MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://monash.edu/business/ebs/research/publications

Detecting distributional differences between temporal granularities for exploratory time series analysis

Sayani Gupta, Rob J Hyndman, Dianne Cook

November 2021

Working Paper ??/2021

Detecting distributional differences between temporal granularities for exploratory time series analysis

Sayani Gupta
Monash University
Email: Sayani.Gupta@monash.edu
Corresponding author

Rob J Hyndman
Monash University

Dianne Cook
Monash University

10 November 2021

JEL classification: C55, C65, C80

Detecting distributional differences between temporal granularities for exploratory time series analysis

Abstract

Cyclic temporal granularities are temporal deconstructions of a time period into units such as hour-of-the-day and work-day/weekend. They can be useful for measuring repetitive patterns in large univariate time series data, and feed new approaches to exploring time series data. One use is to take pairs of granularities, and make plots of response values across the categories induced by the temporal deconstruction. However, when there are many granularities that can be constructed for a time period, there will also be too many possible displays to decide which might be the more interesting to display. This work proposes a new distance metric to screen and rank the possible granularities, and hence choose the most interesting ones to plot. The distance measure is computed for a single or pairs of cyclic granularities and can be compared across different cyclic granularities or on a collection of time series. The methods are implemented in the open-source R package `hakear`.

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

1 Introduction

Cyclic temporal granularities (Bettini et al. 1998; Gupta et al. 2021) are temporal deconstructions that define cyclic repetitions in time, e.g. hour-of-day, day-of-month, or regularly scheduled public holidays. These granularities form ordered or unordered categorical variables. An example of an ordered granularity is day-of-week, where Tuesday is always followed by Wednesday, and so on. An unordered granularity example is week type in an academic semester: orientation, break, exam or regular classes. We can use granularities to explore patterns in univariate time series by examining the distribution of the measured variable across different categories of the cyclic granularities.

As a motivating example, consider Figure 1 which shows electricity smart meter data plotted against two granularities (hour-of-day, month-of-year). The data was collected on a single household in Melbourne, Australia, over a six month period, and was previously used in Wang, Cook & Hyndman (2020b). The

categorical variable (granularity) is mapped to the x-axis, and the distribution of the response variable is displayed using both side-by-side jittered dotplots and boxplots. From panel (a) it can be seen that energy consumption is higher during the morning hours (5–8), when members in the household wake up, and again in the evening hours (17–20) possibly when members get back from work. In addition, the largest variation in energy use is in the afternoon hours (12–16), as seen in the length of the boxes. From panel (b), it is seen that the variability in energy usage is higher in Jan and Feb, probably due to the usage of air conditioners on some days. The median usage is highest in January, dips in February–April and rises again in May–June, although not to the height of January usage. This suggests that the household does not use as much electricity for heating as it does for air conditioning. A lot of households in Victoria use gas heating and hence the heater use might not be reflected in the electricity data.

Many different displays could be constructed using different granularities including day-of-week, day-of-month, weekday/weekend, etc. However, only a few might be interesting and reveal important patterns in energy usage. Determining which displays have "significant" distributional differences between categories of the cyclic granularity, and plotting only these, would make for efficient exploration.

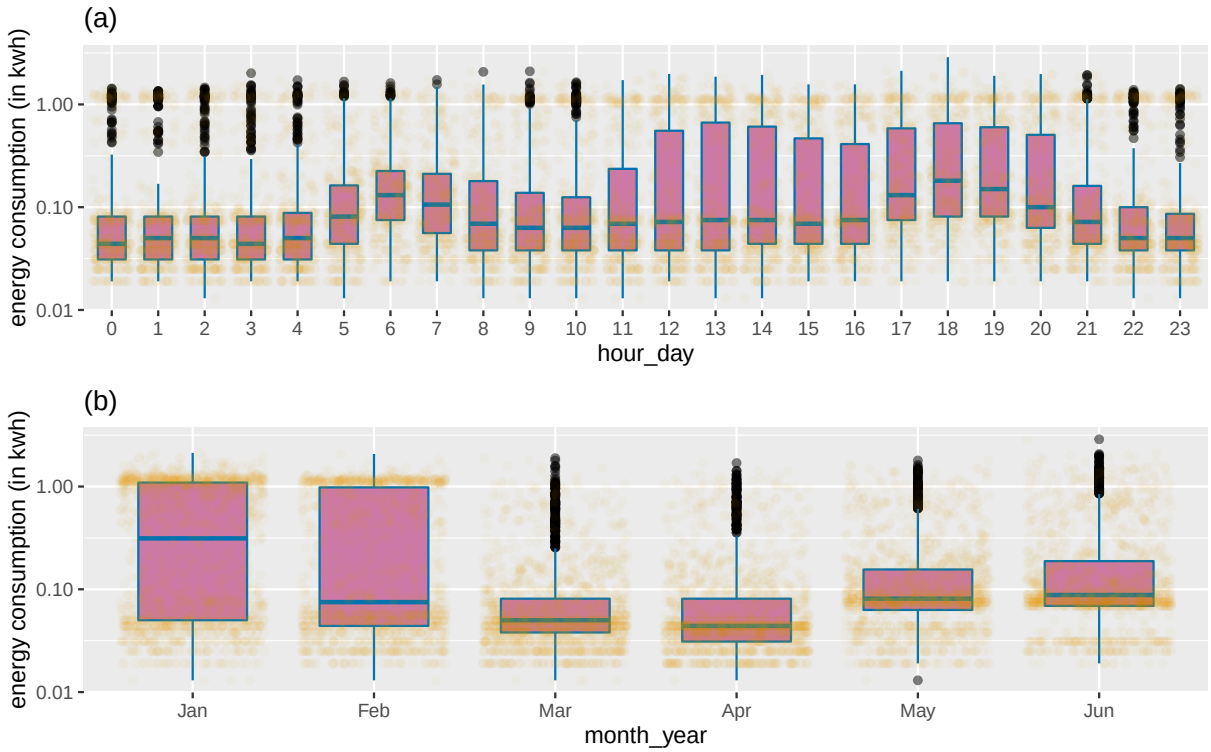

Figure 1: *A cyclic granularity can be considered to be a categorical variable, and used to break the data into subsets. Here, side-by-side boxplots overlaid on jittered dotplots explore the distribution of of energy use by a household for two different cyclic granularities: (a) hour-of-day and (b) and month-of-year. Daily peaks occur in morning and evening hours, indicating a working household, where members leave for and return from work. More volatility of usage in summer months (Jan, Feb) is probably due air conditioner use on just some days.*

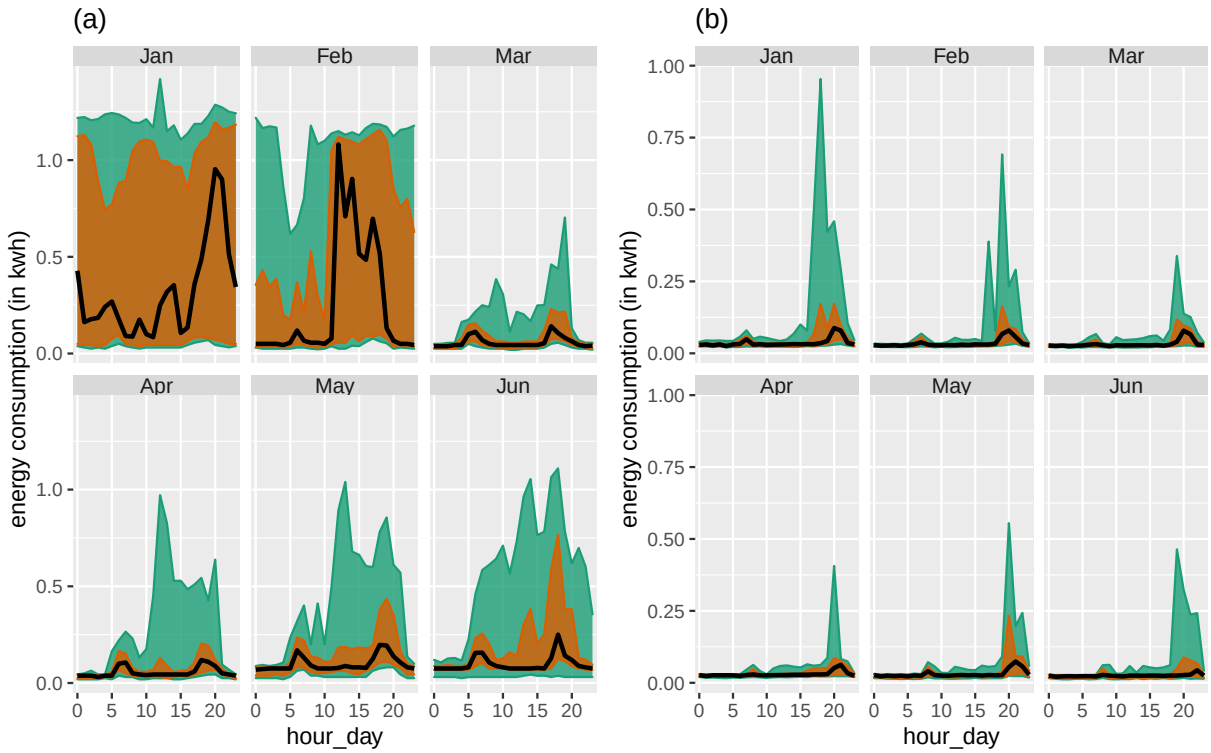

Figure 2: *Distribution of energy consumption displayed through area quantile plots across two cyclic granularities month-of-year and hour-of-day and two households. The black line is the median, whereas the orange band covers the 25th to 75th percentile and the green band covers the 10th to 90th percentile. 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.*

Exploring the distribution of the measured variable across two cyclic granularities provides more detailed information on its structure. For example, Figure 2(a) shows the usage distribution across hour-of-day conditional on month-of-year across two households. 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, similar, 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 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 $N_C(N_C-1)$ possible pairwise plots exhaustively, with one of the two cyclic granularities acting as the conditioning variable. This can be overwhelming for human consumption even for moderately large N_C. It is therefore useful to identify those displays that are informative across at least one cyclic granularity.

This problem is similar to Scagnostics (Scatterplot Diagnostics) by Tukey & Tukey (1988), which are 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) can easily run out of pixels when presenting high-dimensional data. Dang & Wilkinson (2014) and Wilkinson, Anand & Grossman (2005) provide potential solutions to this, where a 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 $N_C(N_C-1)$ conditional distribution 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. 2021) which narrows down the search from $N_C(N_C-1)$ 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 may be too large for exhaustive exploration. Hence, there is a need to reduce the search even further by including only those harmonies that contain useful information.

Buja et al. (2009) and Majumder, Hofmann & Cook (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.

The article is organized as follows. Section 2 introduces a distance measure for detecting distributional difference in temporal granularities, which enables identification of patterns in the time series data; Section 3 devises a selection criterion by choosing a threshold, which results in detection of only significantly interesting patterns. Section 3.3 provides a simulation study on the proposed methodology. Section 4 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.

2.1 Principle

The principle behind the construction of *wpd* is explained through a simple example in Figure 3. Each of these figures describes a panel with two x-axis categories and three facet levels, but with different designs. Figure 3a has all categories drawn from a standard normal distribution for each facet. It is not a particularly interesting display, as the 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 three standard deviations apart. Figure 3c exhibits the opposite situation where distribution between the x-axis categories

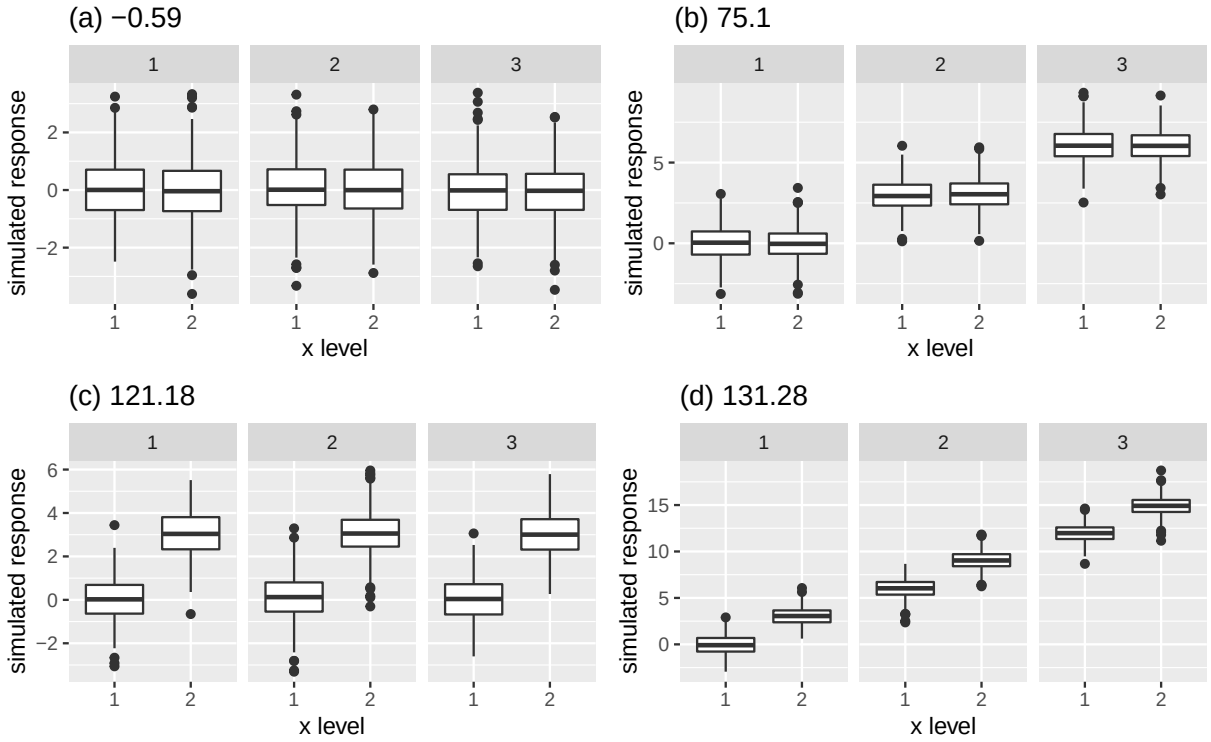

Figure 3: *An example illustrating the principle of the proposed distance measure, displaying the distribution of a normally distributed variable in four panels each with two x-axis categories and three facet levels, but with different designs. Panel (a) is not interesting as the distribution of the variable does not depend on x or facet categories. Panels (b) and (c) are more interesting than (a) since there is a change in distribution either across facets (b) or x-axis (c). Panel (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.*

are three standard deviations apart, but they do not change across facets. In Figure 3d, the distribution varies across both facet and x-axis categories by three 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 (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, Panel (b) is considered less informative compared to Panel (c) in emphasizing the distributional differences.

For displays showing a single cyclic granularity rather than pairs of granularities, we have only two design choices corresponding to no difference and significant differences between categories of that cyclic granularity.

The proposed measure *wpd* is constructed in such a way that it can 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 Notation

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 $A = \{a_j : j = 1,2,\ldots,n_x\}$ and $B = \{b_k : k = 1,2,\ldots,n_f\}$ with A placed across the 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 n_x x-axis levels and n_f facet levels is referred to as a (n_x, n_f) panel. For example, a $(2,3)$ panel will have cyclic granularities with two x-axis levels and three facet levels. Let the four elementary designs as described in Figure 3 be $D_\emptyset$ (referred to as "null distribution") where there is no difference in distribution of v for A or B, D_f denotes the set of designs where there is difference in distribution of v for B and not for A. Similarly, D_x denotes the set of designs where difference is observed only across A. Finally, D_{fx} denotes those designs for which difference is observed across both A and B. We can consider a single granularity ($m = 1$) as a special case of two granularities with $n_f = 1$.

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 weight for different groups of distances and summarizing those weighted distances appropriately to estimate maximum variation.

Table 1: *Nomenclature table*

variable	description
N_C	number of cyclic granularities
H_{N_C}	set of harmonies
n_x	number of x-axis categories
n_f	number of facet categories
λ	tuning parameter
ω	increment (mean or sd)
wpd	raw weighted pairwise distance
wpd_{norm}	normalized weighted pairwise distance
n_{perm}	number of permutations for threshold/normalization
n_{sim}	number of simulations
$wpd_{\text{threshold}}$	threshold for significance
$D_\emptyset$	null design with no distributional difference across categories
D_f	design with distributional difference only across facets categories
D_x	design with distributional difference only across x-axis categories
D_{fx}	design with distributional difference across both facet and x-axis

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.

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 transformed data follows a standard normal distribution. The empirical NQT involves the following steps:

1. The observations of measured variable v are sorted from the smallest to the largest observation $v_{(1)},\ldots,v_{(n)}$.
2. The cumulative probabilities $p_{(1)},\ldots,p_{(n)}$ are estimated using $p_{(i)} = i/(n+1)$ (Hyndman & Fan 1996) so that $p_{(i)} = \Pr(v \leq v_{(i)})$.
3. Each observation $v_{(i)}$ of v is transformed into $v^*(i) = \Phi^{-1}(p(i))$, with Φ denoting the standard normal distribution function.

Characterizing 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 $\{v_{jk}\}$ are chosen as a

way to characterize distributions for the category (a_j, b_k), $\forall j \in \{1,2,\ldots,n_x\}, k \in \{1,2,\ldots,n_f\}$. We use percentiles with $p = 0.01, 0.02, \ldots, 0.99$ to reduce the computational burden in summarizing distributions.

Distance between distributions

A 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. The KL divergence 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 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.

Within-facet and between-facet distances

Pairwise distances could be within-facets or between-facets for $m = 2$. 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 $(a_j b_k, a_j b_{k'})$. There are a total of $\binom{n_f}{2} n_x$ between-facet distances, and $\binom{n_x}{2} n_f$ within-facet distances if they are ordered and $n_f(n_x - 1)$ within-facet distances if they are unordered.

Tuning parameter

For displays with $m > 1$ granularities, we can use a tuning parameter to specify the relative weight given to each granularity. In general, the tuning parameters should be chosen such that $\sum_{i=1}^{m} \lambda_i = 1$.

Following the general principles of Gestalt theory, we wish to weight more heavily granularities that are plotted closer together. For $m = 2$ we choose $\lambda_x = \frac{2}{3}$ for the granularity on the x-axis and $\lambda_f = \frac{1}{3}$ for the granularity mapped to facets, giving a relative weight of $2:1$ for within-facet to between-facet

distances. No human experiment has been conducted to justify this ratio. Specifying $\lambda_x > 0.5$ will weight within-facet distances more heavily, while $\lambda_x < 0.5$ would weight the between-facet distances more heavily. (See the supplements for more details.).

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 $(a_j b_k, a_{j'} b_{k'}) : \{a_j : j = 1,2,\ldots,n_x\}, B = \{b_k : k = 1,2,\ldots,n_f\}$. The pairwise distances between pairs $(a_j b_k, a_{j'} b_{k'})$ denoted by $d_{(jk,j'k')} = JSD(q_{jk}, q_{j'k'})$ are 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:

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

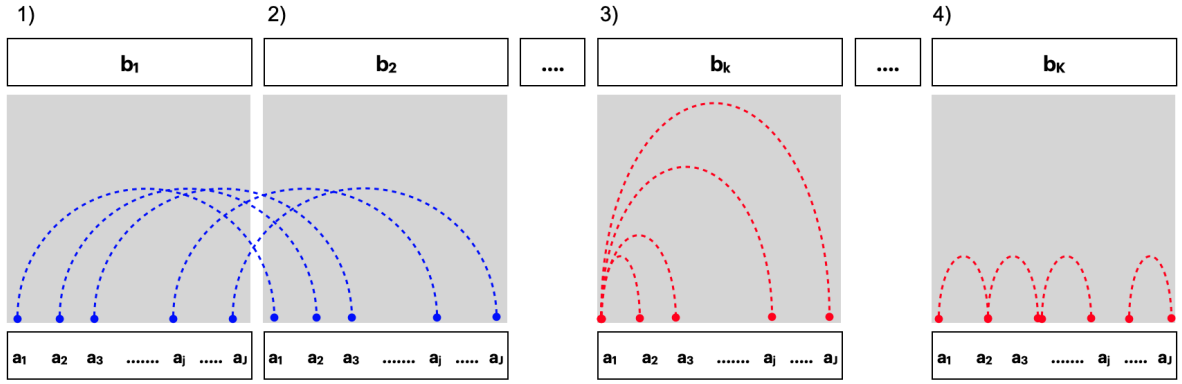

Figure 4: *Within and between-facet distances shown for two cyclic granularities A and B, where A is mapped to x-axis and B is mapped to facets. The dotted lines represent the distances between different categories. Panels 1) and 2) show the between-facet distances. Panels 3) and 4) are used to illustrate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

The wpd_{raw} is then computed as

$$wpd_{\text{raw}} = \max_{j,j',k,k'} (d*_{(jk),(j'k')}) \qquad \forall j, j' \in \{1,2,\ldots,n_x\}, \quad k,k' \in \{1,2,\ldots,n_f\}$$

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 observations 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.

2.4 Adjusting for the number of comparisons

Ideally, it is desired that the proposed distance measure takes a higher value only if there is a significant difference between distributions across categories, and not because the number of categories n_x or n_f is high. That is, under designs like $D_\emptyset$, their distribution 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.

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}) 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 (n_{perm}) permuted data sets and stored in $wpd_{\text{perm-data}}$. Then wpd_{perm} is computed as follows:

$$wpd_{\text{perm}} = \frac{(wpd_{\text{raw}} - \text{mean}(wpd_{\text{perm-data}}))}{\text{sd}(wpd_{\text{perm-data}})},$$

where mean$(wpd_{\text{perm-data}})$ and sd$(wpd_{\text{perm-data}})$ are the mean and standard deviation of $wpd_{\text{perm-data}}$ respectively. Standardizing *wpd* in the permutation approach ensures that the distribution of wpd_{perm} under $D_\emptyset$ has zero mean and unit variance across all comparisons. While this works successfully to make the location and scale similar across different n_x and n_f, it is computationally heavy and time consuming, and hence less user-friendly. 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 the 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 a Gamma distribution. Furthermore, the mean response is bounded, since any JSD is bounded by 1 if a 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 $y = wpd_{\text{raw}}$, $z = (n_x * n_f)$ is the number of groups and e are idiosyncratic errors. Let $\mathrm{E}(y) = \mu$ and $a + b * \log(z) = g(\mu)$ where $g(\mu) = 1/\mu$ 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 defined as the residuals from this model given by

$$wpd_{\text{glm}} = wpd_{\text{raw}} - 1/(\hat{a} + \hat{b} * \log(n_x * n_f))$$

The distribution of wpd_{glm} under $D_\emptyset$ will have approximately zero mean and a constant variance (not necessarily 1).

Combination approach

The simulation results (in Section 3.3) show that the distribution of wpd_{glm} under the null design is similar for high n_x and n_f (levels higher than 5) but less so for lower values of n_x and n_f. Hence, a combination approach is proposed where we use a permutation approach for categories with small numbers of levels, and a modeling approach for categories with higher numbers of levels. This ensures that the computational load of the permutation approach is alleviated while maintaining a 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 $wpd_{\text{glm-scaled}} = wpd_{\text{glm}} * \sigma^2_{\text{perm}} / \sigma^2_{\text{glm}}$ 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:

$$wpd = \begin{cases} wpd_{\text{perm}}, & \text{if } n_x, n_f \leq 5; \\ wpd_{\text{glm-scaled}} & \text{otherwise.} \end{cases} \tag{2}$$

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 non-significant differences in the measured variable. The distance measure *wpd* is used as a test statistic to test the null hypothesis that no harmony/cyclic granularity is significant. We select only those harmonies/cyclic granularities for which the test fails. They are then ranked based on 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 (Edgington & Onghena 2007). 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 $wpd_{\text{threshold}}$, 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 (for two cyclic granularities) is listed as follows:

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

1. For each harmony pair $(A,B) \in H_{N_C}$, the following steps are taken.
    a. Given the measured variable; $\{v_t : t = 0,1,2,\ldots,T-1\}$, *wpd* is computed and is represented by $wpd_{obs}^{A,B}$.

b. For $i = 1,\ldots,m$, randomly permute the original time series: $\{v_t^i : t = 0,1,2,\ldots,T-1\}$ and compute $wpd_i^{A,B}$ from $\{v_t^i\}$.

c. Define $wpd_{\text{sample}} = \{wpd_1^{A,B},\ldots,wpd_M^{A,B}\}$.

2. Stack the wpd_{sample} vectors as $wpd_{\text{sample}}^{\text{all}}$ and compute its $p = 100(1-\alpha)$ percentiles as $wpd_{\text{threshold}p}$.

3. If $wpd_{obs}^{A,B} > wpd_{\text{threshold}p}$, harmony pair (A,B) is selected at the $1-p/100$ level, otherwise rejected.

4. Harmonies selected using the 99^{th}, 95^{th} and 90^{th} thresholds 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 also 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_{fx} (d). Moreover, note the relative differences in *wpd* values between (b) and (c). The value of the tuning parameter λ is set to 2/3, which gives greater emphasis to differences in x-axis categories than facets.

Again consider 1(a) and 1(b) with a *wpd* value of 20.5 and 145 respectively. This is because there is a more gradual increase across hours of the day than across months of the year. If the order of categories is ignored, the resulting *wpd* values are 97.8 and 161 respectively, because differences between any hours of the day tend to be larger than differences only between consecutive hours. Similarly, 1(a) and (b) have *wpd* values of 110.79 and 125.82 respectively. 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 for the first household.

3.3 Simulations

Simulations were carried out to explore the behavior of *wpd* as n_x and n_f were varied, in order to compare and evaluate different normalization approaches. More detailed simulation results, including an exploration of the size and power of the proposed test, are contained in the supplements.

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

m	term	estimate	std.error	statistic	p.value
1	Intercept	26.09	0.54	48.33	0
1	$\log(n_x \times n_f)$	-1.87	0.19	-9.89	0
2	Intercept	23.40	0.22	104.14	0
2	$\log(n_x \times n_f)$	-0.96	0.04	-21.75	0

Simulation design

$m = 1$

Observations were generated from a N(0,1) distribution for $n_x \in \{2,3,5,7,9,14,17,20,24,31,42,50\}$, with $n_{\text{times}} = 500$ observations drawn for each combination of categories. This design corresponds to $D_\emptyset$. For each of the categories, there were $n_{\text{sim}} = 200$ replications, so that the distribution of *wpd* under $D_\emptyset$ could be observed. Let $wpd_{\ell,s}$ denote the value of *wpd* obtained for the ℓ^{th} panel and s^{th} simulation.

$m = 2$

Similarly, observations were generated from a N(0,1) distribution for each combination of n_x and n_f from $\{2,3,5,7,14,20,31,50\}$, with $n_{\text{times}} = 500$ observations drawn for each of the 64 combinations.

Results

Figure 5 shows that both the location and scale of the distributions change across panels. This is not desirable under $D_\emptyset$ as it would mean comparisons of *wpd* values are not appropriate across different n_x and n_f values. Table 2 gives the summary of a Gamma generalized linear model to capture the relationship between wpd_{raw} and the number of comparisons. The intercepts and slopes are similar, independent of the underlying distributions (see supplementary paper for details) and hence the coefficients are shown for the case when observations are drawn from a N(0,1) distribution. Figure 6 shows the distribution of wpd_{perm} and $wpd_{\text{glm-scaled}}$ on the same scale to show that a combination approach could be used for higher values of n_x and n_f to alleviate the computational time of the permutation approach.

These results justify our use of the permutation approach when $n_x < 5$ and $n_f < 5$, and the use of the GLM otherwise.

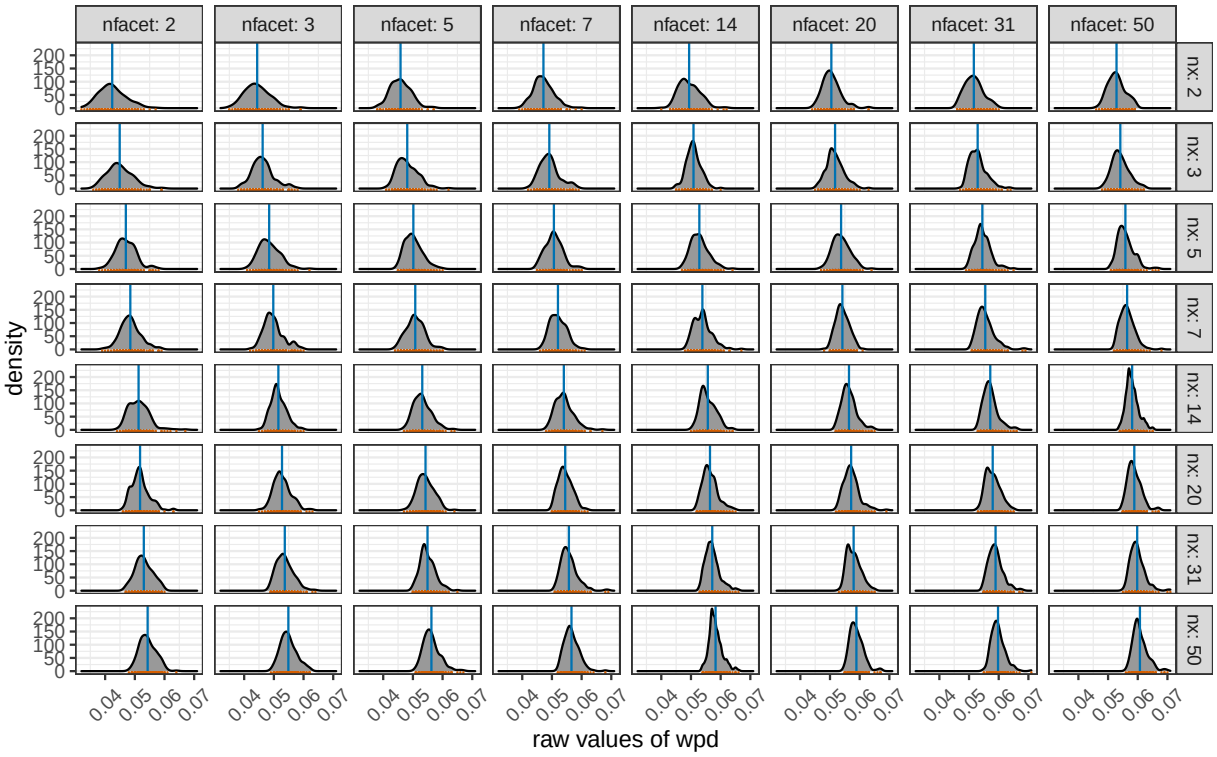

Figure 5: *Distribution of wpd_{raw} is plotted across different n_x and n_f categories under $D_\emptyset$ through density and rug plots. Both location (blue line) and scale (orange marks) of the distribution shifts for different panels. This is not desirable since under the null design, the distribution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

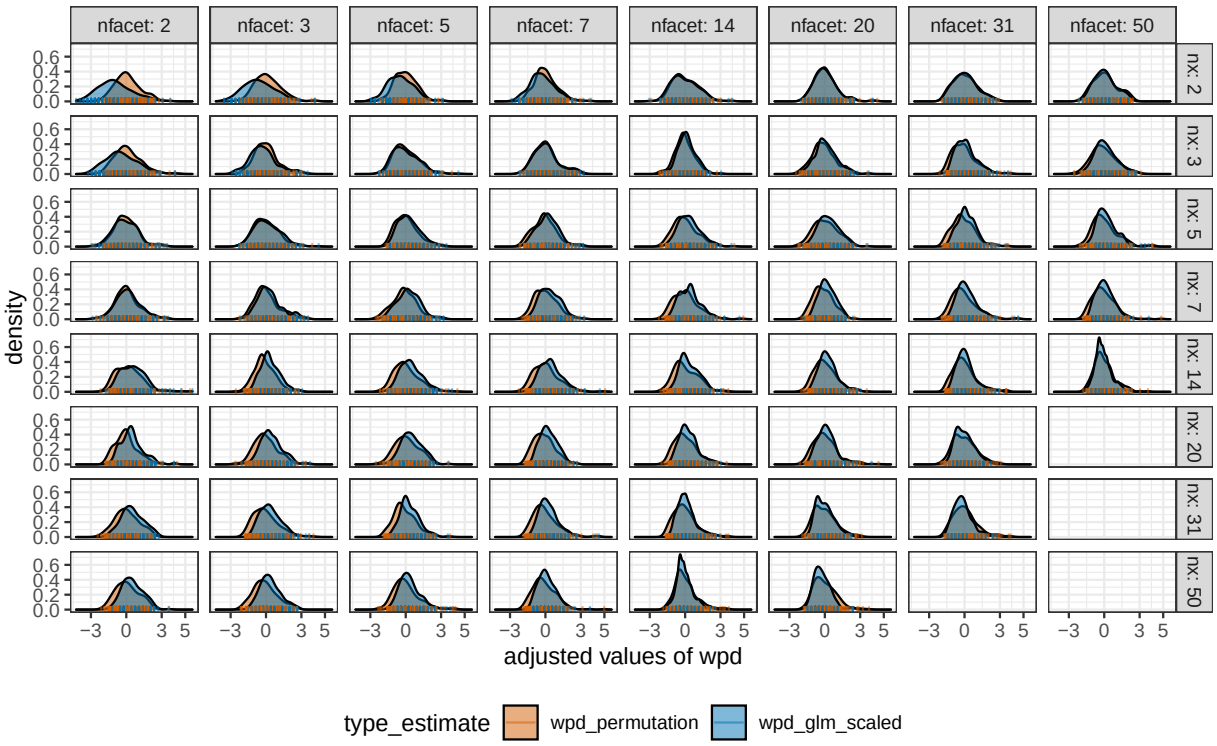

Figure 6: *The distribution of wpd_{perm} and $wpd_{glm\text{-}scaled}$ are overlaid to compare the location and scale across different n_x and n_f. wpd_{norm} takes the value of wpd_{perm} for lower levels, and $wpd_{glm\text{-}scaled}$ for higher levels to to alleviate the problem of computational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

4 Application to residential smart meter dataset

The smart meter data set for eight households in Melbourne were procured by downloading the data from the energy supplier/retailer. The data has been cleaned to form a `tsibble` (Wang, Cook & Hyndman 2020a) containing half-hourly electricity consumption from July to December 2019 for each of the households, No behavioral pattern is likely to be discerned from the time plot of energy usage over the entire period, since the plot will have too many observations squeezed in a linear representation. When we zoom into the September 2019 data in Figure 7(b), 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 daily, weekly and annual seasonal patterns. However, it is not apparent from this view if all households have those patterns, or how strong they are in each case. It is also not clear from this view if any other periodic patterns are present in any household. We start the analysis by choosing a few harmonies, and ranking them for each of these households. 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 the i^{th} household in time period t. The series $\{v_{i,1},\ldots,v_{i,T}\}$ 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. From the four linear granularities of hour, day, week, and month, we obtain $N_C = 4 \times 3/2 = 6$ cyclic granularities: "hour_day", "hour_week", "hour_month", "day_ week", "day_month" and "week_month". Further, we add cyclic granularity day-type ("wknd wday") to capture weekend and weekday behavior. Thus, seven cyclic granularities are considered to be of interest. The pairs of cyclic granularities (C_{N_C}) will have $7 \times 6 = 42$ elements. The set of possible harmonies H_{N_C} from C_{N_C} are chosen by removing clashes using procedures described in (Gupta et al. 2021). Table 3 shows 14 harmony pairs that belong to H_{N_C}.

Table 3: *Ranking of harmonies for the eight households with significance marked for different thresholds. Rankings are different and at most three harmonies are significant for any household. The number of harmonies to explore are reduced from 42 to 3.*

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

facet variable	x variable	id 1	id 2	id 3	id 4	id 5	id 6	id 7	id 8
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

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 household $i \in \{1,2,\ldots,8\}$. 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 three harmonies that are significant for any household. This is a huge reduction in the number of potential harmonies to explore.

Detecting patterns not apparent from linear display

Figure 7 helps to compare households through the heatmap (a) across harmony pairs. Here *dom*, *dow*, *wdwnd* are abbreviations for day-of-month, day-of-week, weekday/weekend and so on. The colors represent the value of *wpd*. Darker cells correspond to higher values of *wpd*. Those with $*$ correspond to *wpd* values above $wpd_{\text{threshold95}}$.

We can now see some patterns that were not discernible in Panel (b), including:

1. id 7 and 8 have the same significant harmonies despite having very different total energy usage.
2. 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.
3. There are no significant periodic patterns for id 5 when we fix the threshold to $wpd_{\text{threshold95}}$.

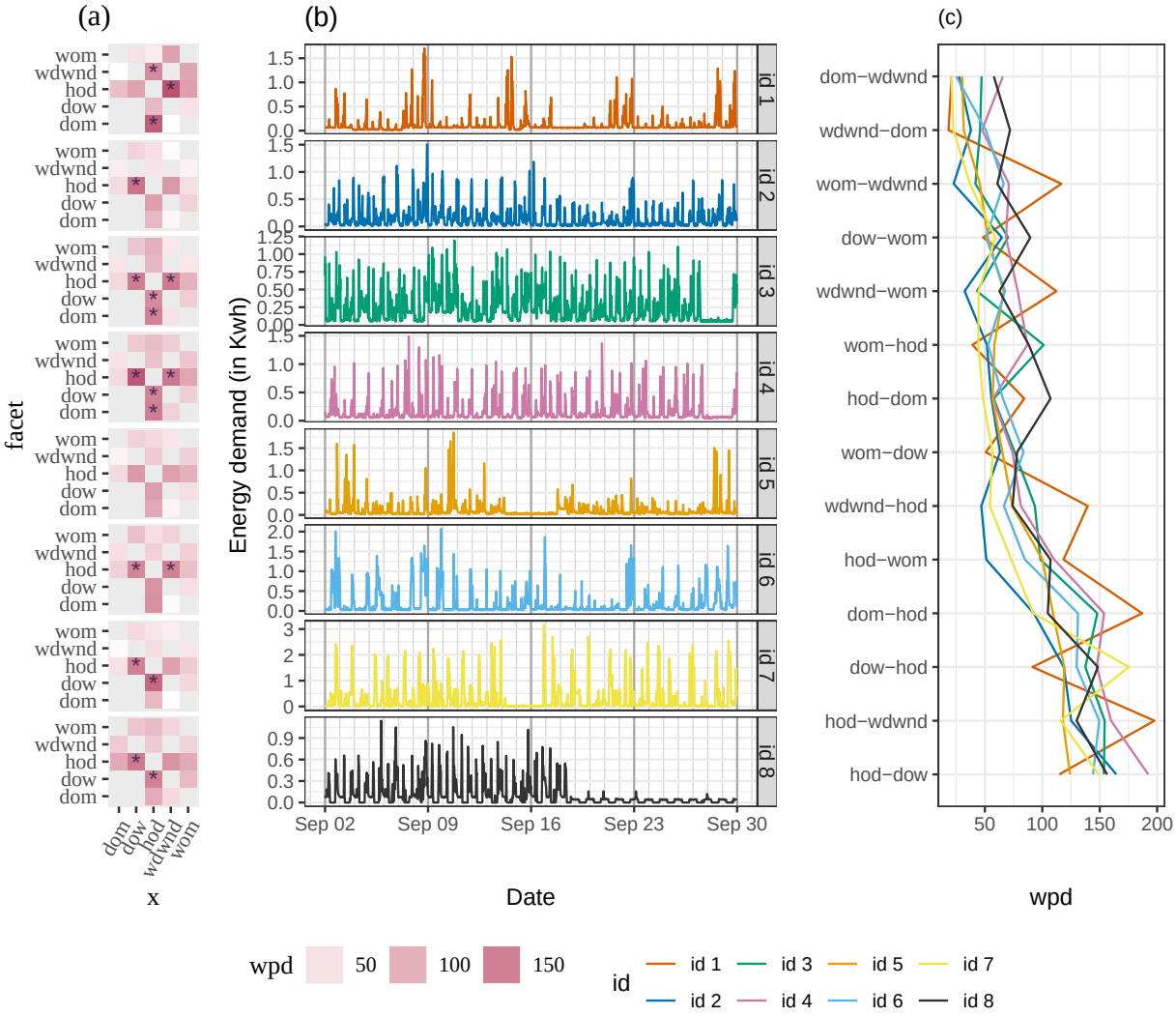

Figure 7: *An ensemble plot with a heatmap (a), line plot (b), parallel coordinate plot (c) to demonstrate energy behavior of the households in different ways. Panel (b) shows the raw demand series for September to highlight the repetitive patterns of energy demand. Panel (a) shows wpd values across harmonies where a darker color indicates a higher ranking harmony. A significant harmony is shown with an asterisk. For example, ids 7 and 8 have significant patterns across (*hod, dow*) and (*dow, hod*). Panel (c) is useful for comparing households across harmonies. For example, for the harmony (*dow-hod*), ids 1 and 7 have the least and highest wpd respectively.*

Note that the *wpd* values are computed over the entire range, but the linear display in (b) is only for September, with the major and minor x-axis corresponding to weeks and days respectively.

Comparing households and validating rank of harmonies

According to Figure 7(c), for the harmony pair (*dow-hod*), household id 7 has the greatest value of *wpd*, while id 1 has the least. From Table 3 it can be seen that the harmony pair (*dow*, *hod*) is important for id 7, however it has been labeled as an inconsequential pair for id 1. The distribution of energy demand for both of these households, with *dow* as the facet and *hod* on the x-axis, may help explain the choice. Figure 8 demonstrates that for id 7, the median (black) and quartile deviation (orange) of energy consumption fluctuates for most hours of the day and days of the week, while for id 1, daily patterns are more consistent within weekdays and weekends. As a result, for id 1, it is more appropriate to examine the distributional difference solely across (*dow*, *wdwnd*), which has been rated higher in Table 3.

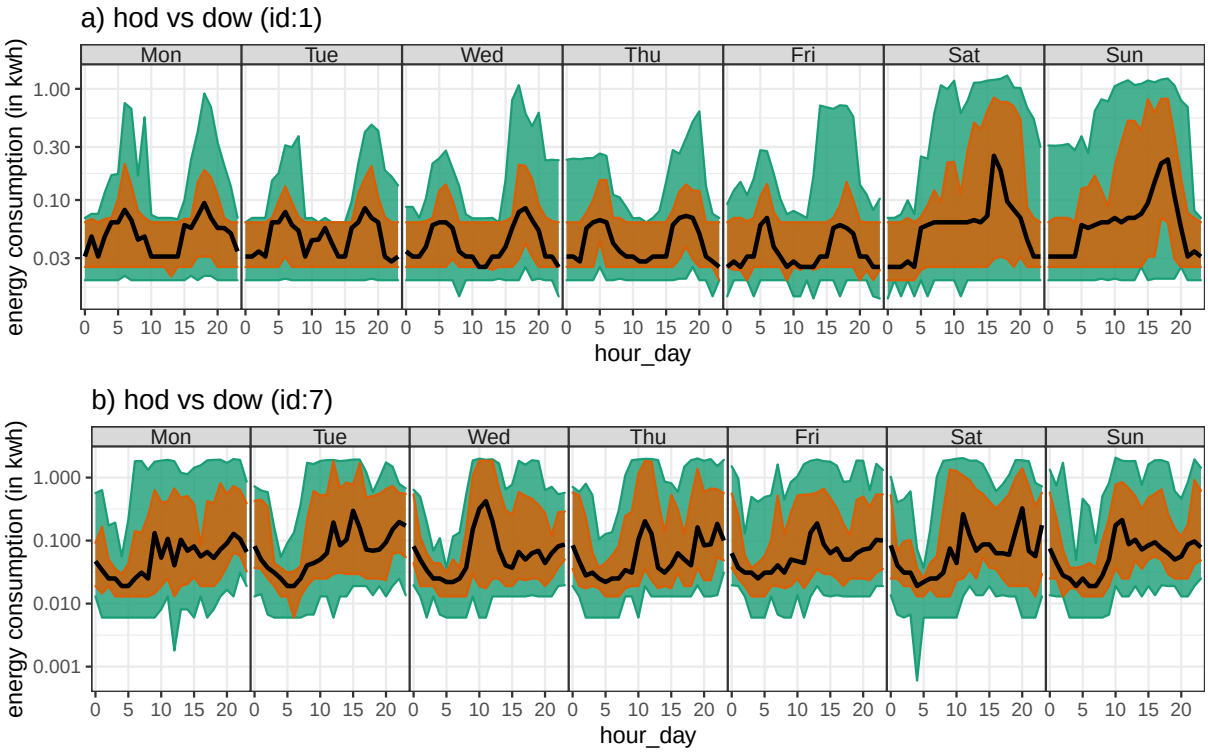

Figure 8: *Comparing distribution of energy demand shown for household id 1 (a) and 7 (b) across* hod *in x-axis and* dow *in facets through quantile area plots. The value of wpd in Table 3 suggests that the harmony pair (*dow, hod*) is significant for household id 7, but not for id 1. This implies that distributional differences are captured more by this harmony for id 7, which is apparent from this display with more fluctuations across median and 75th percentile for different hours of the day and day of week. For id 1, patterns look similar within weekdays and weekends. 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.*

5 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 the usual granularities such as 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.

Acknowledgments

The 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 2021), `knitr` (Xie 2015, 2020) and `rmarkdown` (Xie, Allaire & Grolemund 2018; Allaire et al. 2020). Graphics are produced with `ggplot2` (Wickham 2016).

6 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 package `hakear` is available on Github (https://github.com/Sayani07/hakear) to implement ideas presented in this paper.

References

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

Bettini, C, CE Dyreson, WS Evans, RT Snodgrass & XS Wang (1998). "A glossary of time granularity concepts". In: *Temporal Databases: Research and Practice*. Ed. by O Etzion, S Jajodia & S Sripada. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.406–413.

Buja, A, D Cook, H Hofmann, M Lawrence, EK Lee, DF Swayne & H Wickham (2009). Statistical Inference for Exploratory Data Analysis and Model Diagnostics. *Royal Society Philosophical Transactions A* **367**(1906), 4361–4383.

Dang, TN & L Wilkinson (Mar. 2014). ScagExplorer: Exploring Scatterplots by Their Scagnostics. In: *2014 IEEE Pacific Visualization Symposium*, pp.73–80.

Edgington, E & P Onghena (2007). *Randomization tests*. CRC press.

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

Gupta, S, RJ Hyndman, D Cook & A Unwin (2021). Visualizing probability distributions across bivariate cyclic temporal granularities. *Journal of Computational & Graphical Statistics*. to appear.

Hyndman, RJ & Y Fan (Nov. 1996). Sample Quantiles in Statistical Packages. *Am. Stat.* **50**(4), 361–365.

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

Kullback, S & RA Leibler (1951). On Information and Sufficiency. *Ann. Math. Stat.* **22**(1), 79–86.

Lin, J (Jan. 1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**(1), 145–151.

Majumder, M, H Hofmann & D Cook (Sept. 2013). Validation of Visual Statistical Inference, Applied to Linear Models. *J. Am. Stat. Assoc.* **108**(503), 942–956.

Menéndez, ML, JA Pardo, L Pardo & MC Pardo (Mar. 1997). The Jensen-Shannon divergence. *J. Franklin Inst.* **334**(2), 307–318.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

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

Wang, E, D Cook & RJ Hyndman (2020a). A new tidy data structure to support exploration and modeling of temporal data. *Journal of Computational & Graphical Statistics* **29**(3), 466–478.

Wang, E, D Cook & RJ Hyndman (2020b). Calendar-based graphics for visualizing people's daily schedules. *Journal of Computational & Graphical Statistics* **29**(3), 490–502.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wilkinson, L, A Anand & R Grossman (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. Boca Raton, Florida: Chapman and Hall/CRC. https://yihui.org/knitr/.

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

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