Detecting distributional differences between temporal granularities for exploratory time series analysis

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

May 26, 2021

Abstract

Periodic patterns or associations in large univariate time series data could be discerned by analysing the behaviour across cyclic temporal granularities, which are temporal deconstructions accounting for repetitive behaviour. The temporal granularities form ordered categorical variables and display of distributions of the univariate response variable across two or more combinations of these categorical variable can help explore periodicities, patterns and anomalies. A pair of granularities that can be meaningfully examined together are called "harmonies" and the ones which cannot be are called "clashes". Even after excluding clashes, the list of harmonies that could potentially be displayed is huge and hence overwhelming for human consumption. This work provides a methodology to screen the most informative graphics from the plethora of choices by introducing a distance measure that could be compared across harmonies with varied levels and data sets. Moreover, this distance measure could also be used to rank the s = ted harmonies basis how well they capture the variation in the measured variable. All the methods are implemented in the open source R package `hakear`.

Keywords: data visualization, periodicities, cyclic granularities, permutation tests, Jensen-Shannon distances, smart meter data, R

*Email: `Sayani.Gupta@monash.edu`

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 time series data can become overwhelming with so many possible cyclic temporal granularities (Gupta et al. 2020), which are temporal deconstructions that represent cyclic repetitions in time, e.g. hour-of-day, day-of-month, or any public holidays. These granularities form ordered (for eg. day-of-week where Tue is always followed by Wed, which again is followed by Thu and so on) or unordered categorical variables. Therefore, exploring univariate time series data amounts to exploring the distribution of the measured variable across different categories of the cyclic granularities. Take the example of the electricity smart meter data used in Wang et al. (2020*a*) for four households in Melbourne, Australia. Figure 1 shows the distribution of energy usage of one household (id 2) across cyclic granularity a) hour-of-day and b) month-of-year. Potentially many such displays could be drawn across day-of-week, day-of-month, weekday/weekend, or any other chosen cyclic granularities of interest. However, all of them would not be interesting to discern important patterns in the energy usage. Only those displays, which have "significant" distributional differences between categories of the cyclic granularity would be informative and consequently, the corresponding cyclic granularity can be tagged as a "significant" one.

Exploring the distribution of the measured variable across two cyclic granularities tend to provide more detailed information on its structure. Figure 2(a) shows that energy consumption is higher during the morning hours (5-8) when members in the household wake up and then again in the evening hours (17-20) possibly when members get back from work with maximum variation (large inter-quartile range) in behavior in the afternoon hours (12-16). Figure 2(b) shows the distribution of energy consumption across months January to June. The median and quartile deviation of energy usage in Jan and Feb are generally on a much higher side, possibly due to the usage of air conditioners (Jan, Feb are peak summer in Australia), however for other months (Mar-Jun, autumn and winter), the smaller median and quartile deviation indicate a more regular behavior. It also implies this household do not use as much heater as compared to air conditioner. A lot of households in Victoria use gas heating and hence the usage of heaters might not be reflected here. Figure 2a slice it down further by showing the usage distribution across hour-of-

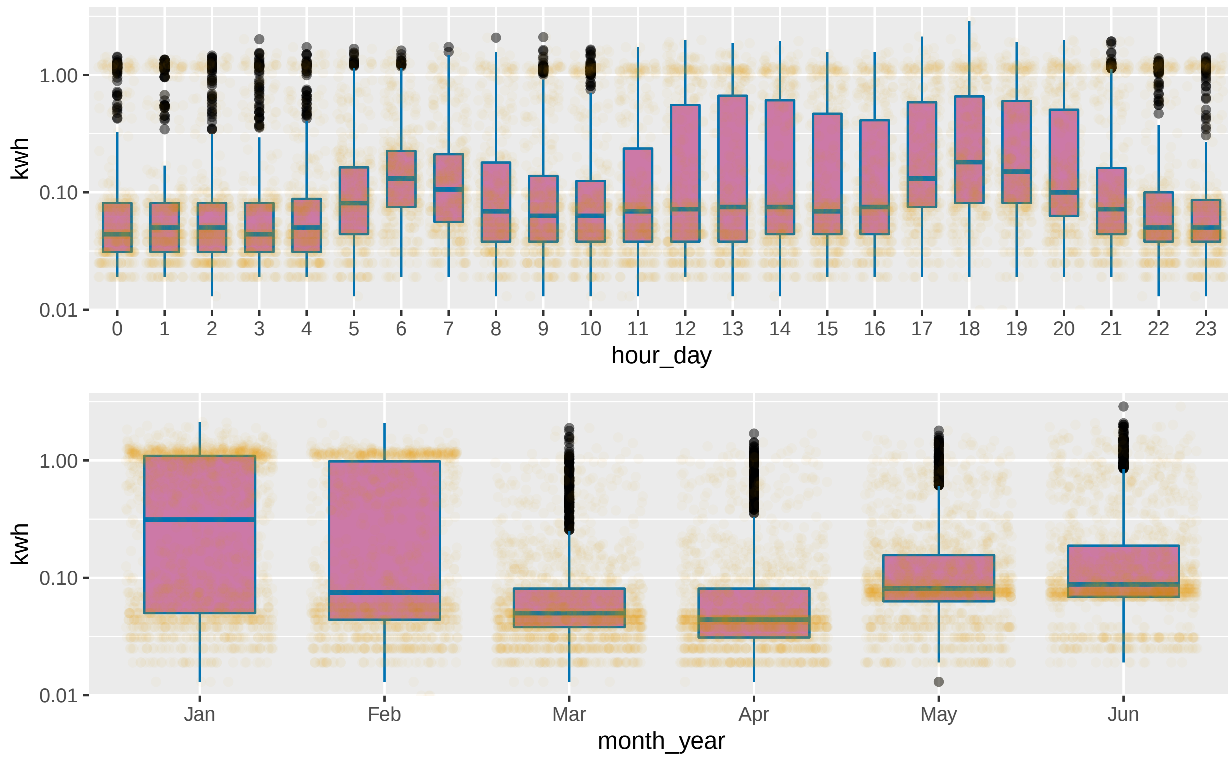

Figure 1: *something*

day conditional on month-of-year. It shows the hourly usage over a day do not remain across months. Unlike other months, the 75th and 90th percentile for all hours of the day in January are high, pretty close and are not characterized by a morning and evening peak. This implies usage of air conditioners is pretty common for this household. The household in Figure 2a has 90th percentile consumption 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 conditioning much less regularly than the first one. The differences seem to be more prominent across month-of-year (facets) than hour-of-day (x-axis) for this households, whereas they are prominent for both cyclic granularities for the first household. It could be immensely useful to make the transition from all potential displays to the ones that are informative across atleast one cyclic granularity.

The dimension of this problem, however, increases when considering two cyclic granularities. Let N_C be the total number of cyclic granularities of interest. That essentially implies there are ${}^{N_C}P_2$ possible pairwise plots exhaustively, with one of the two cyclic granularities acting as the conditioning variable. This is large and overwhelming for human consumption. This problem is

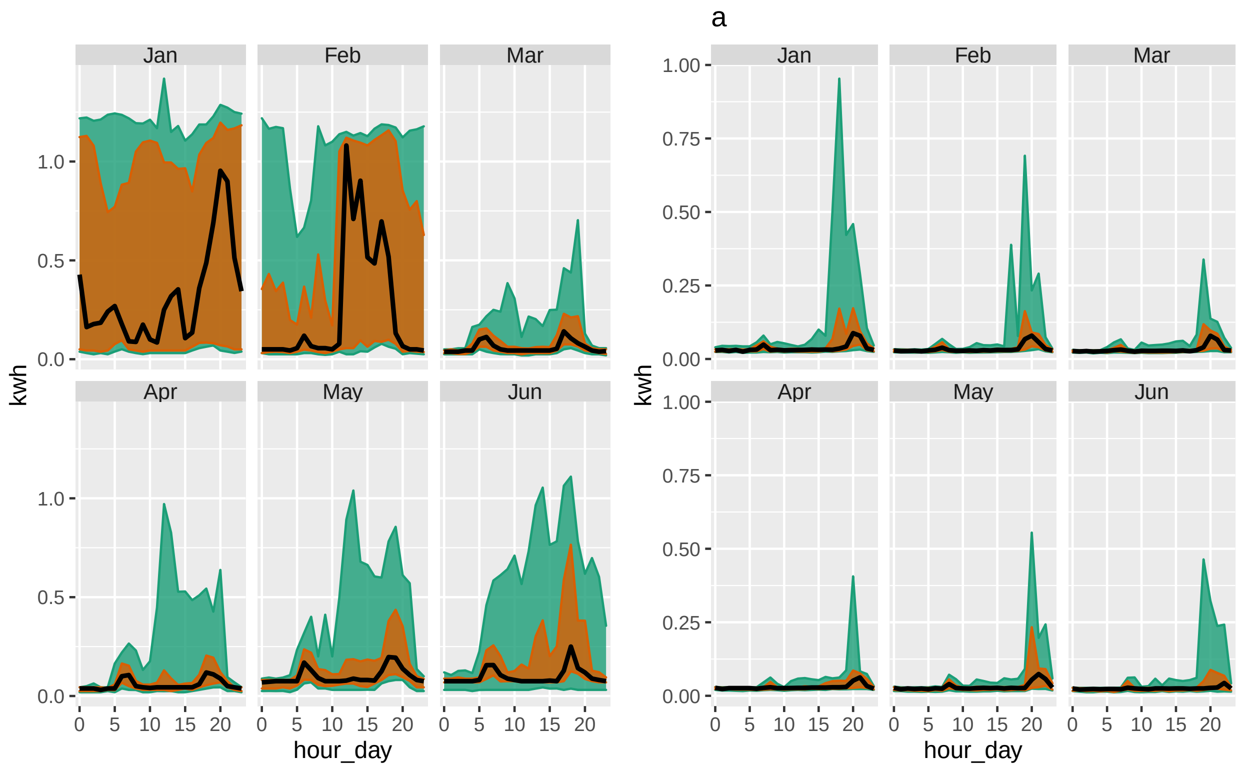

Figure 2: *Distribution of energy consumption displayed through area quantile plots across two cyclic granularities month-of-year and hour-of-day for id2 (left) and id4 (right). The black line is the median, whereas the orange band covers the 25th to 75th percentile and the green band covers the 10th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

similar to Scagnostics (Scatterplot Diagnostics) by Tukey & Tukey (1988), which is used to identify 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 & Wilkinson (2014) and Wilkinson et al. (2005) provides potential solutions to this, where few characterizations can be used to locate anomalies in density, shape, trend, and other features in the 2D point scatters. This work is a natural extension of our work that 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 et al. (2013) present methods for statistical significance testing of visual findings using human cognition as the statistical tests. In this paper, a new distance measure is introduced to enable visual findings that detect significant distributional differences between harmonies.

Our contributions in this paper are:

- introduce a distance measure for detecting distributional difference in temporal granularities, which enables identification of patterns in the time series data;
- devise a framework for choosing a threshold, which results in detection of only significantly interesting patterns;
- show that the proposed distance metric could be used to rank the interesting patterns based on how well they capture the variation in the measured variable since they have been normalized for relevant parameters.

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 2.3.7 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 3 discusses how to choose a threshold to select only significant harmonies. Section 4 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 distributional differences in harmonies

We propose a measure called Weighted Pairwise Distances (wpd) to assess structure in the measured variable across two cyclic granularities.

2.1 Principle

The principle behind the construction of *wpd* is explained through a simple example explained in Figure 3. Each of these figures describes a panel with 2 x-axis categories and 3 facet levels, but with different designs. Figure 3a has all categories drawn from N(0, 1) 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, but across facets the distributions are 3 standard deviations apart. Figure 3c exhibits an exact opposite situation where distribution between the x-axis categories are 3 standard deviations apart, but they do not change across facets. 3d takes a step further by varying the distribution across both facet and x-axis categories by 3 standard deviations. If the panels are to be ranked in order of capturing maximum variation in the measured variable 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. With this principle in mind, display (b) is considered less informative as compared to display (c) in emphasizing the distributional differences. The proposed measure *wpd* is constructed in a way so that it could be used to rank panels of different designs as well as test if a design is interesting. This measure is an estimate of the maximum variation in the measured variable explained by the panel. A higher value of *wpd* would indicate that the panel is interesting to look at, whereas a lower value would indicate otherwise.

2.2 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. This data structure with J x-axis levels and K facet levels is referred to as a (K, J) panel. For example, a $(2, 3)$ panel will have cyclic granularities with 2 x-axis levels and 3 facet levels. 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

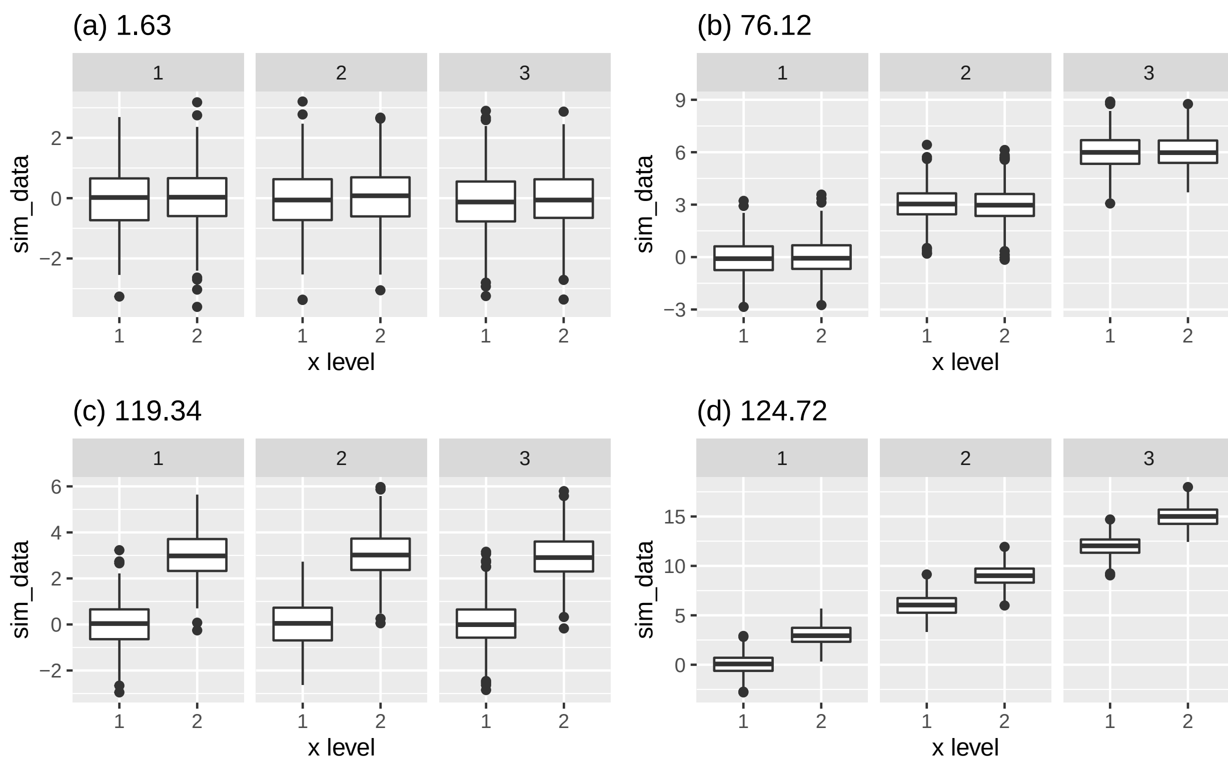

Figure 3: *An example illustrating the principle of the proposed distance measure, displaying the distribution of a normally distributed variable in four panels each with* 2 *x-axis categories and* 3 *facet levels, but with different designs. Display (a) is not interesting as the distribution of the 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 capturing structure in the variable as the distribution of the variable changes across both facet and x-axis variable. The value of our proposed distance measure is presented for each panel, the relative differences between which will be explained later in Section 3.2.*

those designs for which difference is observed across both A and B.

Table 1: *Nomenclature table*

variable	description
N_C	number of cyclic granularities
H_{N_C}	set of harmonies
nx	number of x-axis categories
nfacet	number of facet categories
$lambda$	tuning parameter
$omega$	increment (mean or sd)
wpd	raw weighted pairwise distance
$nperm$	number of permutations for threshold/normalization
$nsim$	number of simulations
wpd_{norm}	normalized weighted pairwise distance
$wpd_{threshold}$	threshold for significance
D_{null}	null design with no distributional difference across categories
D_{var_f}	design with distributional difference only across facets categories
D_{var_x}	design with distributional difference only across x-axis categories
$D_{var_{all}}$	design with distributional difference across both facet and x-axis

2.3 Computation

The computation of the distance measure *wpd* for a panel involves characterizing distributions, computing distances between distributions, choosing a tuning parameter to specify the weightage for different group of distances. Furthermore, the data needs to be appropriately transformed to ensure that the value of *wpd* emphasizes distributional differences across categories and not across different data generating processes.

2.3.1 Data transformation

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 (NQT) (Krzysztofowicz (1997)), so that the quantiles of the transformed data follows a standard normal distribution. This sort of transformation is common in the fields of geo-statistics to make most asymmetrical distributed real world measured variables more treatable and normal-like (Bogner et al. (2012)).The empirical NQT involves the following steps:

1. The sample of measured variable v is sorted from the smallest to the largest observation $v_{(1)},\dots,v_{(i)},..,v_{(n)}$.

2. The cumulative probabilities $p_{(1)},\dots,p_{(i)},..,p_{(n)}$ are estimated using a plotting position like $i/(n+1)$ such that $p_{(i)} = P(v \leq v_{(i)})$.

3. Each observation $v_{(i)}$ of v is transformed into observation $v^*(i) = Q^{-1}(p(i))$ of the standard normal variate v^*, with Q denoting the standard normal distribution and Q^{-1} its inverse.

2.3.2 Characterising distributions

Multiple observations of v correspond to the subset $v_{jk} = \{s : A(s) = j, B(s) = k\}$. The number of observations might vary widely across subsets due to the structure of the calendar, missing observations or uneven locations of events in the time domain. Quantiles of v_{jk}'s are chosen as a way to characterize distributions for the category (a_j, b_k) in the paper $\forall j \in \{1,2,\dots,J\}, k \in \{1,2,\dots,K\}\}$. 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. Sample quantiles (Hyndman & Fan (1996)) are used for estimating population quantiles in a non-parametric setup, which is desirable because of less rigid assumptions made about the nature of the underlying distribution of the data. The default quantile chosen in this paper is percentiles computed for $p = 0.01, 0.02, \dots, 0.99$, 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.

2.3.3 Distance between distributions

Most common way to measure divergence between distributions is the Kullback-Leibler (KL) divergence (Kullback & Leibler 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, the pairwise distances between the distributions of the measured variable are obtained 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 (Menéndez et al. (1997)), 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.

2.3.4 Within-facet and between-facet distances

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).

2.3.5 Tuning parameter

A tuning parameter specifying the weightage given to the within-facet or between-facet categories can help to choose between designs like 3(b) and (c). Following the principle of Gestalt theory, the relative importance of within-facet and between-facet distances are taken to be 2 : 1 and hence $\lambda = \frac{2}{3} = 0.67$. No human experiment is conducted to justify this ratio, however, typically a tuning parameter $\lambda > 0.5$ would tend to upweigh the within-facet distances and that with < 0.5 would upweigh the between-facet distances (refer to the Supplementary section of the paper for more details).

2.3.6 Raw distance measure

The raw distance measure, denoted by wpd_{raw}, is computed after combining all the weighted distance measures appropriately. First, NQT is performed on the measured variable v_t to obtain v^*_t (*data transformation*). Then, for a fixed harmony pair (A,B), percentiles of v^*_{jk} are computed and stored in q_{jk} (*distribution characterization*). This is repeated for all pairs of categories of the form $(a_jb_k, a_{j'}b_{k'}) : \{a_j : j = 1,2,\dots,J\}, B = \{b_k : k = 1,2,\dots,K\}$. The pairwise distances between pairs $(a_jb_k, a_{j'}b_{k'})$ denoted by $d_{(jk,j'k')} = JSD(q_{jk}, q_{j'k'})$ is computed (*distance between distributions*). The pairwise distances (*Within-facet and between-facet*) are 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:

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

The wpd_{raw} is then computed as

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

The statistic "maximum" is chosen to combine the weighted pairwise distances since the distance measure is aimed at capturing the maximum variation of the measured variable within a panel. The statistic "maximum" is, however, affected by the number of comparisons (resulting pairwise distances). For example, for a $(2,3)$ panel, there are 6 possible subsets of obervations corresponding to the combinations $(a_1,b_1),(a_1,b_2),(a_1,b_3),(a_2,b_1),(a_2,b_2),(a_2,b_3)$, whereas for a

$(2,2)$ panel, there are only 4 possible subsets $(a_1,b_1),(a_1,b_2),(a_2,b_1),(a_2,b_2)$. Consequently, the measure would have higher values for the panel $(2,3)$ as compared to $(2,2)$, since maximum is taken over higher number of pairwise distances.

2.3.7 Adjusting for the number of comparisons

Ideally it is desired that wpd takes a higher value 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 panels with different levels, the impact of different number of comparisons need to be eliminated through some adjustment. Henceforth, we call the adjusted measure as wpd_{norm}. Two approaches for adjusting for the number of comparisons are discussed, both of which are substantiated using simulations. The first one, denoted as wpd_{perm}, involves a permutation method to make the distribution of the adjusted distance measure similar for different comparisons. The second one ($wpd_{glm-scaled}$) fits a model to represent the relationship between wpd_{raw} and number of comparisons and defines the adjusted measure as the residual from the model.

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. That is, data is being generated for each of the panels $(2,2),(2,3),(2,5)\ldots,(50,31),(50,50)$. For each of the 64 panels, $ntimes = 500$ observations are drawn for each combination of the categories. That is, if we consider a $(2,2)$ panel, 500 observations are generated for each of the possible subsets, 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.

- *Permutation approach:* This method is somewhat similar in spirit to bootstrap or permutation 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 distribution of the different groups (panels in our case) same under D_{null}. The values of

wpd_{raw} is computed on many permuted ($nperm$) data sets for the l^{th} panel and stored in $wpd_{perm-data}$. $nperm = 100$ is considered and the adjusted distance measure wpd_{perm} is computed as follows:

$$wpd_{perm} = \frac{(wpd_{raw} - \bar{wpd}_{perm-data})}{sd(wpd_{perm-data})}$$

where $\bar{wpd}_{perm-data}$ and $sd(wpd_{perm-data})$ are the mean and standard deviation of $wpd_{perm-data}$ 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, another approach to adjustment, with potentially less computational time, is proposed.

- *Modeling approach:* wpd_{raw} being a Jensen-Shannon distance, follows a Chi-square distribution (Lin (1991)), which is a special case of Gamma distribution. Hence, a Gamma generalized linear model (GLM) for wpd_{raw} is fitted with number of comparisons as the explanatory variable, which is of the following form:

$$y_l = a + b * log(z_l) + e_l, \quad l = (1, 2, \ldots, L)$$

,

where $y_l = median_s(wpd_{l,s})$, $wpd_{l,s}$ denotes the value of wpd_{raw} obtained for the l^{th} panel and s^{th} simulation, z_l is the number of groups $(nx * nfacet)$ in 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: $wpd_{glm} = wpd_{raw} - 1/(\hat{a} + \hat{b} * log(nx * nfacet))$.

- *Combination approach:* The simulation results show that adjustment 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

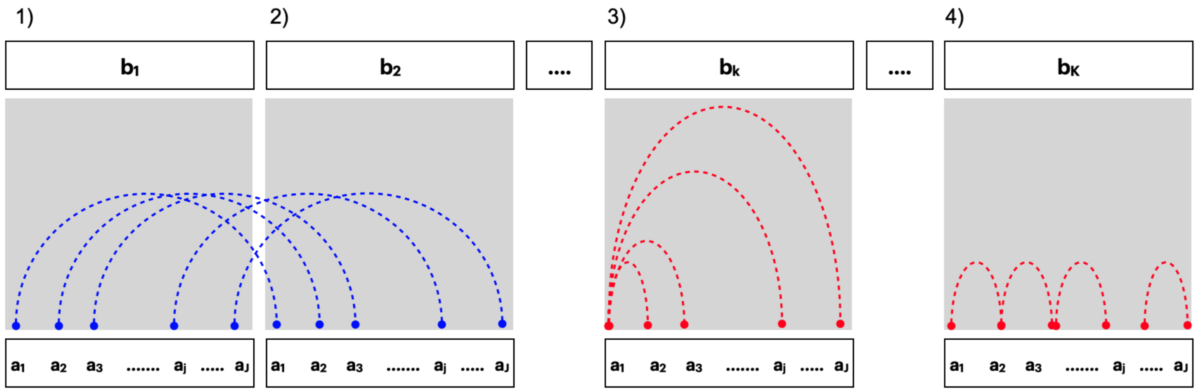

Figure 4: *Within and between-facet distances shown for two cyclic granularities A and B, where A is mapped to x-axis and B is mapped to facets. The dotted lines represent the distances between different categories. Panel 1) and 2) show the between-facet distances. Panel 3) and 4) are used to illustrate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

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 $nfacet$. The adjusted 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 adjusted measure from the combination approach, denoted by wpd_{norm} is defined as follows:

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

where $wpd_{glm-scaled}$ is the transformed wpd_{glm} with a similar scale as wpd_{perm}.

```
#> [1] 1.003985
```

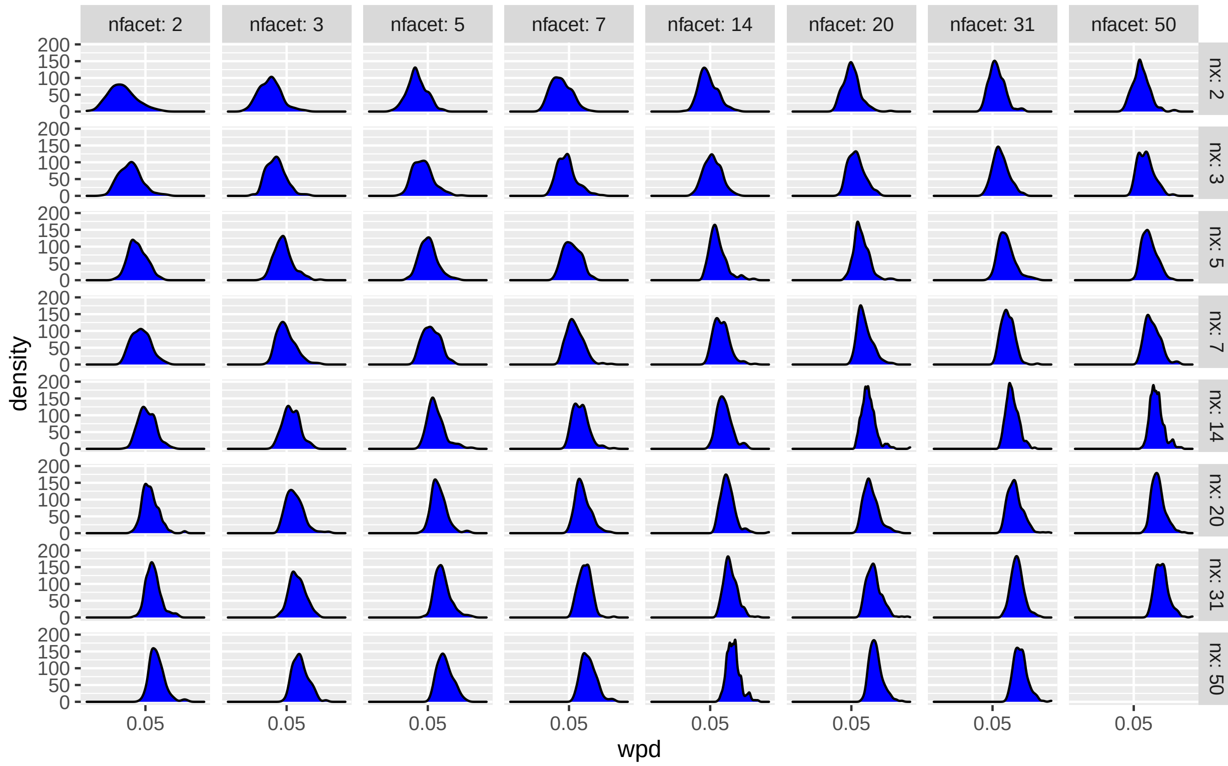

Figure 5: *Distribution of wpd is plotted across different nx and $nfacet$ categories under D_{null}. Both shape and scale of the distribution changes for different comparisons. This is not desirable since under null design, the distribution of the distance measure is not expected to capture any differences.*

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

term	estimate	std.error	statistic	p.value
(Intercept)	23.69448	0.2399014	98.76757	0
log(‘nx * nfacet‘)	-1.02357	0.0481998	-21.23596	0

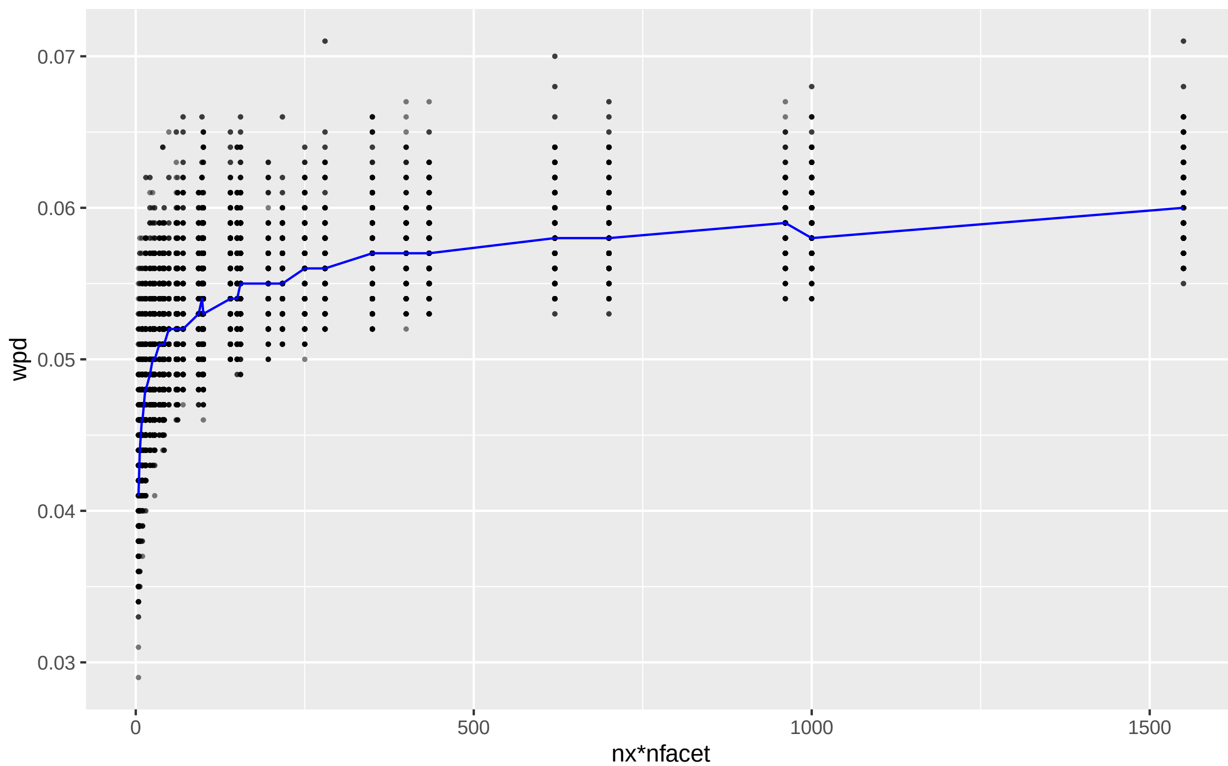

Figure 6: *wpd is plotted against* $nx * nfacet$ *(the maximum number of pairwise comparisons) and the blue line represents the median of the multiple values for each* $nx * nfacet$*. The median increases abruptly for lower values of* $nx * nfacet$ *and slowly for higher* $nx * nfacet$*. Thus, the measure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

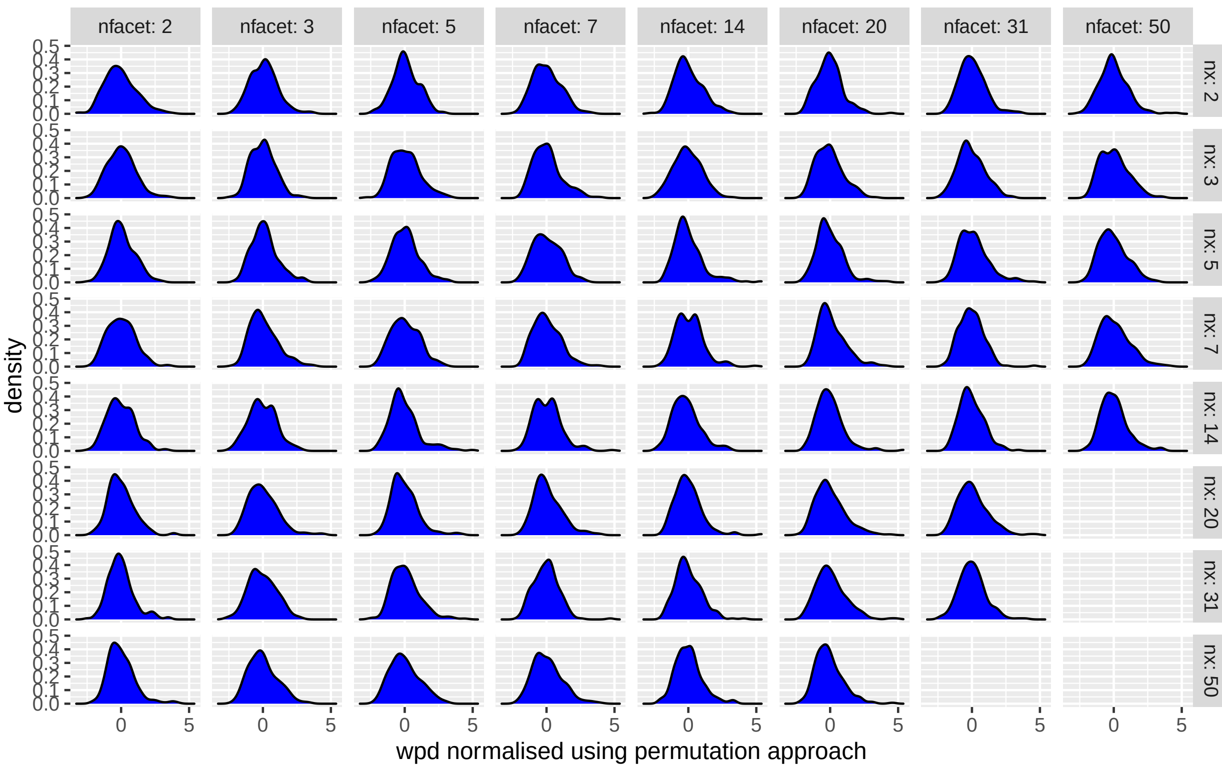

Figure 7: *Distribution of wpd_{perm} is plotted across different nx and $nfacet$ categories. Both shape and scale of the distributions are now similar for different panels under the null design.*

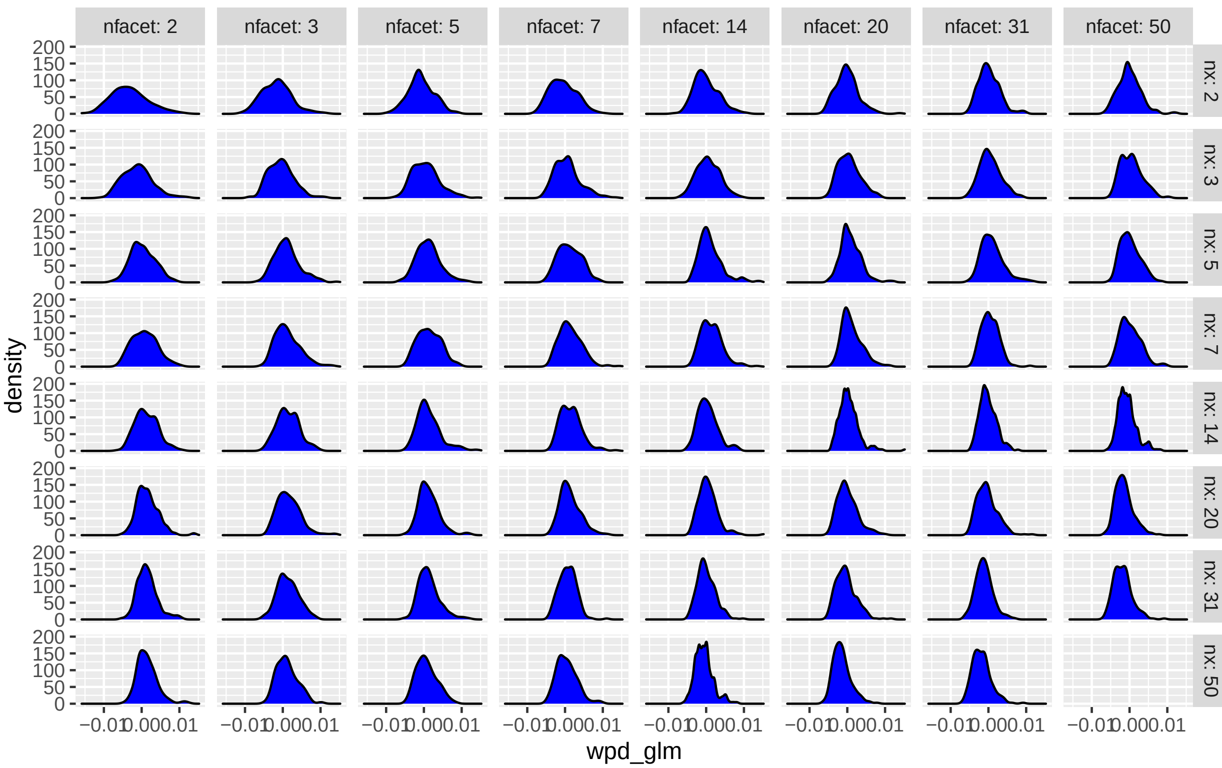

Figure 8: *The distribution of wpd_{glm} is plotted. The distributions are more similar across higher nx and nfacet and dissimilar for fewer nc and nfacets.*

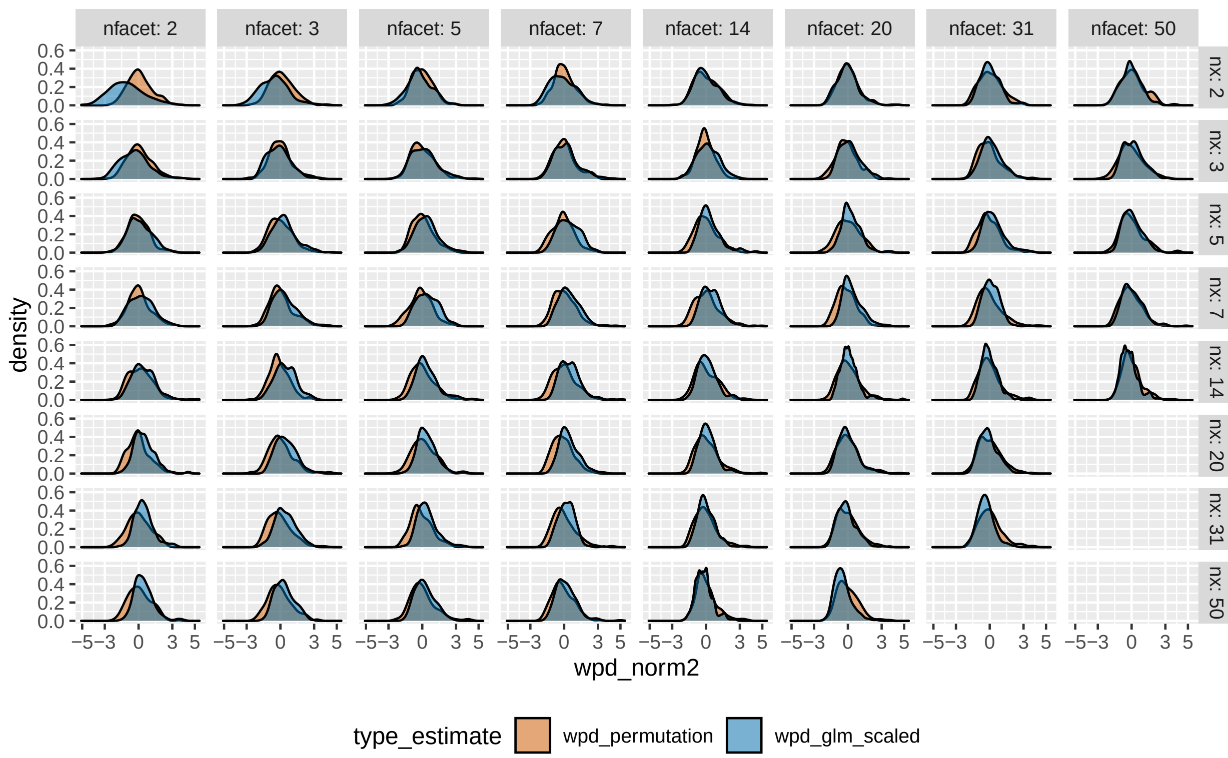

Figure 9: *wpd_perm and $wpd_{glm-scaled}$ are plotted together on the same scale. They also have the same location and hence the values from these two approaches could be compared across panels. $wpd_{glm-scaled}$ would be used to normalise wpd_{raw} for higher nx and $nfacet$ and $wpd_{glm-scaled}$ would be used for smaller levels to alleviate the problem of computational time.*

3 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 by employing 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.

3.1 Choosing a threshold

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. But we will use this procedure to obtain a threshold such that harmony pairs with a wpd_{norm} value higher than this threshold will only be considered significant. The process of choosing a threshold is described as follows:

- **Input:** All harmonies of the form $\{(A,B), A = \{a_j : j = 1,2,\dots,J\}, B = \{b_k : k = 1,2,\dots,K\}\}$ 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; $\{v_t : t = 0,1,2,\dots,T-1\}$, the wpd_{norm} is computed and is represented by wpd_{obs}.
3. From the original sequence a random permutation is obtained: $\{v_t^* : t = 0,1,2,\dots,T-1\}$.
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 $wpd_{sample} = \{wpd_{perm_1}, wpd_{perm_2}, \dots, wpd_{perm_M}\}$.

7. Repeat Steps (1-6) for all harmony pairs $(A,B) \in H_{N_C}$ and store it in wpd^{all}_{sample}.

8. 99^{th} percentiles of wpd^{all}_{sample} is computed and stored in $wpd_{threshold99}$

9. If $wpd_{obs_{A,B}} > wpd_{threshold99}$, harmony pair (A,B) is selected with a 99% level of significance and otherwise rejected.

Similarly, a harmony pair (A,B) is selected with a 95% and 90% level of significance if $wpd_{obs_{A,B}} > wpd_{threshold95}$ and $wpd_{obs_{A,B}} > wpd_{threshold90}$, where $wpd_{threshold95}$ and $wpd_{threshold90}$ denote the 95^{th} and 90^{th} percentile of wpd^{all}_{sample} respectively.

3.2 Simulation design

Observations are generated from a N(0,1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = \{3,7,14\}$ and $nfacet = \{2,9,10\}$. The panel $(3,2),(7,9),(14,10)$ are considered to have design D_{null}. The panels $(7,2),(14,9)$ have design of the form D_{var_f}. $(14,2),(3,10)$ have design of the form D_{var_x} and the rest are under $D_{var_{null}}$. We generate only one data set for which all these designs were simulated and consider this as the original data set. We generate 200 repetitions of this experiment with different seeds and compute the proportion of times a panel is rejected when it is under D_{null}. We also compute the proportion of times a panel is rejected when it actually belongs to a non-null design. The first proportion is desired to be as small as possible and a higher value of the later is expected. Also, these would constitute to be the estimated size and power of the test.

3.2.1 Results

The results for this section is WIP and to be included in details in the Supplementary paper. Also, Figure **??** presents the results of wpd_{norm} from the illustrative designs in Section 2. As expected, the value of wpd_{norm} under null design is the least (a), followed by (b, c and d). Moreover, note the relative difference in wpd_{norm} values as we move from a to d, which aligns with the idea of wpd_{norm}, since the differences between x categories are up-weighed by design.

3.3 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.

4 Application to residential smart meter dataset

The smart meter data set for eight households in Melbourne has been utilized to see the use of wpd_{norm} proposed in the paper. The data has been cleaned to be a `tsibble` (Wang et al. (2020*b*)) 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. The line display of energy usage will have too many measurements squeezed in a linear representation. No behavioral pattern is likely to be discerned from this view. When we zoom into the linear representation of this series in Figure **??** (b) for September, some patterns are visible in terms of peaks and troughs, but we do not know if they are regular or what is their period. Electricity demand, in general, has a daily and weekly periodic pattern. However, it is not apparent from this view if all of these households have those patterns and in case they have if they are significant enough. Also, it is not clear if any other periodic patterns are present in any household which might have been hidden with this view. We start the analysis by choosing few harmonies, ranking them for each for these households, compare households to get more insights into what these rankings imply and if they could be used to remove some non-interesting harmonies. Furthermore, the ranking and selection of significant harmonies could be validated by analyzing the distribution of the energy usage across significant harmonies.

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 the tsibble is 30 minutes. We consider coarser 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 3 shows 14 harmony pairs that belong to H_{N_C}.

Selecting and Ranking harmonies for all households

wpd_{norm} is computed for all harmony pairs $\in H_{N_C}$ and for each households $i \in i = \{1, 2, \ldots, 8\}$. 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 3 shows the rank of the harmonies for different households. The rankings are different for different households, which is a reflection of their varied behaviors. Most importantly, there are at most 3 harmonies that are significant for any household. This is a huge reduction in number of potential harmonies to explore closely, starting from 42.

Comparing households and validating patterns from linear display

Figure 10 helps to compare households through the heatmap (a) and validate the results of the heatmap through linear display in (b). (b) contains the raw data for a month (Sep-2019). In (a), the value of wpd_{norm} filled as colors, implying darker cells correspond to more significant a harmony pair. Also, the ones with $*$ are significant at 95% level. This plot suggests that there are no significant periodic patterns for id 5. Household id 6 and 7 differ in the sense that for id 6, the difference in patterns only during weekday/weekends, whereas for id 7, all or few other days of the week are also important. This might be due to their flexible work routines or different day-off. The periodic pattern is very similar to id 8, which has a very different total energy usage. Households id 2 and 3 are similar, in terms of linear display and heatmap, which implies that similar periodic behavior can stem from households with different demographics.

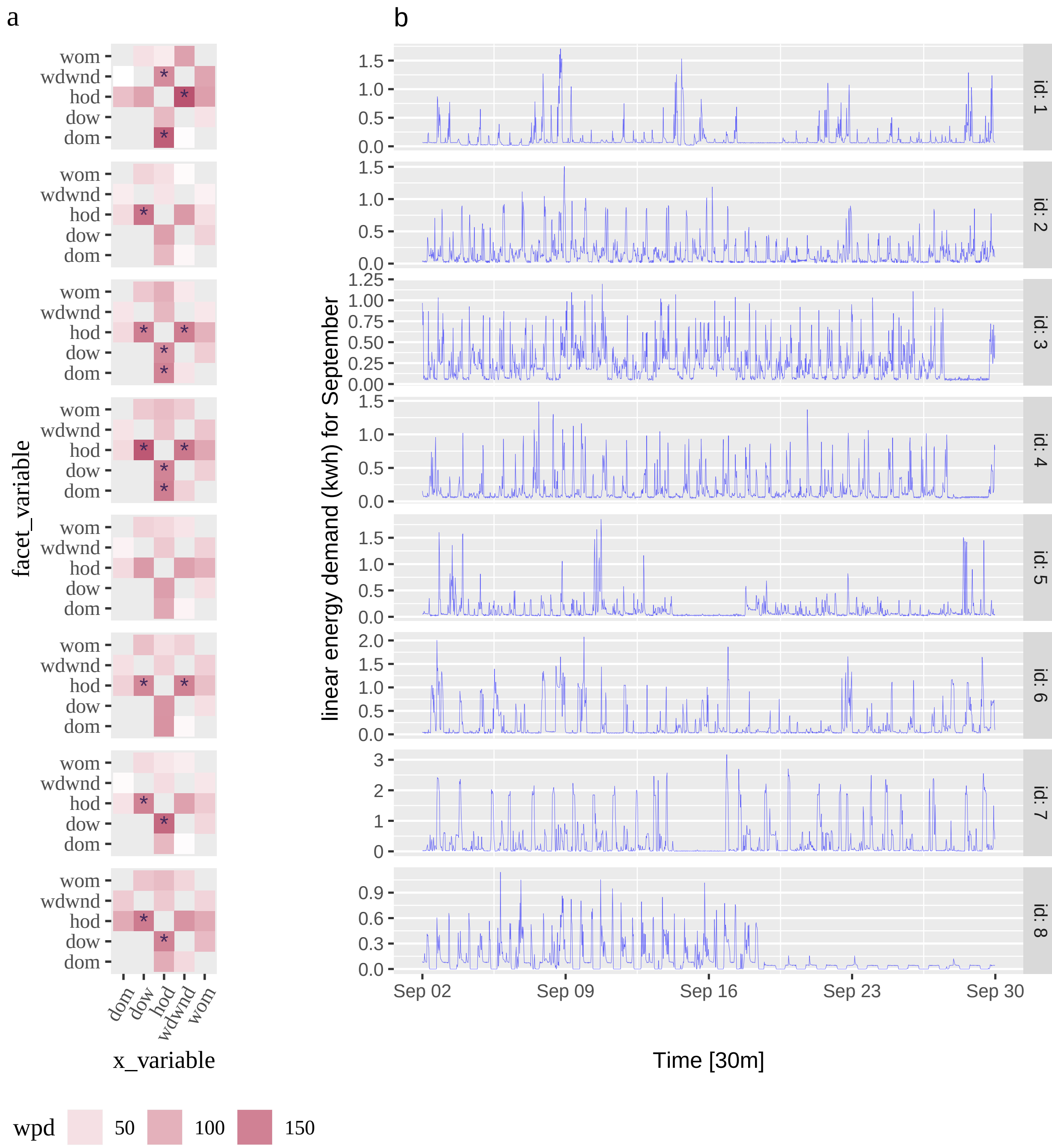

Figure 10: *Harmony pairs are shown for all household ids. The darker the colour, the higher the importance of the harmony. Also, the ones bordered red are selected with 90 percent significance level. Visualizing the pairs in this way helps us to see the important cyclic granularities along the x-axis and facet along with the information that which households should be analyzed together.*

Table 3: *Ranking of harmonies for the eight households with significance levels.*

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

Validating rank of household id:4 and 5

From table 3, it could be seen that the harmony pair (dow, hod) is significant for household id 4, but for 5 it has been tagged as an insignificant pair. The distribution of energy demand with dow as the x-axis and hod as the facets for both of these households can help justify the selection. Figure 11 show that the distribution of energy consumption changes across both dow and hod for id 4, but it is not so different for id 5, as apparent from the scale of the usage. Although the differences might seem significant at first, with closer inspection it could be seen that the scale of the demand is lower in this case and hence the differences are not large enough to cross the threshold for significance.

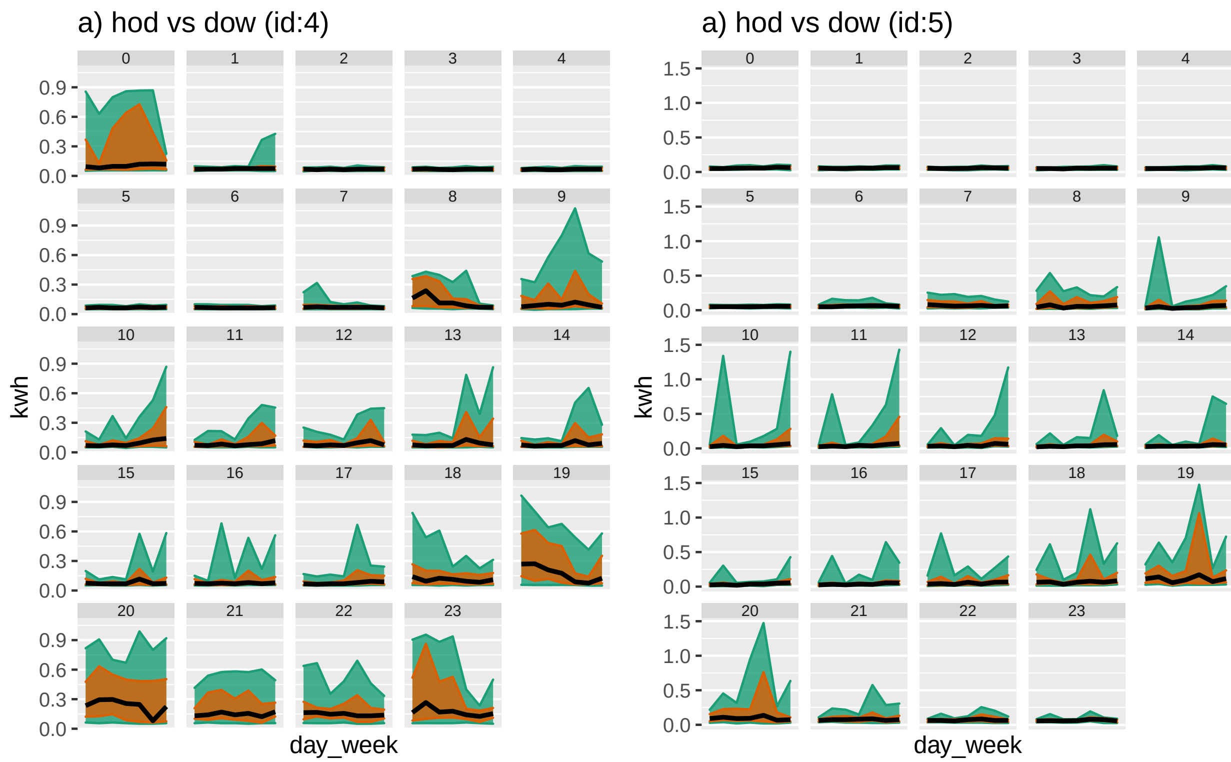

Figure 11: *Distribution of energy demand shown for household id 1 across hod in x-axis and wd-wnd in facets in a) and just the reverse in b). In c), distribution of energy demand for household id:4 shown across hod and wd-wnd. It can be seen that the differences in distributions are more apparent when 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.*

5 Discussion

Exploratory data analysis involve many iterations of finding and summarizing patterns. With temporal data available at more and more finer scales, exploring periodicity has become overwhelming with so many possible granularities to explore. A common solution would be to zoom into "interesting" segments, but there is no way to know the "interesting" segments a priori. This work refines the search of periodic patterns by identifying those for which the differences between the displayed distributions is greatest, and rating them 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 harmonies would be varying for subjects and it would be hard to track the behavior when the number of individuals rise. One way to find groups would be to actually locate clusters who have similar periodic behavior.

References

Bogner, K., Pappenberger, F. & Cloke, H. L. (2012), 'Technical note: The normal quantile transformation and its application in a flood forecasting system', *Hydrol. Earth Syst. Sci.* **16**(4), 1085–1094.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F. & Wickham, H. (2009), 'Statistical inference for exploratory data analysis and model diagnostics', *Royal Society Philosophical Transactions A* **367**(1906), 4361–4383.

Dang, T. N. & Wilkinson, L. (2014), ScagExplorer: Exploring scatterplots by their scagnostics, *in* '2014 IEEE Pacific Visualization Symposium', pp. 73–80.

Gupta, S., Hyndman, R. J., Cook, D. & Unwin, A. (2020), 'Visualizing probability distributions across bivariate cyclic temporal granularities'.

Hyndman, R. J. & Fan, Y. (1996), 'Sample quantiles in statistical packages', *Am. Stat.* **50**(4), 361–365.

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

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *Ann. Math. Stat.* **22**(1), 79–86.

Lin, J. (1991), 'Divergence measures based on the shannon entropy', *IEEE Transactions on Information Theory* **37**(1), 145–151.

Majumder, M., Hofmann, H. & Cook, D. (2013), 'Validation of visual statistical inference, applied to linear models', *J. Am. Stat. Assoc.* **108**(503), 942–956.

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

Tukey, J. W. & Tukey, P. A. (1988), 'Computer graphics and exploratory data analysis: An introduction', *The Collected Works of John W. Tukey: Graphics: 1965-1985* **5**, 419.

Wang, E., Cook, D. & Hyndman, R. J. (2020*a*), 'Calendar-based graphics for visualizing people's daily schedules', *Journal of Computational and Graphical Statistics* . to appear.

Wang, E., Cook, D. & Hyndman, R. J. (2020*b*), 'A new tidy data structure to support exploration and modeling of temporal data', *Journal of Computational and Graphical Statistics* . to appear.

Wilkinson, L., Anand, A. & Grossman, R. (2005), Graph-theoretic scagnostics, *in* 'IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.', IEEE, pp. 157–164.