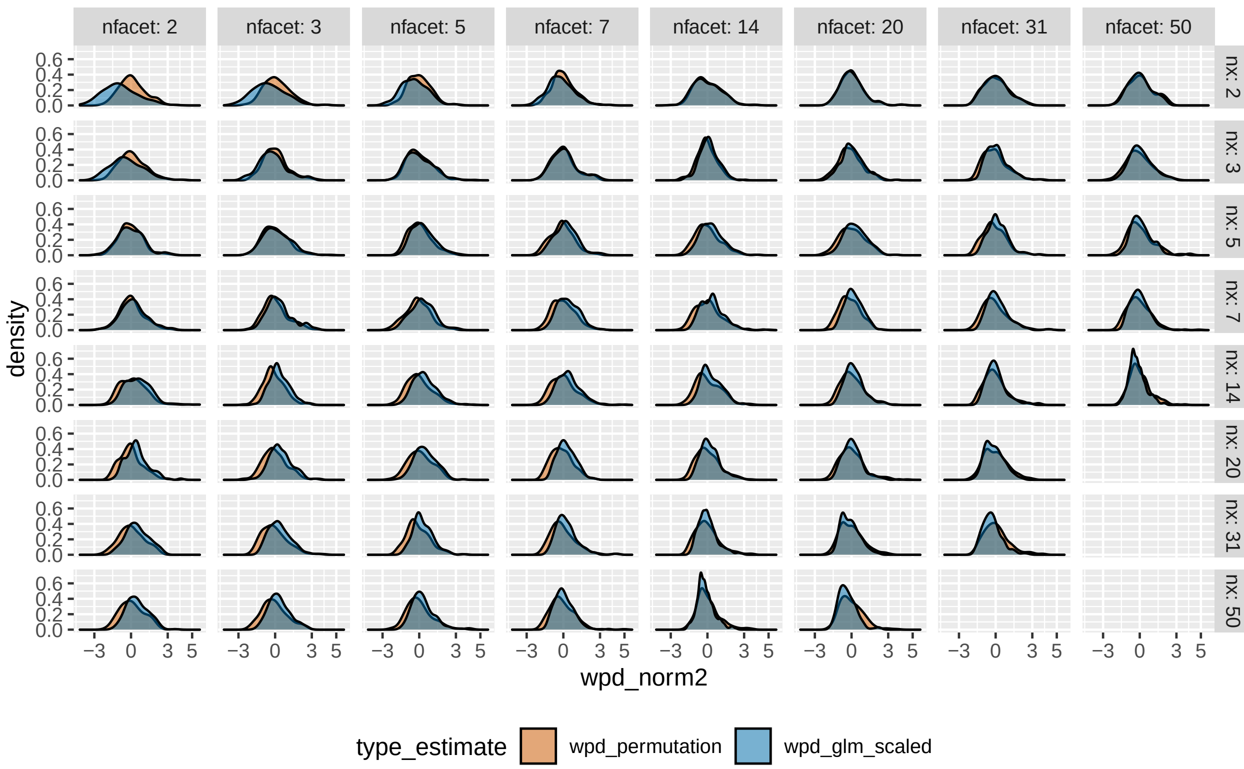


Figure 11: wpd_perm and $wpd_{glm-scaled}$ are plotted together on the same scale. They also have the same location and hence the values from these two approaches could be compared across panels. $wpd_{glm-scaled}$ would be used to normalise wpd_{raw} for higher nx and $nfacet$ and $wpd_{glm-scaled}$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

$$wpd_{norm} = \begin{cases} wpd_{perm}, & \text{if } J, K <= 5 \\ wpd_{glm-scaled} & \text{otherwise} \end{cases} \tag{2}$$

3.3 Properties

This section reports the results of a simulation study that was carried out to evaluate the behavior of wpd_{norm} under different designs and other potential factors. The behavior of wpd_{norm} 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 the Supplementary paper. The learning from the simulations are as follows: The values of wpd_{norm} 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_{norm} was designed such that this relationship holds. Furthermore, the distribution of the measure wpd_{norm} does not change for different facet and x categories. The distribution of wpd_{norm} looks 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 initial 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_{norm} could be found in the Supplementary paper.

4 Ranking and selecting significant harmonies

In this section, we provide a method to select important harmonies by eliminating all harmonies for which patterns are not significant across x or facet categories through randomization test. Randomization tests (permutation tests) generates a random distribution by re-ordering our observed data and allow to test if the observed data is significantly different from any random distribution. 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.

4.1 Choosing a threshold

Typically, a randomization test involves calculating a test statistic, randomly shuffling the data and calculating the test statistic several times to obtain a distribution of the test statistic. We will use this procedure to test if there is any interesting pattern captured by the harmonies, which essentially implies if wpd_{norm} is significantly different from zero. The percentages of times the wpd_{norm} obtained from the permuted data is greater than or equal to the observed wpd_{norm} is the p-value. The randomization test is described as follows:

- **Input:** All harmonies of the form $\{(A, B), A = \{a_j : j = 1, 2, \ldots, J\}, B = \{b_k : k = 1, 2, \ldots, K\}\}$ with A placed across x-axis and B across facets $\forall (A, B) \in N_C$.
- **Output:** Harmony pairs (A, B) for which wpd_{norm} is significant.

1. Fix harmony pair (A, B).
2. Given the data; $\{v_t : t = 0, 1, 2, \ldots, T-1\}$, the wpd_{norm} is computed and is represented by wpd_{obs}.
3. From the original sequence a random permutation is obtained: $\{v_t^* : t = 0, 1, 2, \ldots, T-1\}$.
4. wpd_{norm} is computed for the permuted sequence of the data and is represented by wpd_{perm_1}.
5. Steps (3) and (4) are repeated a large number of times M (M = 200).
6. For each permutation, one wpd_{perm_i} is obtained. Define $wpd_{sample} = \{wpd_{perm_1}, wpd_{perm_2}, \ldots, wpd_{perm_M}\}$.
7. Repeat Steps (1-6) for all harmony pairs $(A, B) \in H_{N_C}$ and store it in wpd_{sample}^{all}.
8. 95^{th} percentile of wpd_{sample}^{all} is computed and stored in $wpd_{threshold}$.
9. If $wpd_{obs_{A,B}} > wpd_{threshold}$, harmony pair (A, B) is selected, otherwise rejected.

4.2 Simulation design

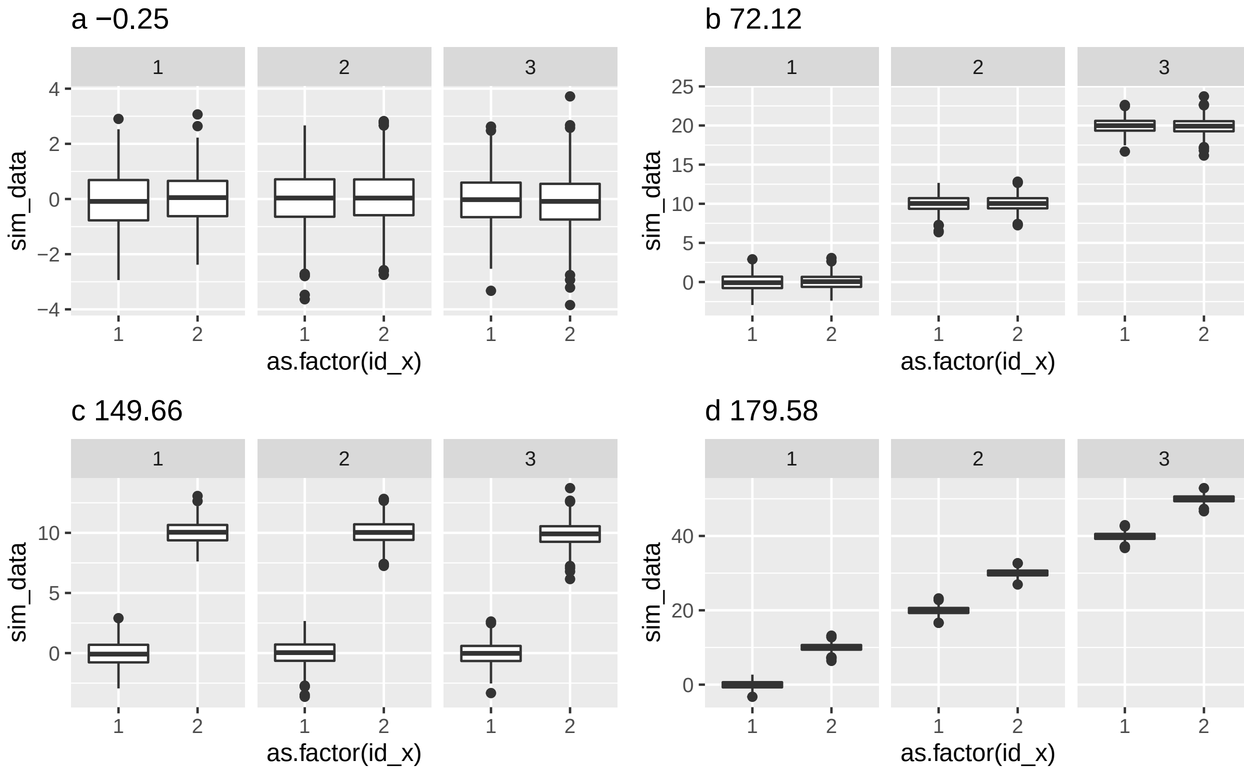
Observations are generated from a N(0,1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = nfacet = \{2, 3, 5, 7, 14, 20, 31, 50\}$. For each of the 64 panels, $ntimes = 500$ observations are drawn for each combination of the categories. The values of wpd_{norm} is obtained for each of the panels for the designs D_{null}, D_{var_x}, D_{var_f} and $D_{var_{all}}$. $wpd_{threshold}$ is computed from all of these panels together and the number of times a harmony pair $(A, B) \in H_{N_C}$ is selected when in fact it was of the design D_{null} is noted. This entire process is repeated for several null data sets to see the number of times any harmony pair $(A, B) \in H_{N_C}$ is selected under null.

nx = (3, 7, 14) nfacet = (2, 9, 10)

(3, 2), (7,9), (14, 10) being null sets and others under some other design. Generate 200 repetitions of different data sets with the same design for all the pairs and compute size and power.

4.2.1 Results

Figure 5 shows the distribution of wpd plotted across different nx and $nfacet$ categories. Since under D_{null}, there is no difference in distributions across different categories, we expect the distance measure wpd to reflect that as well and have the same null distribution across categories. But Figure 5 shows that both the location and scale of the distributions change across panels. This is not desirable under D_{null} as it would mean comparisons of wpd values is not appropriate across different nx and $nfacet$. Figure 6 shows how the median of wpd varies with the total number of distances $nx * nfacet$ for each panel. The median increases abruptly for lower values of $nx * nfacet$ and slowly for higher $nx * nfacet$.



4.3 Simulation environment

Simulation studies were carried out to study the behavior of *wpd*, build the normalization method as well as compare and evaluate different normalization approaches. R version 4.0.1 (2020-06-06) is used with the 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.

5 Application to residential smart meter dataset

The smart meter data set for eight households in Melbourne has been utilized to see the use of the wpd_{norm} proposed in the paper. The data has been cleaned to be a `tsibble` (Wang, Cook, and Hyndman (2020a)) containing half-hourly electricity consumption from Jul-2019 to Dec-2019 for each of the households, which is procured by them by downloading their data from the energy supplier/retailer. Demand data for these households are shown in a linear time scale in Figure 12. It is evident from the range of the demand data that these households vary in consumption levels as well as in their temporal patterns. In the left panel of Figure 12 (a), the linear representation of the entire time period is shown, whereas in the right panel (b) a particular month is shown and furthermore a week has been highlighted to inspect if there is any daily or weekly periodic patterns in their behavior that is reflected when we zoom into the the linear representation of the time series. We start the analysis by asking if the ranking of the harmonies make sense for the households, then compare households to get more insights of what these rankings imply and if they could be used to remove some non-interesting harmonies. Furthermore, we see if the display of the significant harmonies could be validated by zooming in the linear representation of the time series.

Choosing cyclic granularities of interest and removing clashes

Let $v_{i,t}$ denote the electricity demand for i^{th} household for time period t. The series $v_{i,t}$ is the linear granularity corresponding to half-hour since the interval of this data is 30 minutes. We consider coarser

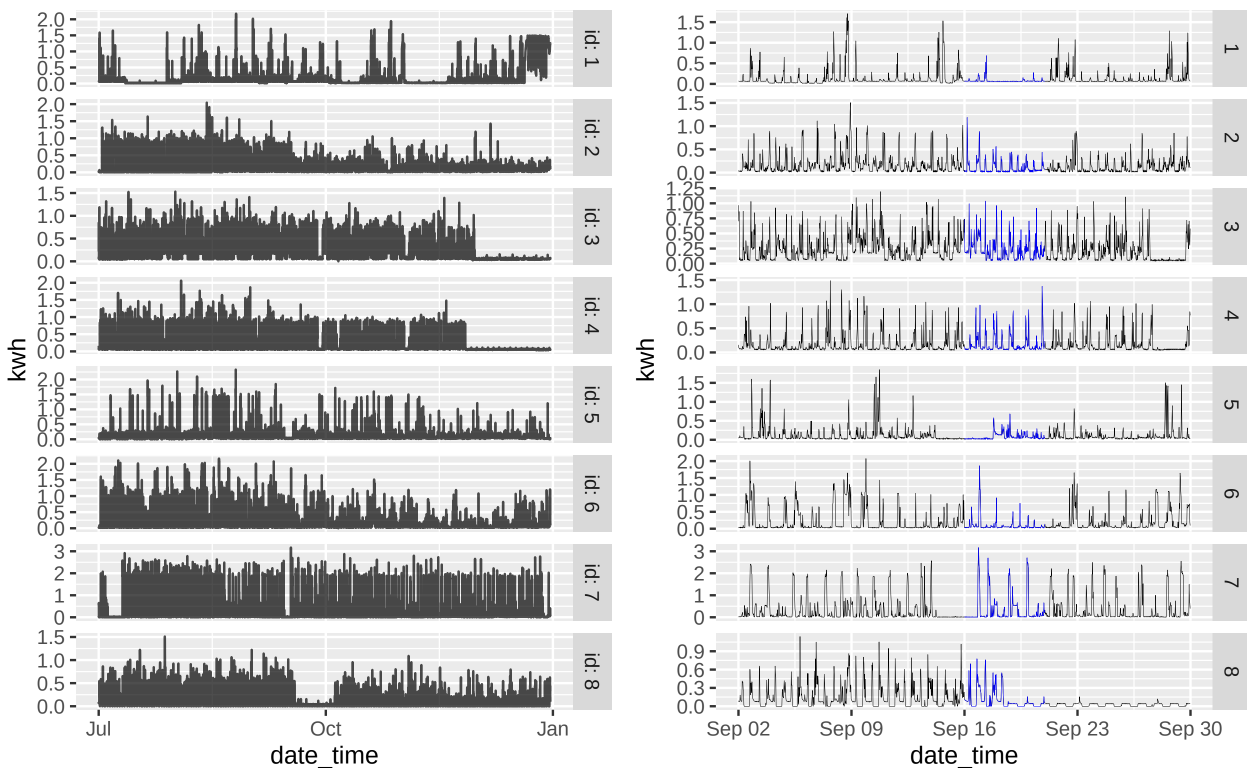


Figure 12: Electricity demand for eight households are shown in different facets from Jul-19 to Dec-19 in Fig a and it has been zoomed in for Sep-19 in Fig b, where a week in Sep-19 has been highlighted. From the scales of Fig a, it is apparent that they have different level of consumption but all of them have some periodic behavior in terms of regular peaks and troughs. It is not clear which all periodic patterns exist. In fig b, periodic pattern is zoomed in for a month and we can see weekly patterns for the entire period and daily pattern for the highlighted week.

linear granularities like hour, day, week and month from the commonly used Gregorian calendar. Considering 4 linear granularities hour, day, week, month in the hierarchy table, the number of cyclic granularities is $N_C = (4 * 3/2) = 6$. We obtain cyclic granularities namely "hour_day", "hour_week", "hour_month", "day_ week", "day_month" and "week_month", read as "hour of the day", etc. Further, we add cyclic granularity day-type("wknd wday") to capture weekend and weekday behavior. Thus, 7 cyclic granularities are considered to be of interest. The set consisting of pairs of cyclic granularities (C_{N_C}) will have $7_{P_2} = 42$ elements which could be analyzed for detecting possible periodicities. The set of possible harmonies H_{N_C} from C_{N_C} are chosen by removing clashes using procedures described in (Gupta et al. 2020). Table 4 shows 14 harmony pairs that belong to H_{N_C}.

Choosing and Ranking harmonies for all households

vi, t has a asymmetrical distribution as could be seen in 15 and the Normal score transform has been applied to make it more symmetric. Let $v*_{i,t}$ denote the normal-quantile transformed electricity demand for i^{th} household for time period t. Suppose $(A, B) \in H_{N_C}$ be a harmony pair where $A = \{a_j : j = 1, 2, \dots, J\}$ and $B = \{b_k : k = 1, 2, \dots, K\}$ with A placed across x-axis and B across facets. Suppose $q^{i,p}_{A,j}$ denote the quantiles with probability p for the of the i^{th} household for j^{th} category of the cyclic granularity A. Similarly, $q^{i,p}_{B,k}$ denotes the same for the k^{th} category of the cyclic granularity B. Sample quantiles were computed at $p = 0.01, 0.02, \dots, 0.99$. Jensen-Shannon distances are computed between $q^{i,p}_{A,j}$ and $q^{i,p}_{B,k}$ for each $j \in J, k \in K$ to obtain within-facet and between-facet distances. A tuning parameter of $\lambda = 0.67$ has been considered to upweigh the within-facet distances and down-weigh the between facet distances and the maximum of them are obtained to compute *wpd*. It is further normalized using the approach described in Section **??**. This entire process is repeated for all harmony pairs $\in H_{N_C}$ and for each households $i \in i = \{1, 2, \dots, 8\}$. The harmony pairs are then arranged in descending order and the important ones with significance level 1%, 5% and 10% are highlighted with ***, ** and * respectively. Table 4 shows the rank of the harmonies for different households 16 shows the heatmap for the eight households with the value of wpd_{norm} filled as colors.

Validating rank of household id:1

From table 4, it could be seen that for household id:1, (hod, wdwnd) has been ranked higher than (wdwnd, hod), both of these being significant. Further, we see that (wom, wdwnd) has been ranked 5^{th} and tagged as an insignificant pair. Figure 13 is used to show if this selection and ranking of harmony pairs makes sense for this household. Panel a) of Figure 13 shows the distribution of energy demand with weekday/weekend as the x-axis and hour-of-day as the facets and helps to compare the weekend/weekday patterns for different hours of the day. It could be observed that the difference between weekend and weekday is the highest from 15 to 19 hours of the day. Panel b) shows the distribution of energy demand with the variables swapped and helps to compare the daily patterns within weekday and weekend. It could be observed that the daily pattern is similar for weekdays and weekends with a morning and evening peak. However,the difference between morning and evening peaks are higher for weekends. Since wpd_{norm} is designed to put more weightage on within-facet differences, it makes sense that the pair (hod, wdwnd) has been ranked higher than (wdwnd, hod). Panel c) shows the distribution of energy demand with weekday/weekend as the x-axis and week-of-month as the facets. Although the differences might seem significant at first, with closer inspection it could be seen that the range of the demand is low in this case and hence the differences are not large enough to cross the threshold for significance.

From Figure 14, it could further be observed that id5 has only one significant harmony (hod, dow). Apparently, $(hod, wnd/wday)$ which is an important harmony for most households is not important for this one.

Comparing similar households and spotting anomalous ones

Validating patterns from linear display

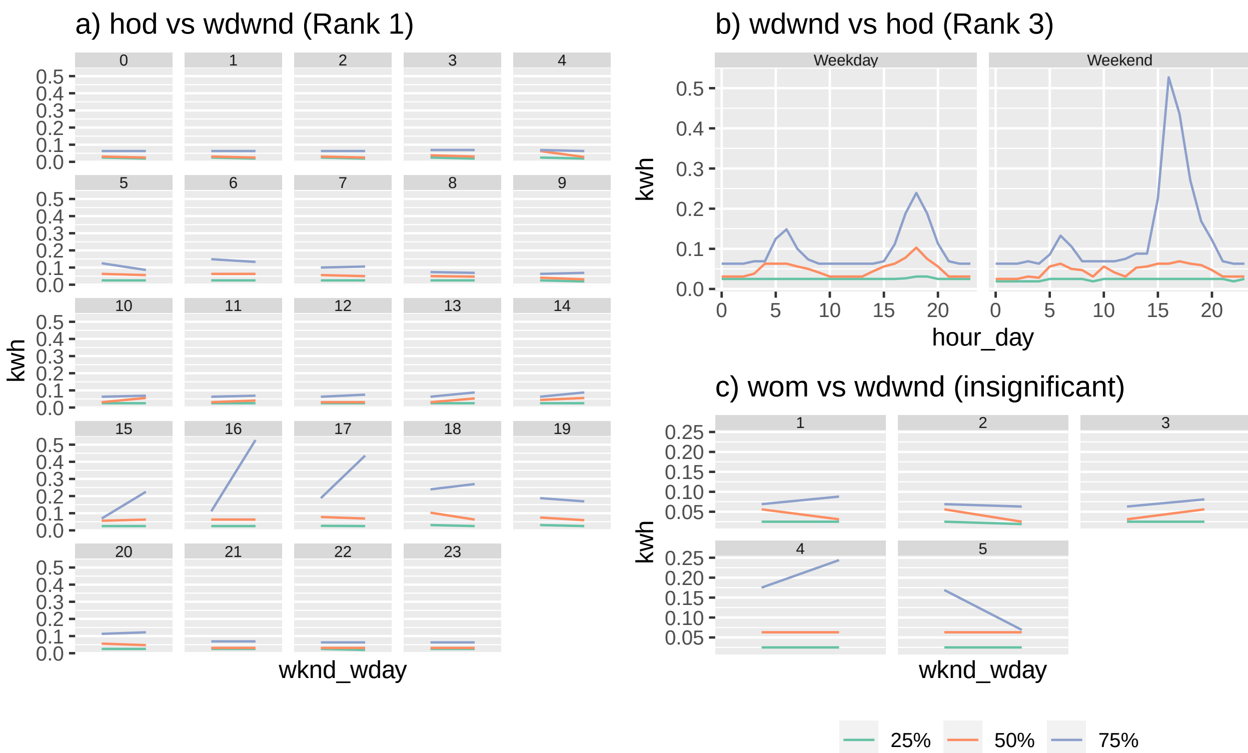


Figure 13: Distribution of energy demand shown for household id 1 across hod in x-axis and wd-wnd in facets in a) and just the reverse in b). In c), distribution of energy demand for household id:4 shown across hod and wd-wnd. It can be seen that the differences in distributions are more apparent when viewed in a) as compared to b). It seems like there is more difference in the distributions of hod for b) compared to c). This also confers with the value of the normalised measure shown in Figure 11.

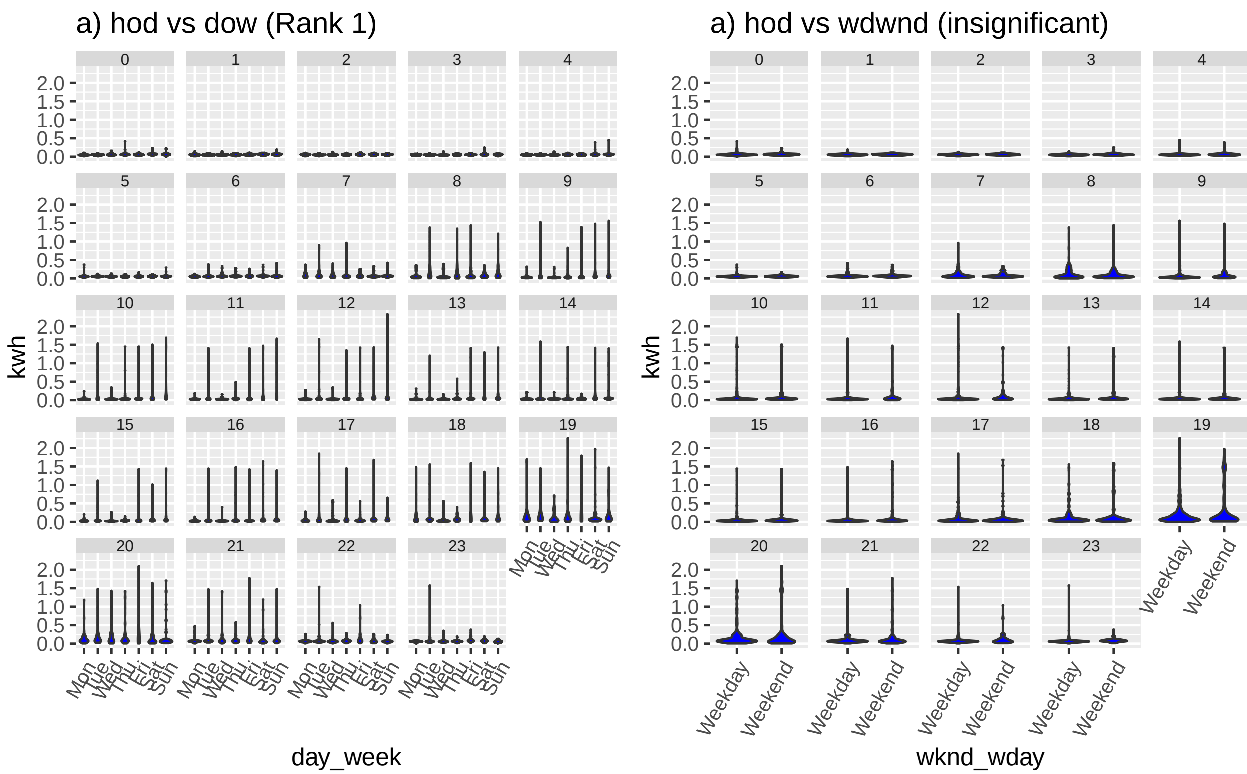


Figure 14: something

Table 4: Ranking of the harmonies for the four households are shown. The first four harmonies are unanimously important for all households, after which there is differences in the rankings for most households.

facet variable	x variable	facet_levels	x levels	id 1	id 2	id 3	id 4	id 5	id 6	id 7	id 8
hod	wdwnd	24	2	1 ***	2 *	1 **	2 **	3	1 **	3	3 *
dom	hod	31	24	2 ***	4	3 **	3 **	4	3 *	4	6
wdwnd	hod	2	24	3 **	10	7	7	6	8	8	10
hod	wom	24	5	4	9	6	5	5	5	5	5
wom	wdwnd	5	2	5	14	14	10	12	9	12	13
hod	dow	24	7	6	1 ***	2 **	1 ***	1 *	2 **	2 **	1 **
wdwnd	wom	2	5	7	12	13	8	7	7	10	12
dow	hod	7	24	8	3	4 **	4 **	2	4 *	1 ***	2 **
hod	dom	24	31	9	7	10	13	10	10	9	4
wom	dow	5	7	10	6	8	9	8	6	7	9
dow	wom	7	5	11	5	9	11	11	12	6	7
wom	hod	5	24	12	8	5	6	9	11	11	8
dom	wdwnd	31	2	13	13	11	12	14	14	14	14
wdwnd	dom	2	31	14	11	12	14	13	13	13	11

6 Discussion

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

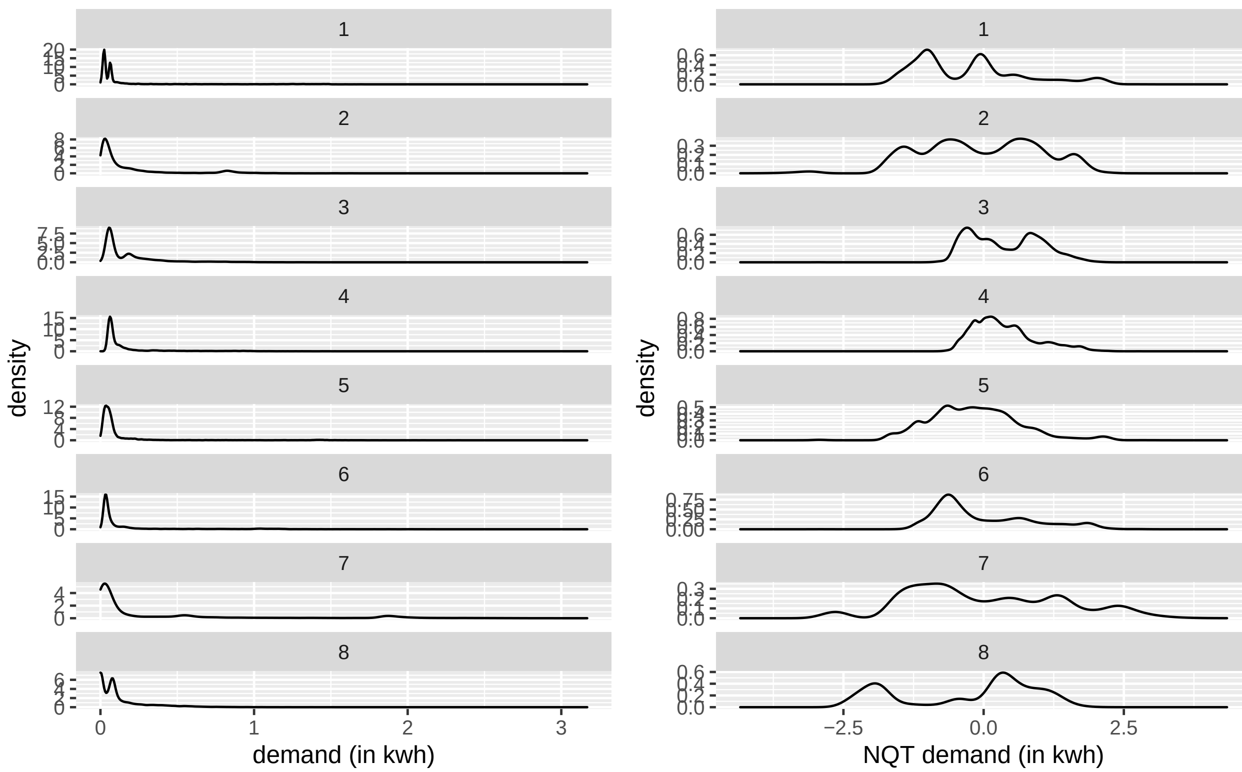


Figure 15: The raw density of the half-hourly demand for the eight households in Panel a. Panel b shows the normal-score-transform half-hourly demand for the same households which has resulted in more symmetric distribution of half-hourly demand.

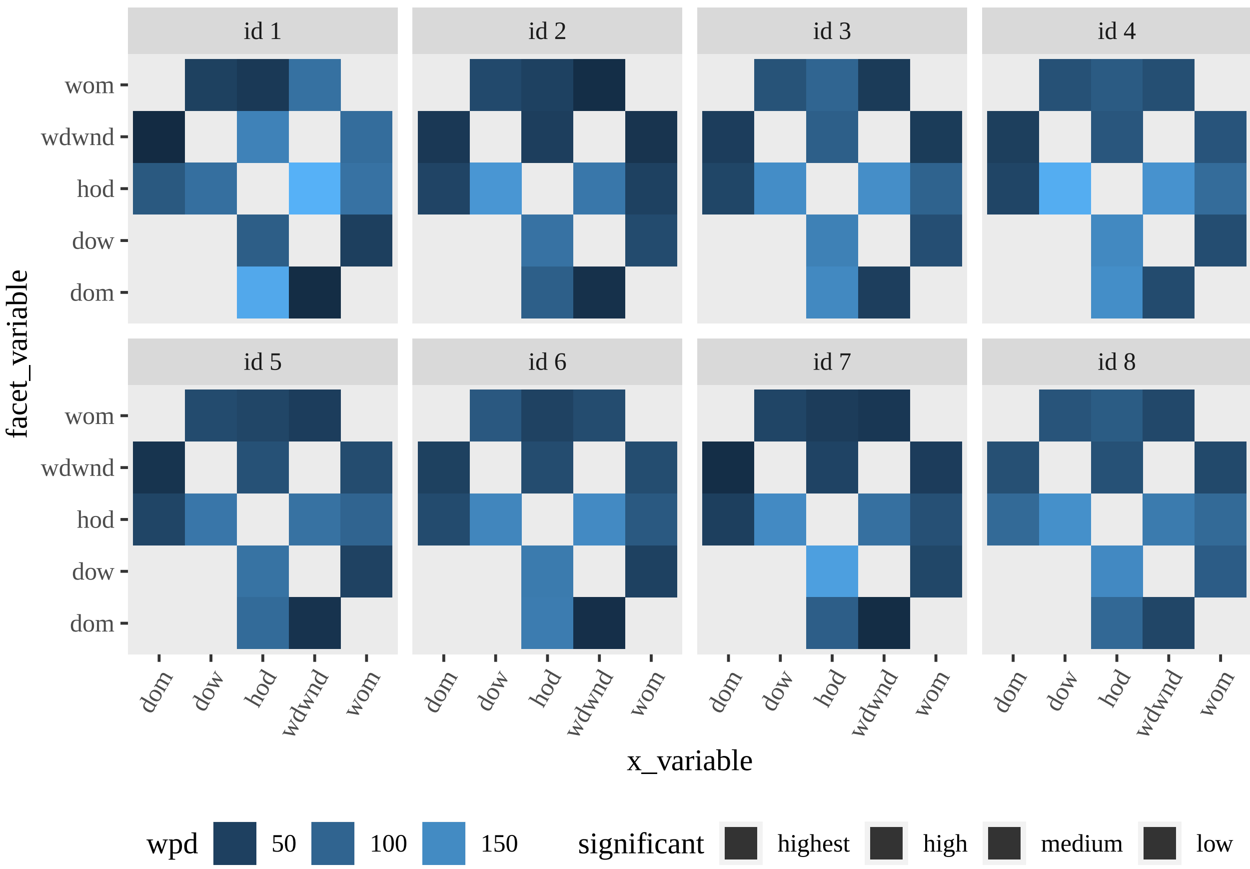


Figure 16: Harmony pairs are shown for all household ids. The darker the colour, the higher the importance of the harmony. Visualizing the pairs in this way helps us to see the important cyclic granularities along the x-axis and facet along with the information that which ones amongst them should be analyzed together.

Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." *J. Am. Stat. Assoc.* 108 (503): 942–56.

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

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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. 2020a. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics*. https://doi.org/10.1080/10618600.2019.1695624.

———. 2020b. "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.