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

June 14, 2021

Abstract

Patterns or associations in large univariate time series data could be explored by analyzing the behavior across cyclic temporal granularities, which are temporal deconstructions accounting for repetitive behavior, for eg, hour-of-the-day, work-day/weekend, or holidays. This way of exploring time series analysis, however, presents itself with a plethora of displays that are potentially overwhelming for human consumption. This work provides a methodology to screen and rank the displays that are most informative in discerning distributional differences. This is done by introducing a distance measure for one or a pair of cyclic granularities that could be compared across different cyclic granularities and data sets. 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 1977) 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 (Bettini et al. 1998, 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. Figure 1(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 interquartile range) in behavior in the afternoon hours (12-16). Figure 1(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 consistent behavior. It might also imply that this household does 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. 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 energy usage. Only those displays, which have "significant" distributional differences between categories of the cyclic granularity would be informative.

Exploring the distribution of the measured variable across two cyclic granularities tends to provide more detailed information on its structure. For example, Figure 2(a) slice down further by showing the usage distribution across hour-of-day conditional on month-of-year across two

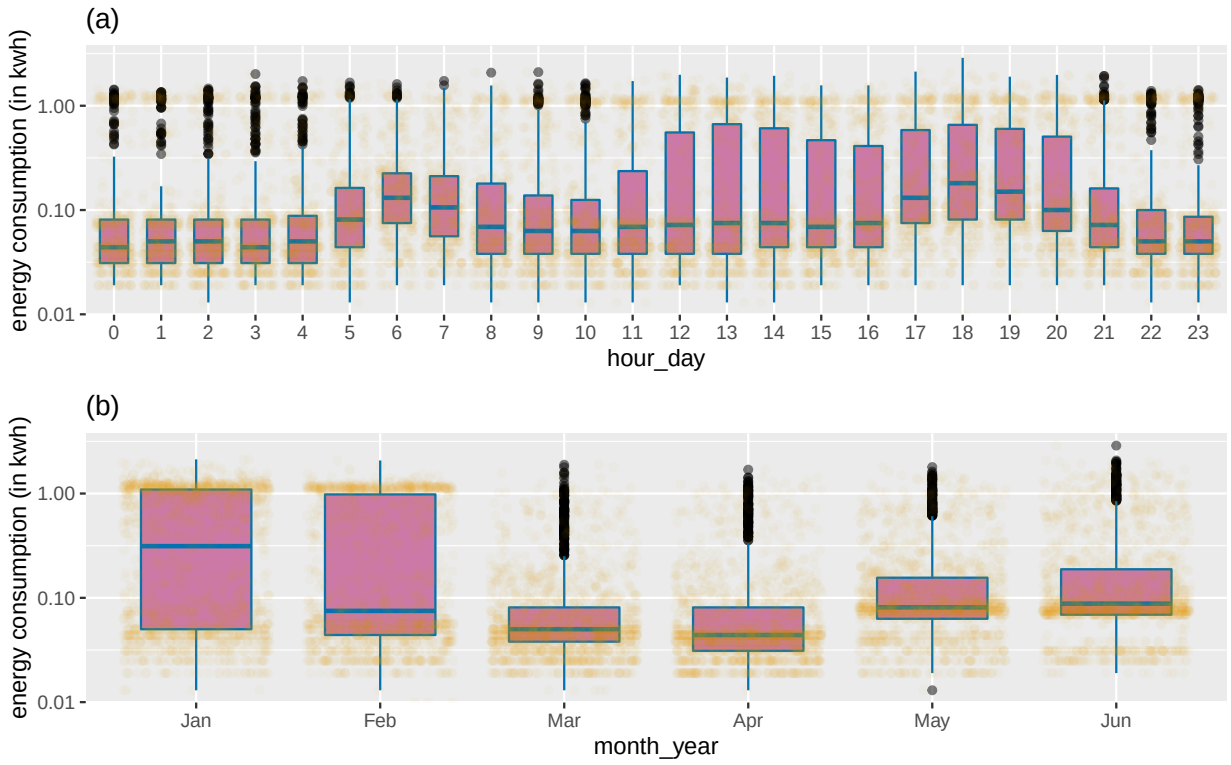

Figure 1: *Boxplots showing the distribution of one household across one cyclic granularity at a time - (a) hour-of-day and (b) and month-of-year. The daily and annual periodic behavior of energy is apparent in (a) and (b) respectively, with daily peaks occurring in morning and evening hours when members in the house are present and active, more volatility (usage of air conditioner) in summer months (Jan, Feb) due to air conditioners and more consistent behavior in winter and autumn months (Mar, Apr, May, June).*

households (id 2 and 4). It shows the hourly usage over a day does not remain the same 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. The household in Figure 2(b) 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 household, whereas they are prominent for both cyclic granularities for the first household.

Are all of these four displays in Figures 1 and 2 useful in understanding the distributional difference in energy usage? Which ones are more useful than others? If N_C is the total number of cyclic granularities of interest, the number of displays that could be potentially informative is N_C when considering displays of the form in Figure 1. The dimension of the problem, however, increases when considering more than one cyclic granularity. When considering displays of the form in Figure 2, there are ${}^{N_C}P_2$ possible pairwise plots exhaustively, with one of the two cyclic granularities acting as the conditioning variable. This is huge and overwhelming for human consumption even for moderately large N_C. It could be immensely useful to make the transition from all potential displays to the ones that are informative across atleast one cyclic granularity.

This problem is 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) provide 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. In this paper, we provide a solution to narrowing down the search from ${}^{N_C}P_2$ plots by introducing a new distance measure that can be used to detect significant distributional differences across cyclic granularities. This work is a natural extension of our previous work (Gupta et al. 2020) 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, the visual discovery of distributional differences is facilitated by choosing a threshold for the proposed numerical distance measure, eventually selecting only those cyclic granularities for which the distributional differences are sufficient to make it an interesting display.

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 selection criterion by 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 across different datasets and temporal granularities since they have been normalized for relevant parameters.

The article is organized as follows. Section 2 introduces a new distance measure, discusses the reasoning and details the computation. Section 3 discusses how to choose a threshold to select only useful displays. Section 4 provides some simulation study on the proposed methodology. Section 5 presents an application to residential smart meter data in Melbourne to show how the proposed methodology can be used to automatically detect temporal granularities along which distributional differences are significant.

2 Proposed distance measure

We propose a measure called Weighted Pairwise Distances (wpd) to detect distributional differences in the measured variable across cyclic granularities.

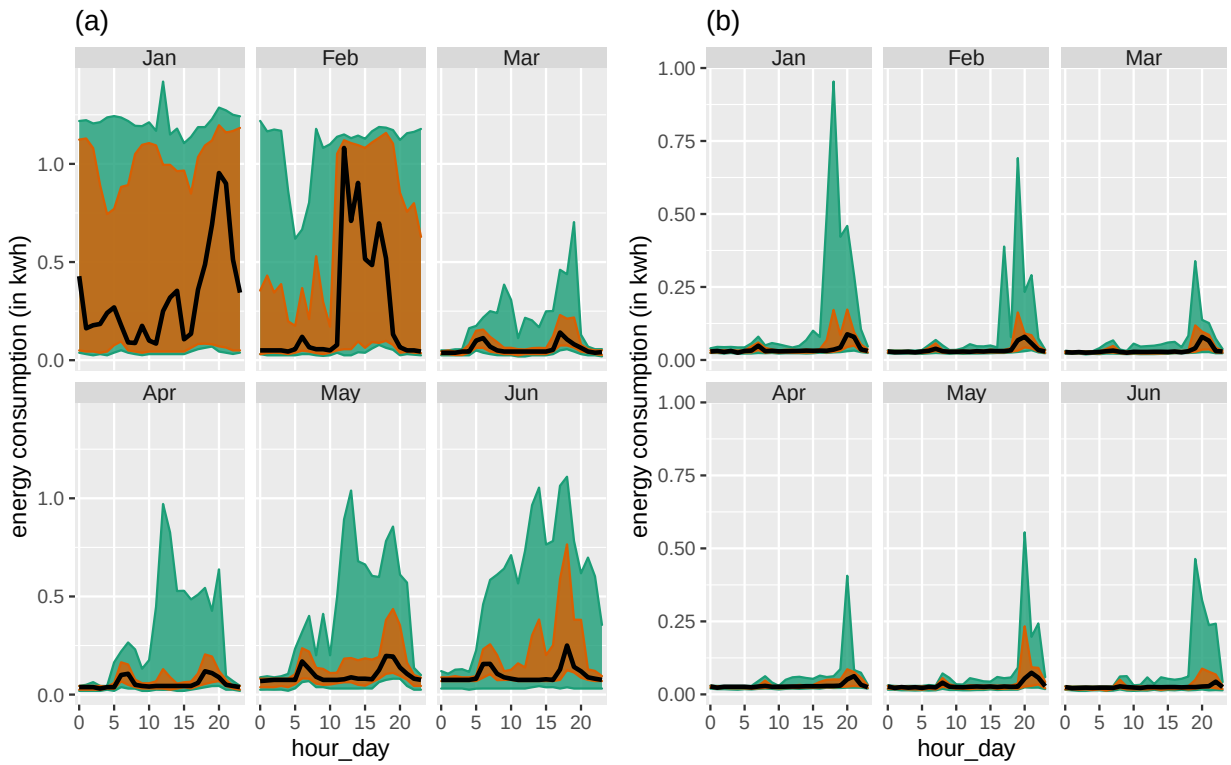

Figure 2: *Distribution of energy consumption displayed through area quantile plots across two cyclic granularities month-of-year and hour-of-day and two households. The black line is the median, whereas the orange band covers the 25th to 75th percentile and the green band covers the 10th to 90th percentile.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

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. Figure 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 items placed at close proximity can be compared more easily, because people assume that they are in the same group 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. Considering one cyclic granularity, we would have only two design choices similar to (a) and (c), corresponding to no difference and significant differences between categories of that cyclic granularity only. 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 aimed to be 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

Let the number of cyclic granularities considered in the display be m. The notations and methodology are described in detail for $m = 2$. But it can be easily extended to $m > 2$. Consider two cyclic granularities A and B, such that $A = \{a_j : j = 1,2,\ldots,J\}$ and $B = \{b_k : k = 1,2,\ldots,K\}$ with A placed across x-axis and B across facets. Let $v = \{v_t : t = 0,1,2,\ldots,T-1\}$ be a continuous variable observed across T time points. This data structure with J x-axis levels and K facet

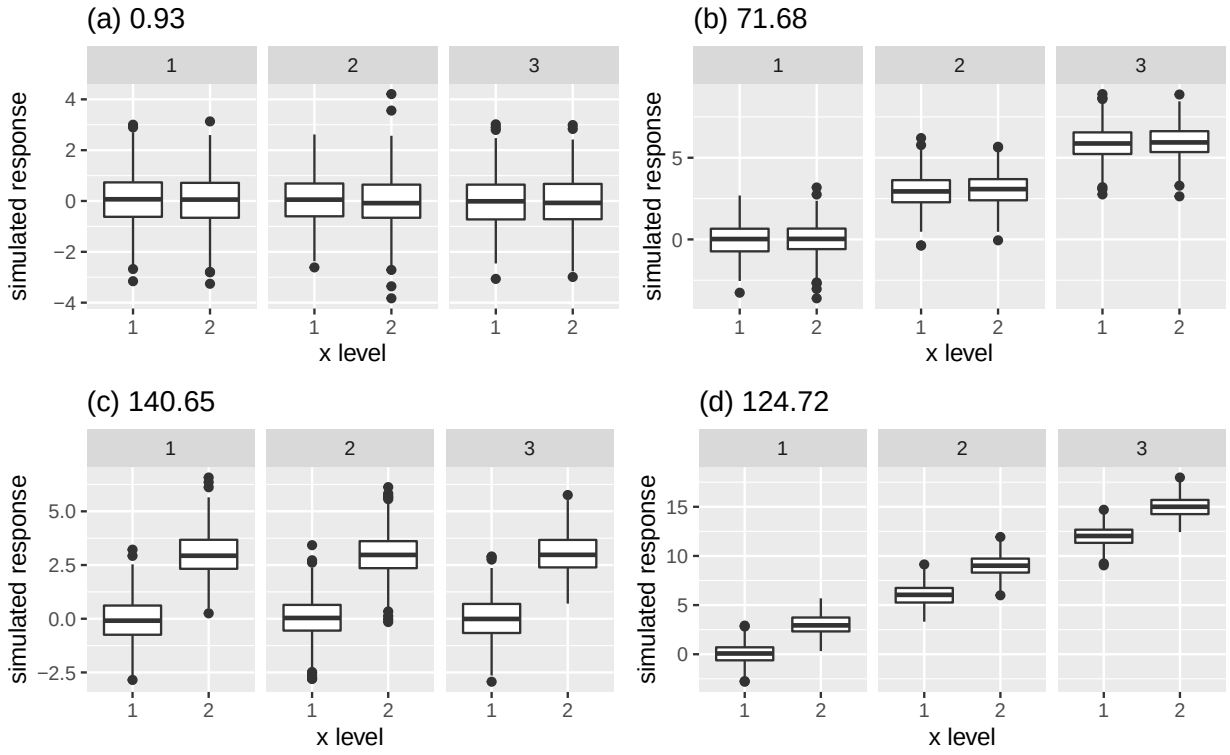

Figure 3: *An example illustrating the principle of the proposed distance measure, displaying the distribution of a normally distributed variable in four panels each with* 2 *x-axis categories and* 3 *facet levels, but with different designs. Display (a) is not interesting as the distribution of the 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.*

levels is referred to as a (J,K) 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} (referred to as "null distribution") 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. $m = 1$ is a special case of $m = 2$ with $J = 1$.

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
λ	tuning parameter
ω	increment (mean or sd)
wpd	raw weighted pairwise distance
wpd_{norm}	normalized weighted pairwise distance
$nperm$	number of permutations for threshold/normalization
$nsim$	number of simulations
$wpd_{threshold}$	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 and summarizing those weighted distances appropriately to estimate maximum variation. Furthermore, the data needs to be appropriately transformed to ensure that the value of *wpd* emphasizes detection of 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 $v_{(1)},\ldots,v_{(i)},..,v_{(n)}$.

2. The cumulative probabilities $p_{(1)},\ldots,p_{(i)},..,p_{(n)}$ 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. in this paper, quantiles of v_{jk}'s are chosen as a way to characterize distributions for the category (a_j, b_k), $\forall j \in \{1,2,\ldots,J\}, k \in \{1,2,\ldots,K\}\}$. 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

$p = 0.01, 0.02, \ldots, 0.99$, 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.

2.3.3 Distance between distributions

One of the most common ways 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. The KL divergence, however, is not symmetric and hence can not be considered as a "distance" measure. The Jensen-Shannon divergence (Menéndez et al. 1997) based on the Kullback-Leibler divergence is symmetric and has a finite value. Hence, in this paper, the pairwise distances between the distributions of the measured variable are obtained through the square root of the Jensen-Shannon divergence, called Jensen-Shannon distance (JSD) and is 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. Other common measures of distance between distributions are Hellinger distance, total variation distance and Fisher information metric.

2.3.4 Within-facet and between-facet distances

Pairwise distances could be within-facets or between-facets for $m \geq 2$. Figure 4 illustrates how they 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 balance weightage between designs like 3(b) and (c). The tuning parameters should be chosen such that $\sum_{i=1}^{m} \lambda_i = 1$. When $m = 2$, following the principle of Gestalt theory, $\lambda = \frac{2}{3} = 0.67$ is chosen to put a relative weightage of $2:1$ for within-facet and between-facet distances. 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). For $m = 1$, since there are no conditioning variables or groups, $\lambda = 1$.

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 $(a_j b_k, a_{j'} b_{k'}) : \{a_j : j = 1,2,\ldots,J\}, B = \{b_k : k = 1,2,\ldots,K\}$. 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 (*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

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

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 categories, and not because the number of categories J or K is high. That is, under designs like D_{null}, the distribution of the wpd values should not differ for a different number of categories. Only then the distance measure could be compared across panels with different levels. This calls for an adjusted measure, which normalizes for the different number of comparisons. We denote it by wpd. Two approaches for adjusting the number of comparisons are discussed, both of which are substantiated using simulations. The first one defines an adjusted measure wpd_{perm} based on the permutation method to remove the effect of different comparisons. The second approach fits a model to represent the relationship between wpd_{raw} and the number of comparisons and defines the adjusted measure ($wpd_{glm-scaled}$) as the residual from the model.

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 accomplishes a different goal of finding the null distribution for different groups (panels in our case) and standardizing the raw values using that distribution. The values of wpd_{raw} is computed on many ($nperm$) permuted data sets stored in $wpd_{perm-data}$. Then 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} under D_{null} has the same $mean = 0$ and $\sigma^2_{perm} = 1$ across all comparisons. While this works successfully to make the location and scale similar across different nx and $nfacet$, 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

In this approach, a Gamma generalized linear model (GLM) for wpd_{raw} is fitted with number of comparisons as the explanatory variable. Since, wpd_{raw} is a Jensen-Shannon distance, it follows a Chi-square distribution (Menéndez et al. (1997)), which is a special case of Gamma distribution. Furthermore, the mean response is bounded, since any JSD is bounded by 1 given that base 2 logarithm is used (Lin (1991)). Hence, by Faraway (2016), an inverse link is used for the model, which is of the form $y = a + b * log(z) + e$, where $y = wpd_{raw}$, $z = (nx * nfacet)$ is the number of groups and e 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:

$$wpd_{glm} = wpd_{raw} - 1/(\hat{a} + \hat{b} * log(nx * nfacet))$$

The distribution of wpd_{glm} under D_{null} will have $mean = 0$, since it is the residuals from the model, and a constant variance σ^2_{glm}, which might not equal 1.

Combination approach

The simulation results (4) show that the distribution of wpd_{glm} under null is similar for high nx and $nfacet$ (levels higher than 5) and not so much for lower nx and $nfacet$. Hence, a combination approach is proposed which chooses permutation approach for categories with smaller levels and modeling approach for categories with higher levels. This ensures that the computational load of the permutation approach is alleviated while maintaining similar null distribution across different categories. This approach, however, requires that the adjusted variables from the two approaches are brought to the same scale. We define $wpd_{glm-scaled} = wpd_{glm} * \sigma^2_{perm}/\sigma^2_{glm}$ as the transformed wpd_{glm} with a similar scale as wpd_{perm}. The adjusted measure from the combination approach, denoted by wpd is then defined as follows:

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

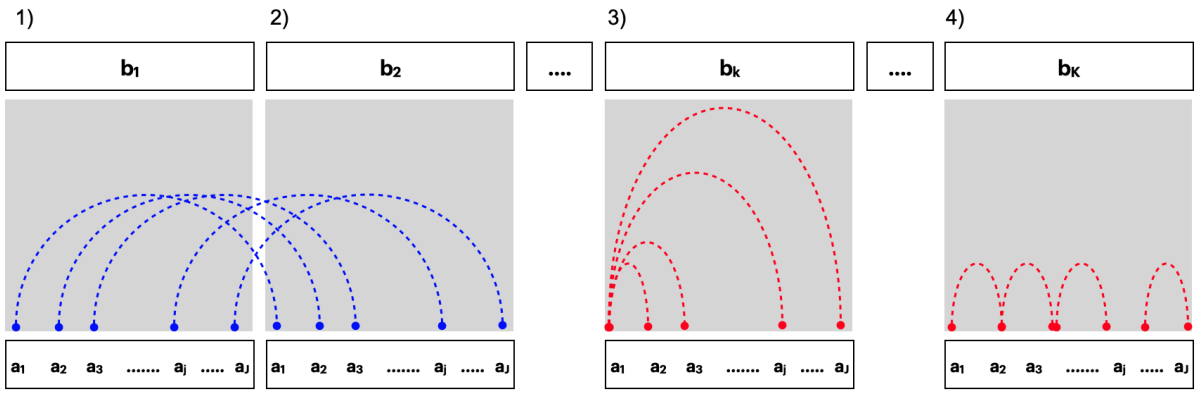

Figure 4: *Within and between-facet distances shown for two cyclic granularities A and B, where A is mapped to x-axis and B is mapped to facets. The dotted lines represent the distances between different categories. Panel 1) and 2) show the between-facet distances. Panel 3) and 4) are used to illustrate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

3 Ranking and selection of cyclic granularities

A cyclic granularity is referred to as "significant" if there is a significant distributional difference of the measured variable between different categories of the harmony. In this section, a selection criterion to choose significant harmonies is provided, thereby eliminating all harmonies that exhibit complete randomness in the measured variable. The distance measure *wpd* is used as a test statistic to test the null hypothesis that any given harmony is not significant. We select only those harmonies for which the test fails. The significant harmonies are then ranked basis how well they capture variation in the measured variable.

3.1 Selection

A threshold and consequently a selection criterion is chosen using the notion of Randomization tests. The data is permuted several times and *wpd* is computed for each of the permuted data sets to obtain the sampling distribution of *wpd* under the null hypothesis. If the null hypothesis is true, then *wpd* obtained from the original data set would be a likely value in the sampling distribution. But in case the null hypothesis is not true, then it is less probable that *wpd* obtained for the original data will be from the same distribution. This idea is utilized to come up with a threshold for selection, denoted by $wpd_{threshold}$, defined as the 99^{th} percentile of the sampling distribution. A harmony is selected if the value of *wpd* for that harmony is greater than the chosen threshold. The detailed algorithm for choosing a threshold and selection procedure is listed as follows:

- **Input:** All harmonies of the form $\{(A,B), A = \{a_j : j = 1,2,\ldots,J\}, B = \{b_k : k = 1,2,\ldots,K\}\}$, $\forall (A,B) \in H_{N_C}$.
- **Output:** Harmony pairs (A,B) for which *wpd* is significant.

1. Fix harmony pair (A,B).
2. Given the measured variable; $\{v_t : t = 0,1,2,\ldots,T-1\}$, *wpd* is computed and is represented by $wpd_{obs}^{A,B}$.
3. From the original sequence a random permutation is obtained: $\{v_t^1 : t = 0,1,2,\ldots,T-1\}$.
4. *wpd* is computed for the permuted sequence of the data and is represented by $wpd_1^{A,B}$.

5. Steps (3) and (4) are repeated a large number of times M ($M = 200$).

6. For each permutation, one $wpd_i^{A,B}$ is obtained. Define $wpd_{sample} = \{wpd_1^{A,B} wpd_2^{A,B}, \ldots, wpd_M^{A,B}\}$.

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

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

9. If $wpd_{obs}^{A,B} > wpd_{threshold99}$, harmony pair (A,B) is selected, otherwise rejected.

Similarly, a harmony pair (A,B) could be selected 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_{sample}^{all} respectively. A harmony selected using 99^{th}, 95^{th} and 90^{th} threshold are tagged as ***, **, * respectively.

3.2 Ranking

The distribution of *wpd* is expected to be similar for all harmonies under the null hypothesis, since they have been adjusted for different number of categories for the harmonies or underlying distribution of the measured variable. Hence, the values of *wpd* for different harmonies are comparable and can be used to rank the significant harmonies. A higher value of *wpd* for a harmony indicates that higher maximum variation in the measured variable is captured through that harmony. Figure 3 presents the results of *wpd* from the illustrative designs in Section 2. The value of *wpd* under null design (a) is the least, followed by (b), (c) and (d). This aligns with the principle of *wpd*, which is expected to have lowest value for null designs and highest for designs of the form $D_{var_{all}}$ (d). Moreover, note the relative differences in *wpd* values between (b) and (c). The value of the tuning parameter λ is set to 0.67, which has resulted in giving more emphasis to differences in x-axis categories. Again consider 1(a) and 1(b) with a *wpd* value of 20.5 and 145 respectively. This is because there is more gradual increase across hours of the day than months of the year. If order is not considered, they result in a *wpd* value of 97.8 and 161 respectively, which follows from the fact that if we consider difference between any hours of the day, the maginitude will be much larger than if we consider difference between consecutive categories. Simularly, for 1(a) and (b) has 110.79 and 125.82 as the *wpd* values and both are significant. The

ranking implies that the distributional differences are more prominent for the second household, as is also seen from the bigger fluctuations in the 90^{th} percentile than the first household.

4 Simulations

4.1 Behavior of raw and adjusted distance measures

Simulation design

$m = 1$

Observations are generated from a N(0,1) distribution for each $nx = \{2,3,5,7,9,14,17,20,24,31,42,50\}$ to cover a wide range of levels from very low to moderately high. $ntimes = 500$ observations are drawn for each combination of the categories, that is, for a panel with $nx = 3$, 500 observations are simulated for each of the categories. 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 categories $nsim = 200$ times, so that the distribution of wpd under D_{null} could be observed. The values of wpd is obtained for each of the panels. $wpd_{l,s}$ denotes the value of wpd obtained for the l^{th} panel and s^{th} simulation.

$m = 2$

Similarly, observations are generated from a N(0,1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = nfacet = \{2,3,5,7,14,20,31,50\}$. 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, $\{(1,1),(1,2),(2,1),(2,2)\}$.

Results

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 $nfacet$. 2 gives the summary of the generalized linear model to capture the relationship between wpd_{raw} and number of comparisons. Thus, the model considered is $wpd_{l,s} = 23.69448 + -1.02357 * log(nx * nfacet)) + e$. The intercepts are similar independent of the starting distribution (Please refer to supplementary). Figure 6 shows the distribution of

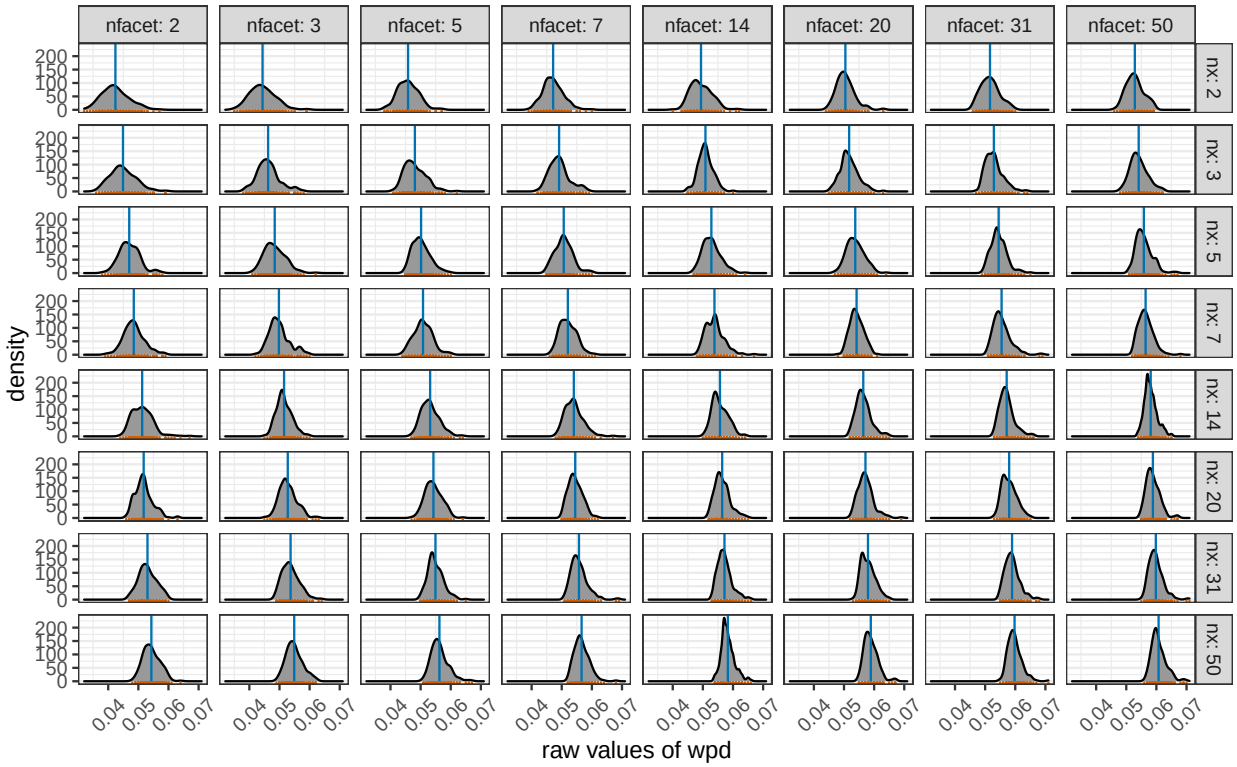

Figure 5: *Distribution of wpd_{raw} is plotted across different nx and $nfacet$ categories under D_{null} through density and rug plots. Both location (blue line) and scale (orange marks) of the distribution shifts for different panels. This is not desirable since under null design, the distribution is not expected to capture any differences.*

wpd_{perm} and $wpd_{glm-scaled}$ in the same scale to show that a combination approach could be used for higher values of levels to alleviate the computational time of permutation approach.

4.2 Choosing threshold

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 $nfacet = \{2,9,10\}$. This would result in 9 panels, viz, $(3,2),(3,9),(3,10),\ldots,(14,9),(14,10)$. Few experiments were conducted. In the first scenario, data for all panels are simulated using the null design D_{null}. In other scenarios, data simulated from the panel $(14,2)$ and $(3,10)$ are under $D_{vary_{all}}$. Moreover, $\omega = \{0.5,2,5\}$ are considered to examine if the proposed test is able to capture subtle differences and non-subtle

Table 2: *Results of generalised linear model to capture the relationship between wpd_{raw} and number of comparisons.*

m	term	estimate	std.error	statistic	p.value
1	(Intercept)	26.0863516	0.5397440	48.330973	0.0e+00
1	log(‘nx * nfacet‘)	-1.8745975	0.1894537	-9.894751	1.8e-06
2	(Intercept)	23.3996082	0.2247005	104.136879	0.0e+00
2	log(‘nx * nfacet‘)	-0.9571158	0.0439971	-21.754076	0.0e+00

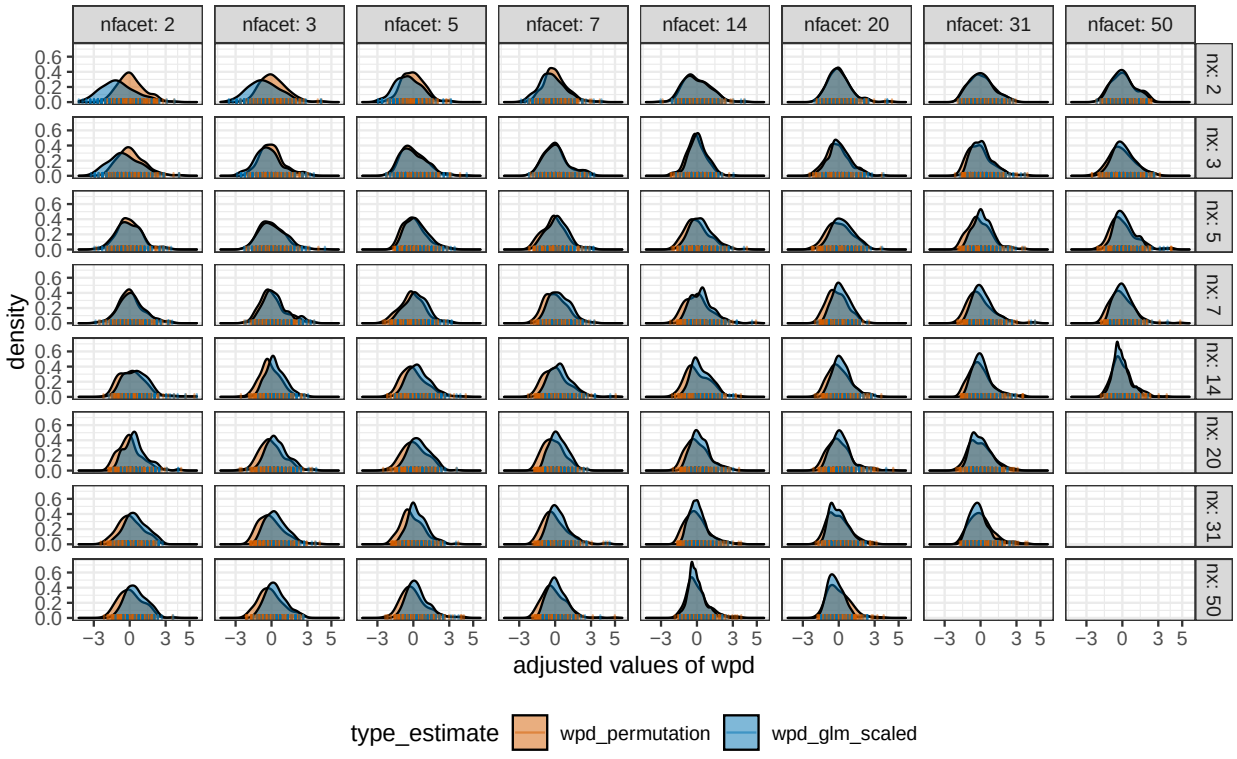

Figure 6: *The distribution of wpd_{perm} and $wpd_{glm-scaled}$ are overlaid to compare the location and scale across different nx and nfacet. wpd_{norm} takes the value of wpd_{perm} for lower levels, and $wpd_{glm-scaled}$ for higher levels to to alleviate the problem of computational time in permutation approaches. This is possible as the distribution of the adjusted measure looks similar for both appraches for higher levels.*

differences when we shift from the null design. In the last scenario, we consider the panel $(3,2),(7,9),(14,10)$ to be under D_{null}, the panels $(7,2),(14,9)$ to be under D_{var_f}. $(14,2),(3,10)$ under D_{var_x} and the rest under $D_{var_{null}}$. This is done to check if the consequent ranking procedure leads to designs like $D_{vary_{all}}$ to be chosen first followed by $D_{vary_{all}}$. We generate only one data set each for which these scenarios were simulated and consider this as the original data set. We generate 1000 repetitions of this experiment with different seeds.

Results

The proportion of times a panel is rejected when it is under D_{null} is computed with $wpd_{threshold99}$ and it is around 0.1. This is obvious, since if the level of significance for each test is less than 0.01 (as a result of choosing 99^{th} percentile as the threshold) and we have 9 tests, the size for the entire test should be close to $0.01 * 9 \approx 0.1$. 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. It is found that as we increase from low to high changes from the null distribution, the power increased. The results and graphics are included in details in the Supplementary paper.

4.3 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 High Power Computing (HPC) Cluster, addressing the 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 *wpd* 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.

No behavioral pattern is likely to be discerned from the line display of energy usage over the entire period, since the plot will have too many measurements squeezed in a linear representation. When we zoom into the linear representation of this series in Figure 7 (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 of these households, compare households to get more insights into what these rankings imply. Furthermore, the ranking and selection of significant harmonies is validated by analyzing the distribution of 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_i is computed on $v_{i,t}$ 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 highlighted with $***$, $**$ and $*$ corresponding to the 99^{th}, 95^{th} and 90^{th} percentile threshold. 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 the number of potential harmonies to explore closely, starting from 42.

Detecting patterns for households not apparent from linear display

Figure 7 helps to compare households through the heatmap (a) across harmony pairs with the cyclic granularity mapped to x-axis and facet being plotted on the x-axis and y-axis of the heatmap. Here dom, dow, wdwnd are abbreviations for day-of-month, day-of-week and weekday/weekend and so on. (a) emphasizes patterns not discernible through (b), which is a linear display containing the raw data for one month (Sep-2019), with the major and minor x-axis corresponding to weeks and days respectively. In (a), the colors represent the value of wpd, implying darker cells correspond to more significant harmony pairs. Also, the ones with $*$ corresponds to the ones above $wpd_{threshold95}$. 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. id 7 and 8 have the same significant harmonies despite having very different total energy usage. Note that the wpd values are computed over the entire range, but the linear display is zoomed into September.

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

facet variable	x variable	id 1	id 2	id 3	id 4	id 5	id 6	id 7	id 8
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 8 shows that the median (black) and quartile deviation (orange) of energy consumption changes across both dow and hod for id 4 for most hours, but it is not so different for id 5 with differences captured only for 90^{th} percentile.

6 Discussion

Exploratory data analysis involves many iterations of finding and summarizing patterns. With temporal data available at finer scales, exploring time series has become overwhelming with so many possible granularities to explore. A common solution is to aggregate and look at the patterns across usual granularities like hour-of-day or day-of-week, but there is no way to know the "interesting" granularities a priori. A huge number of displays need to be analyzed or we might end up missing informative granularities. This work refines the search of informative granularities by identifying those for which the differences between the displayed distributions are greatest and rating them in order of importance of capturing maximum variation.

The significant granularities across different datasets (individuals/subjects) do not imply similar patterns across different datasets. They simply mean that maximum distributional differences are being captured across those granularities. A future direction of work is to be able to explore and compare many individuals/subjects together for similar patterns across significant granularities.

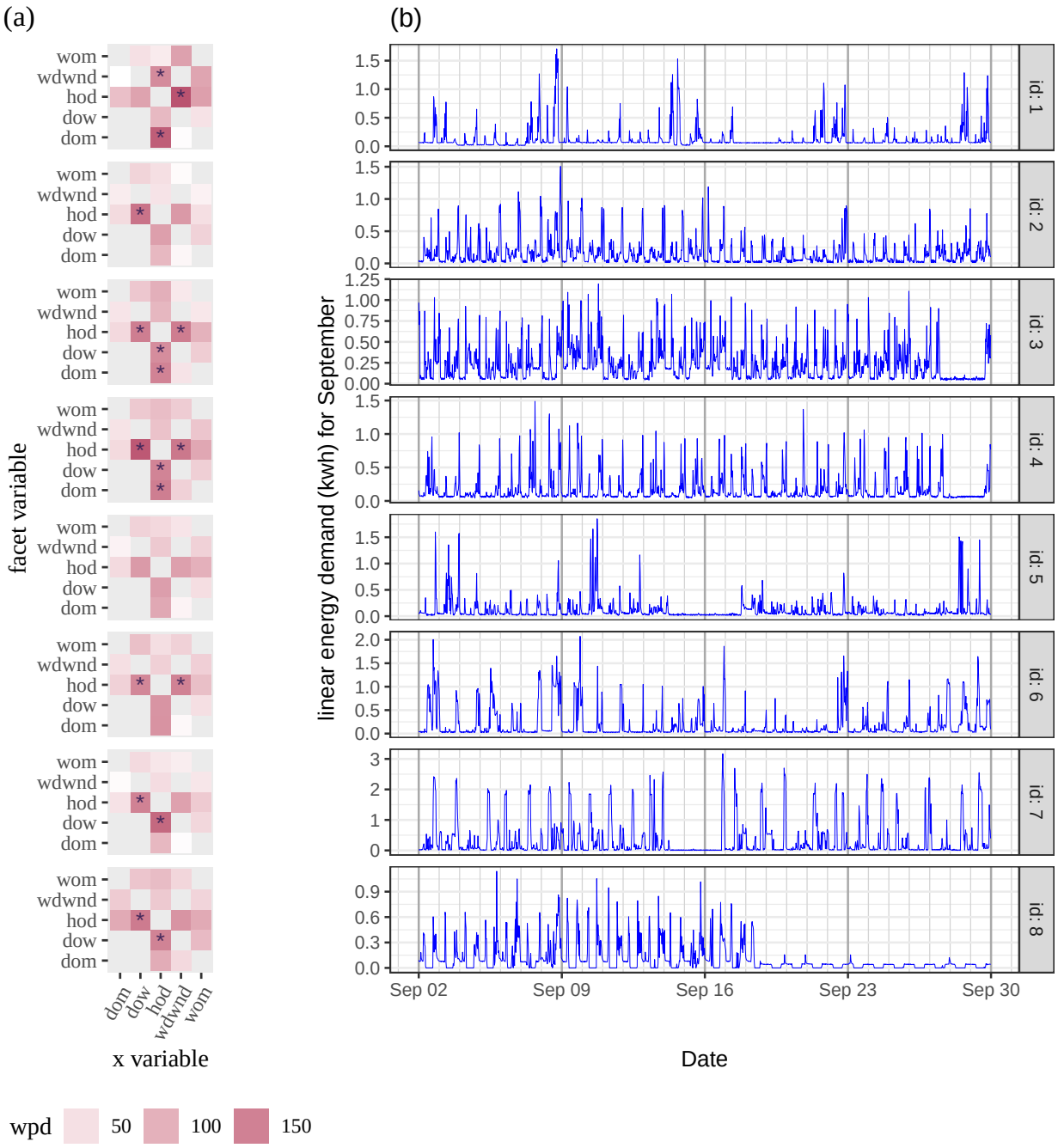

Figure 7: *A heatmap across harmonies (a) and line plot (b) showing energy usage for all household ids. The darker the colour in (a), the higher the harmony is ranked. Visualizing the harmony pairs in this way helps us to see the important cyclic granularities along the x-axis, facet or both. It further shows which households have significant patterns across similar temporal granularities. For eg, id 7 and 8 have significant patterns across (hod, dow) and (dow, hod), which was not apparent in (a).*

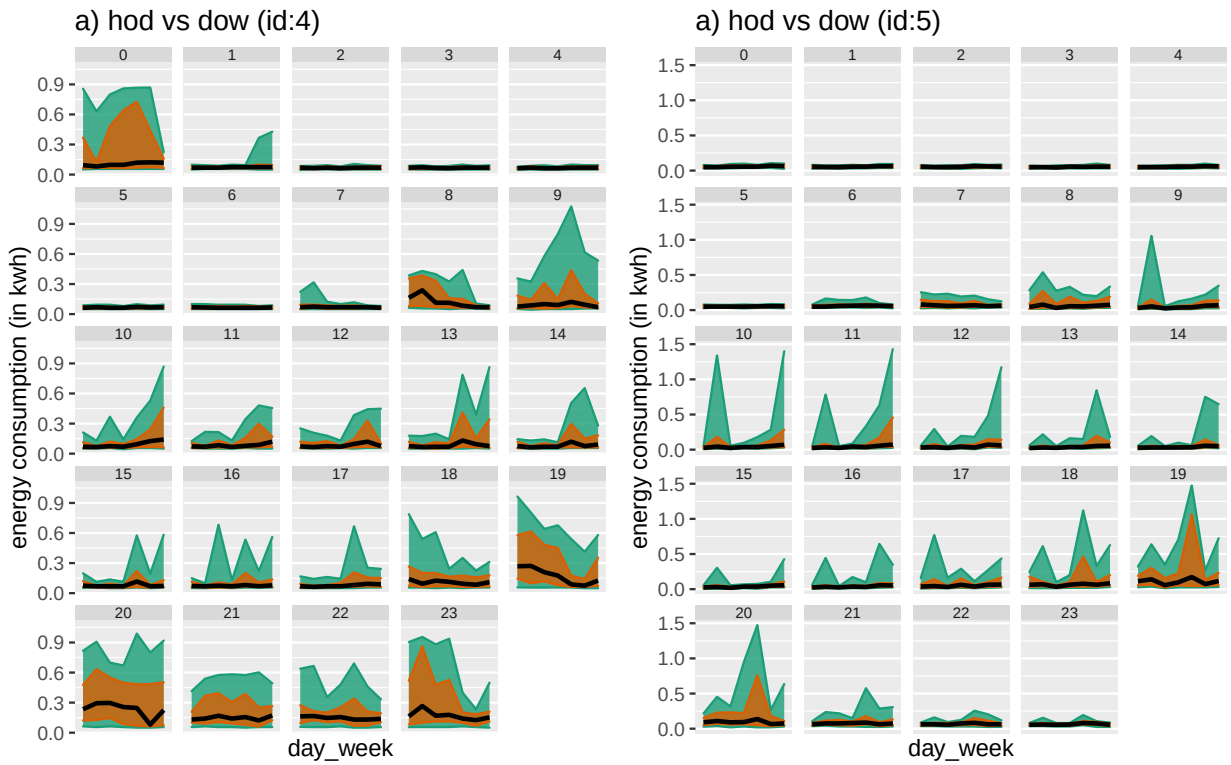

Figure 8: *Comparing distribution of energy demand shown for household id 4 (a) and 5 (b) across dow in x-axis and hod in facets through quantile area plots. The value of wpd in Table 3 suggests that the harmony pair (dow, hod) is significant for household id 4, but not for 5. This implies that distributional differences are captured more by this harmony for id 4, which is apparent from the display with more fluctuations across median and 75th percentile. Here, the median is represented by the black line, the orange area corresponds to quartile deviation and the green area corresponds to area between* 10^{th} *and* 90^{th} *quantile.*

Acknowledgments

The Australian authors thank the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) for supporting this research. Sayani Gupta was partially funded by Data61 CSIRO during her PhD. The Github repository, github.com/Sayani07/paper-hakear, contains all materials required to reproduce this article and the code is also available online in the supplemental materials. This article was created with R (R Core Team 2019), `knitr` (Xie 2015, Xie (2020)) and `rmarkdown` (Xie et al. 2018, Allaire et al. (2020)).

7 Supplementary Materials

Data and scripts: Data sets and R code to reproduce all figures in this article (main.R).

Simulation results and scripts: All simulation table, graphics and and R code to reproduce it (paper-supplementary.pdf, paper-supplementary.Rmd)

R-package: The open-source R (R Core Team 2019) package `hakear` (**?**) is available on Github (`https://github.com/Sayani07/hakear`) to implement ideas presented in this paper.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W. & Iannone, R. (2020), *rmarkdown: Dynamic Documents for R*. R package version 2.1.

URL: *https://github.com/rstudio/rmarkdown*

Bettini, C., Dyreson, C. E., Evans, W. S., Snodgrass, R. T. & Wang, X. S. (1998), A glossary of time granularity concepts, *in* O. Etzion, S. Jajodia & S. Sripada, eds, 'Temporal Databases: Research and Practice', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 406–413.

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.

Faraway, J. J. (2016), *Extending the Linear Model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second Edition*, 2nd edition edn, Chapman and Hall/CRC.

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 Trans. Inf. 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.

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

Tukey, J. W. (1977), *Exploratory data analysis*, Addison-Wesley, Reading, Mass.

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.

Xie, Y. (2015), *Dynamic Documents with R and knitr*, 2nd edn, Chapman and Hall/CRC, Boca Raton, Florida.
URL: *https://yihui.org/knitr/*

Xie, Y. (2020), *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.28.
URL: *https://yihui.org/knitr/*

Xie, Y., Allaire, J. & Grolemund, G. (2018), *R Markdown: The Definitive Guide*, Chapman and Hall/CRC, Boca Raton, Florida.
URL: *https://bookdown.org/yihui/rmarkdown*