Median Maximum Pairwise Distance

Contents

1 Introduction

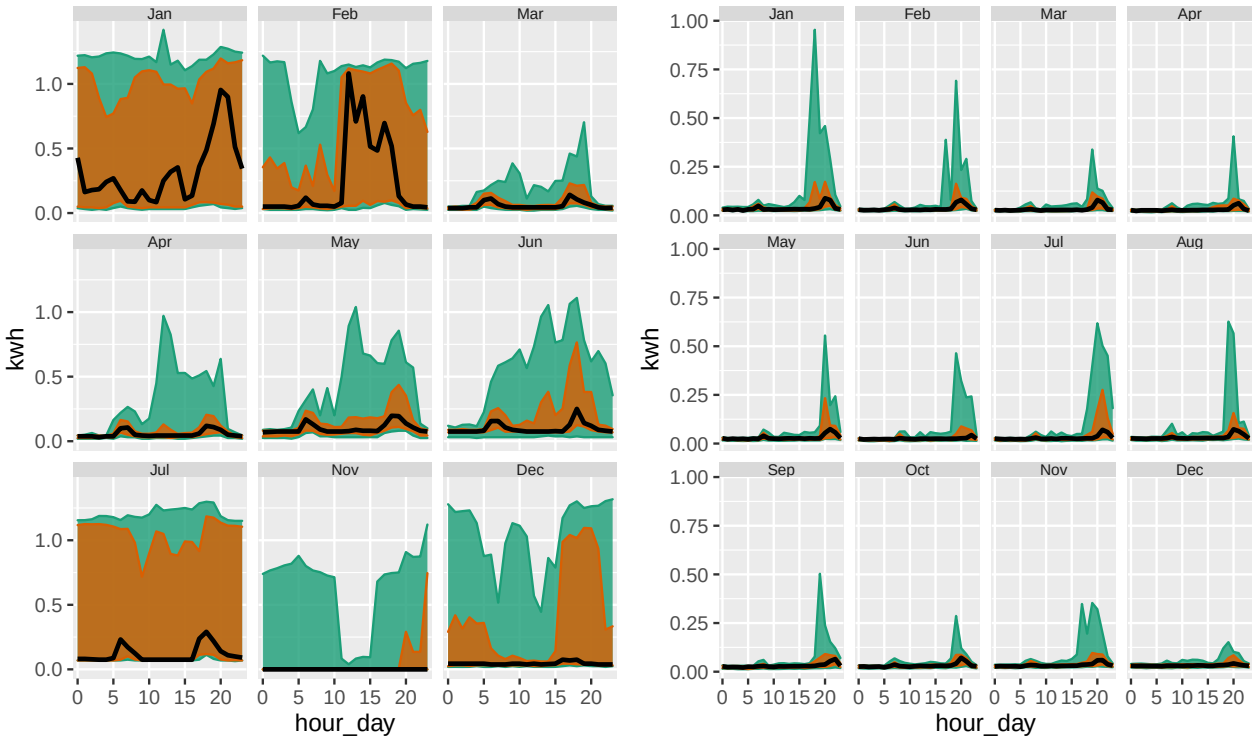
Take the example of the calendar display of electricity smart meter data used in Wang, Cook, and Hyndman (2020) for four households in Melbourne, Australia. The authors show how hour-of-the-day interact with weekday and weekends and then move on to use calendar display to show daily schedules. The calendar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

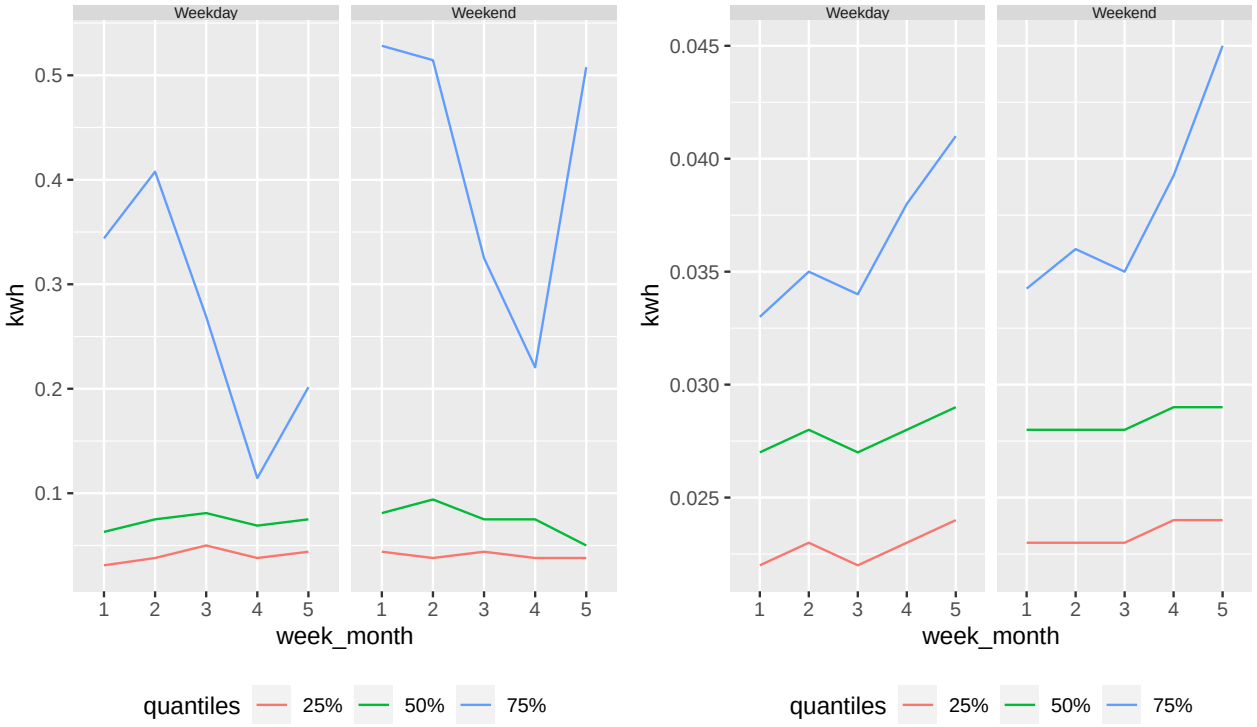
useful to make the transition from all possible ways to only ways that could potentially be important given a situation.

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

This is similar to Scagnostics (Scatterplot Diagnostics)(**???**), which is used to discern meaningful patterns in large collections of scatterplots. Given a set of v variables, there are $v(v-1)/2$ pairs of variables, and thus the same number of possible pairwise scatterplots. Therefore for even small v, the number of scatterplots can be large, and scatterplot matrices (SPLOMs) could easily run out of pixels when presenting high-dimensional data. (**???**), (**???**) provides a solution to this, where few characterizations help us to locate anomalies for further analysis or search for similar distributions in a "large" SPLOM with more than a hundred dimensions.

The paper (gravitas) 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 for which variation within different categories is significant. Also, ranking the remaining harmonies based on how well they capture the variation in the measured variable could be potentially useful.





Here, we aim to build a new measure to follow through these two main objectives:

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

2 Median Maximum Pairwise Distances (MMPD)

2.1 Principle

The principle employed for building a new metric is explained through a simple example explained in Figure 1. Each of these figures have the same panel design with 2 x-axis categories and 4 facet levels. Figure 1a has all x categories drawn from N(5, 10) distribution for each facet. It is not an interesting display particularly, as distributions do not vary across x-axis or facet categories. Figure 1b has x categories drawn from the same distribution within a facet and different for different facet categories. Figure 1b exhibits an exact opposite situation where distribution between the x-axis categories within each facet is different but they are same across facets. Figure 1d takes a step further by varying the distribution across both facet and x-axis categories. If we are asked to rank the displays in order of importance from minimum to maximum, we might order it as a, b, c and then d. It might be argued that it is not clear if b should precede or succeed c. Gestalt theory suggests that when items are placed in close proximity, people assume that they are in the same group because they are close to one another and apart from other groups. Hence, displays that capture more variation within different categories in the same group would be important to bring out different patterns of the data. With this principle, display b could be considered less informative as compared to display c.

With reference to the graphical design in **??**, therefore the idea would be to rate a harmony pair higher if the variation between different levels of the x-axis variable is higher on an average across all levels of the facet variables. Thus the metric could be obtained by computing maximum pairwise distances between distributions of the continuous random variable across x-axis categories for all facets and then taking the median of those maximum pairwise distances across facets. This would help capture the average maximum difference in distribution of the measurement variable explained by the two cyclic granularities together. We

call this metric MMPD which stands for Median Maximum Pairwise Distances. In the next section we shall see how we go about computing this measure.

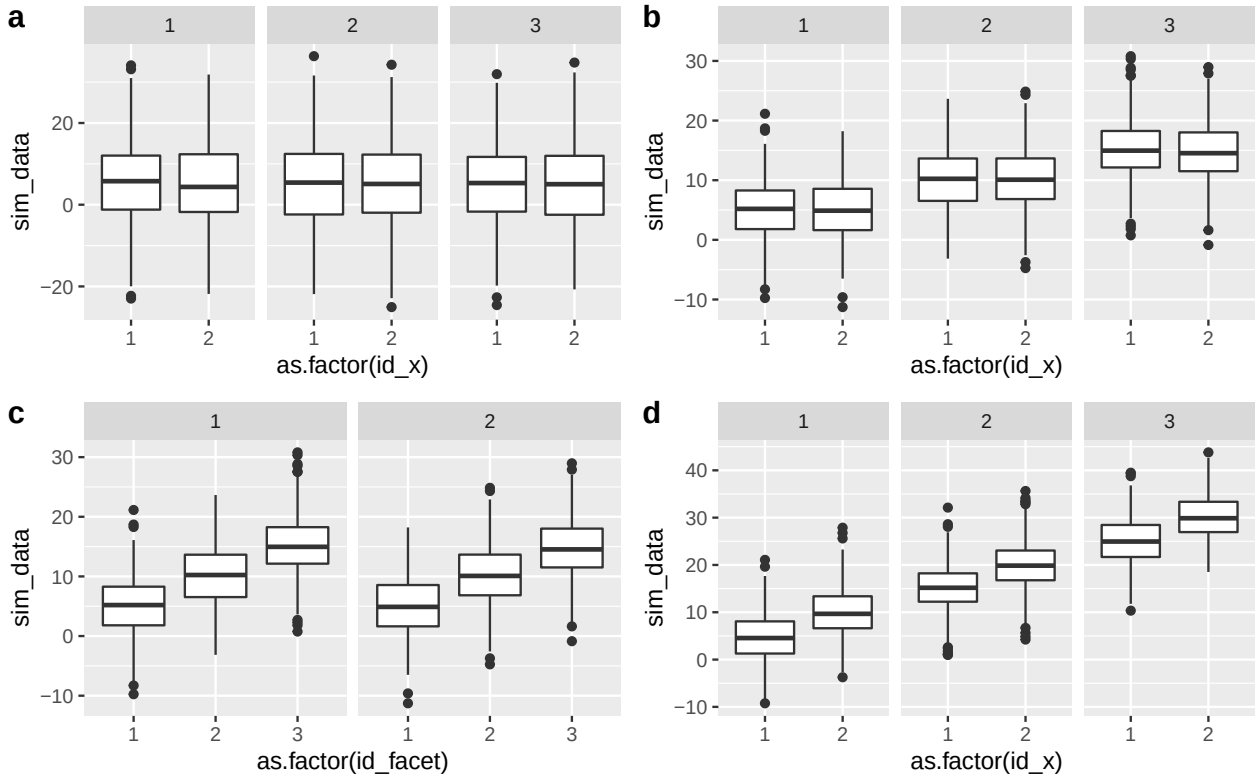


Figure 1: A graphical display with two categories mapped to x-axis and 4 categories mapped to facets with the distribution of a continuous random variable plotted on the y-axis. Display a is not interesting as the distribution of the continuous rv does not depend across x-axis or facet categories. Display b and c are more interesting than a since there is a change in distribution either across facets(b) or x-axis(a). Display d is most interesting as distribution of the rv changes across both facet and x-axis variable.

2.2 Computation

2.2.1 Distance between distributions

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

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

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

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

2.2.2 Distribution of Jensen-Shannon distances

Jensen-Shannon distances (JSD) are distributed as chi-squared with m df where we discretize the continuous distribution with m discrete values. Taking sample percentiles to approximate the integral would mean taking $m = 99$. With large m, chi-squared is asymptotically normal by the CLT. Thus, by CLT, ${\chi^2}_m \tilde{N}(m, 2m)$, which would depend on the number of discretization used to approximate the continuous distribution. Then $b_n = 1 - 1/n$ quantile of the normal distribution and $a_n = 1/[n * \phi(b_n)]$ where ϕ is the normal density function. n is the number of pairwise comparisons being made.

2.2.3 Normalize distances

Suppose that X_1, X_2 , ... , X_n are i.i.d. random variables with expected values $E(X_i) = \mu < \infty$ and variance $Var(X_i) = \sigma^2 < \infty$. Let $Y = max(X_1, X_2, \ldots, X_n)$.

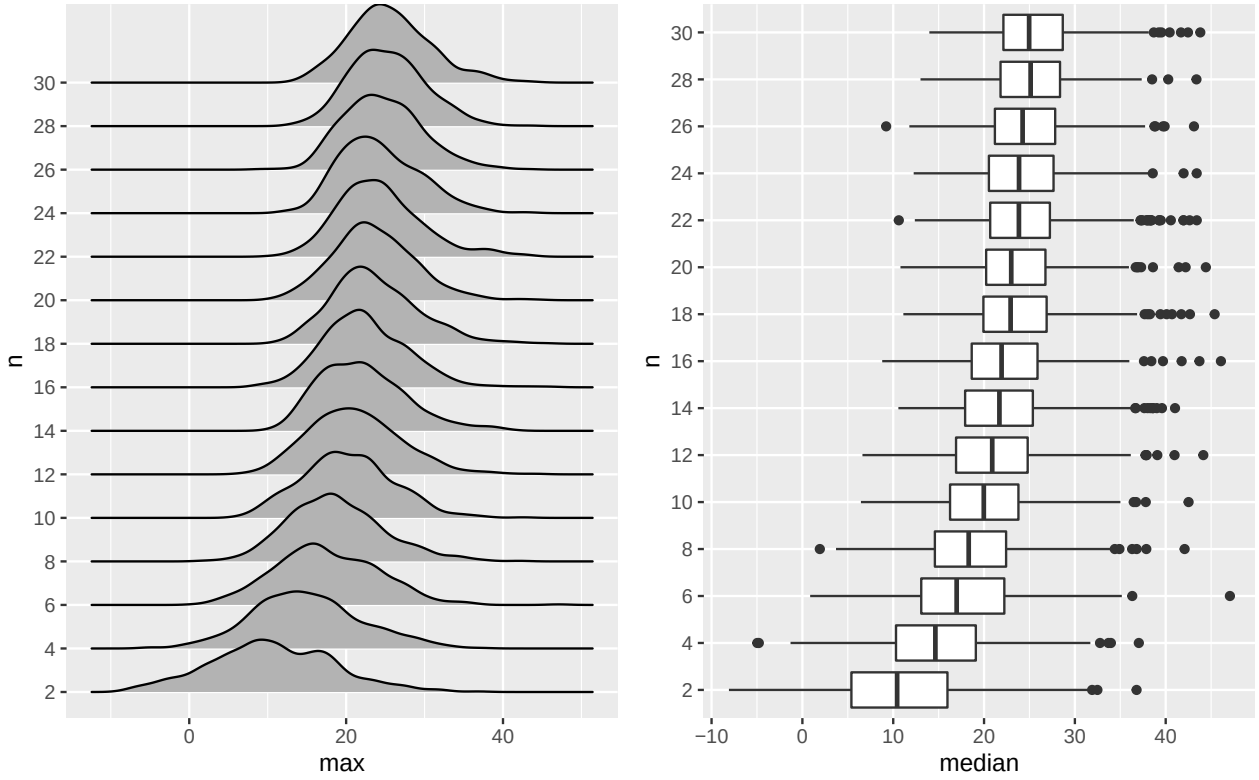
Let $F_X(x)$ be the common distribution of the variables X_i and let $F_Y(y)$ be the corresponding distribution of Y. $F_Y(y)$ could be obtained from $F_X(x)$ simply by using: $F_Y(y) = P[(X_1 \leq y) \cap (X_2 \leq y) \cap ... \cap (X_n \leq y)] = F_X(y)^n$. For large n, the distribution of Y approaches a standard shape, which does not depend on F_X. But what about the case when n is not large enough? The distribution of maximum in that case will indeed depend on n and the underlying distribution of X. If $F_X(x)$ is the CDF of X, then $F_Y(y) = F_x(y)^n$. Suppose Φ nd ϕ are the cdf and pdf of a standard normal distribution, then $f_Y(y) = n\Phi(y)^{n-1}\phi(y)$, which depends on n. Hence, we are trying to normalise for n. Also, it depends on the underlying distribution of X, which we have assumed as normal in our case. As n grows, we can see the right tail growing, which implies that the probability that we will get a higher maximum is more. Now, for large n, we used EVT to normalise for n, that is, we brought them to the same scale without distorting the range of the distribution. But in our case, we will mostly have small n. It is important to ensure that they have the same mean and variation, for being able to compare the maximum value across n. We observe from the following graphs that our normalisation works after $n = 6$, after which the difference in mean and standard deviation flattens out a lot.

Maximum pairwise distances are not robust to the number of levels and is higher for harmonies with higher levels. Thus these maximum pairwise distances need to be normalized for different harmonies in a way that eliminates the effect of different levels. The Fisher–Tippett–Gnedenko theorem in the field of Extreme Value Theory states that the maximum of a sample of iid random variables after proper re-normalization can converge in distribution to only one of Weibull, Gumbel or Freschet distribution, independent of the underlying data or process.

More formally, $d_1, d_2 \ldots, d_n$ be a sequence of independent and identically-distributed pairwise distances and $M_n = \max\{d_1, \ldots, d_n\}$. Then Fisher–Tippett–Gnedenko theorem (Haan and Ferreira 2007) suggests that if a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ and $\lim_{m\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$, where F is a non-degenerate distribution function, then the limit distribution F belongs to either the Gumbel, Fréchet or Weibull family. The normalizing constants (a_n, b_n) vary depending on the underlying distribution of the pairwise distances. Hence to normalize appropriately, it is important to assume a distribution of these distances.

2.2.4 why normalize

2.2.5 Mean and standard deviation of the distribution of maximum



2.2.6 Distribution of distances

2.2.7 Theoretical evidence

JS distances are distributed as chi-squared with m df where we discretize the continuous distribution with m discrete values. Taking sample percentiles to approximate the integral would mean taking $m = 99$. With large m, chi-squared is asymptotically normal by the CLT. Thus, by CLT, ${\chi^2}_m \tilde{} N(m, 2m)$, which would depend on the number of discretization used to approximate the continuous distribution. Then $b_n = 1 - 1/n$ quantile of the normal distribution and $a_n = 1/[n * \phi(b_n)]$ where ϕ is the normal density function. n is the number of pairwise comparisons being made.

2.2.8 Empirical evidence

Distribution of JS distances is assumed to be normal but the mean and variance are estimated from the sample, rather than deducing it from the number of discretization used to approximate the continuous distribution. We look at different scenarios, where observations are collected from Normal, Exponential, Chi-squared and Gumbel distribution and found the distribution of JS distances are similar, irrespective of which distribution they are drawn from.

2.2.8.1 Initial distribution of observed variables shown in plot title

2.3 Algorithm

The algorithm employed for computing MMPD is summarized as follows:

- **Input:** Data corresponding to all harmony pairs, i.e., data sets of the form (C_i, C_j, v) $\forall i, j \in N_C$
- **Output:** MMPD (Median Maximum Pairwise Distances) measuring the average variation across different levels of C_i and C_j $\forall i, j \in N_C$

1. Fix harmony pair (C_i, C_j).
2. Fix k. Then there are L groups corresponding to level A_k of C_i.
3. Compute $m = \binom{L}{2}$ pairwise distances between distributions of L unordered levels and $m = L - 1$ pairwise distances for L ordered categories.
4. Identify maximum within the m computed distances.
5. Compute normalized maximum distance (NM) using appropriate norming constants.
6. Use Steps 1-5 to compute normalized maximum distance for $\forall k \in \{1, 2, \ldots, K\}$.
7. Compute MMPD = median $(NM_1, NM_2, \ldots, NM_K)/\log(K)$.
8. Repeat Steps 1 to 7 for all harmony pairs.

2.3.1 Bounds

This is not correct because MMPD should be median of standardized Gumbel distribution. So no bound?

By Lin (1991),

$$0 \leq JSD(P||Q) \leq ln(2)$$

.

Thus,

$$0 \leq MMPD \leq \frac{ln(2)}{ln(k)}$$

. Now, by assumption $k \geq 2$ and hence,

$$\frac{ln(2)}{ln(k)} : \begin{cases} 1 & if\ k = 2 \\ < 1 & if\ k \geq 2 \end{cases}$$

Thus,

$$0 \leq MMPD \leq\ 1$$

3 The statistical test

3.1 Definition

3.1.1 Algorithm for computation for all harmony pairs

Assumption: random permutation without considering ordering (global)

1. Given the data; $\{v_t : t = 0, 1, 2, \ldots, T-1\}$, the MMPD is computed and is represented by $MMPD_{obs}$.

2. From the original sequence a random permutation is obtained: $\{v_t^* : t = 0, 1, 2, \ldots, T-1\}$.
3. MMPD is computed for all random permutation of the data and is represented by $MMPD_{sample}$.
4. Steps (2) and (3) are repeated a large number of times M (e.g. 1000).
5. For each permutation, one $MMPD_{sample}$ value is obtained.
6. 95^{th} percentile of this $MMPD_{sample}$ distribution is computed and stored in $MMPD_{threshold}$.
7. If $MMPD_{obs} > MMPD_{threshold}$, harmony pairs are accepted. Only one threshold for all harmony pairs.

Pros: Considering thresholds global for all harmony pairs would imply less computation time.

Cons: Only one threshold for all harmony pairs means we are assuming distribution of all harmonies pairs are similar, which might not be the case.But nevertheless, it is a good benchmark.

3.2 Null distribution of MMPD

3.2.1 Normalised maximum distances follow standard Gumbel distribution

3.2.2 Limiting distribution of median of normalised maximum distances is normal

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

3.2.3 Confidence interval of test statistic

3.3 Simulation design

Behavior of the statistic - control simulation

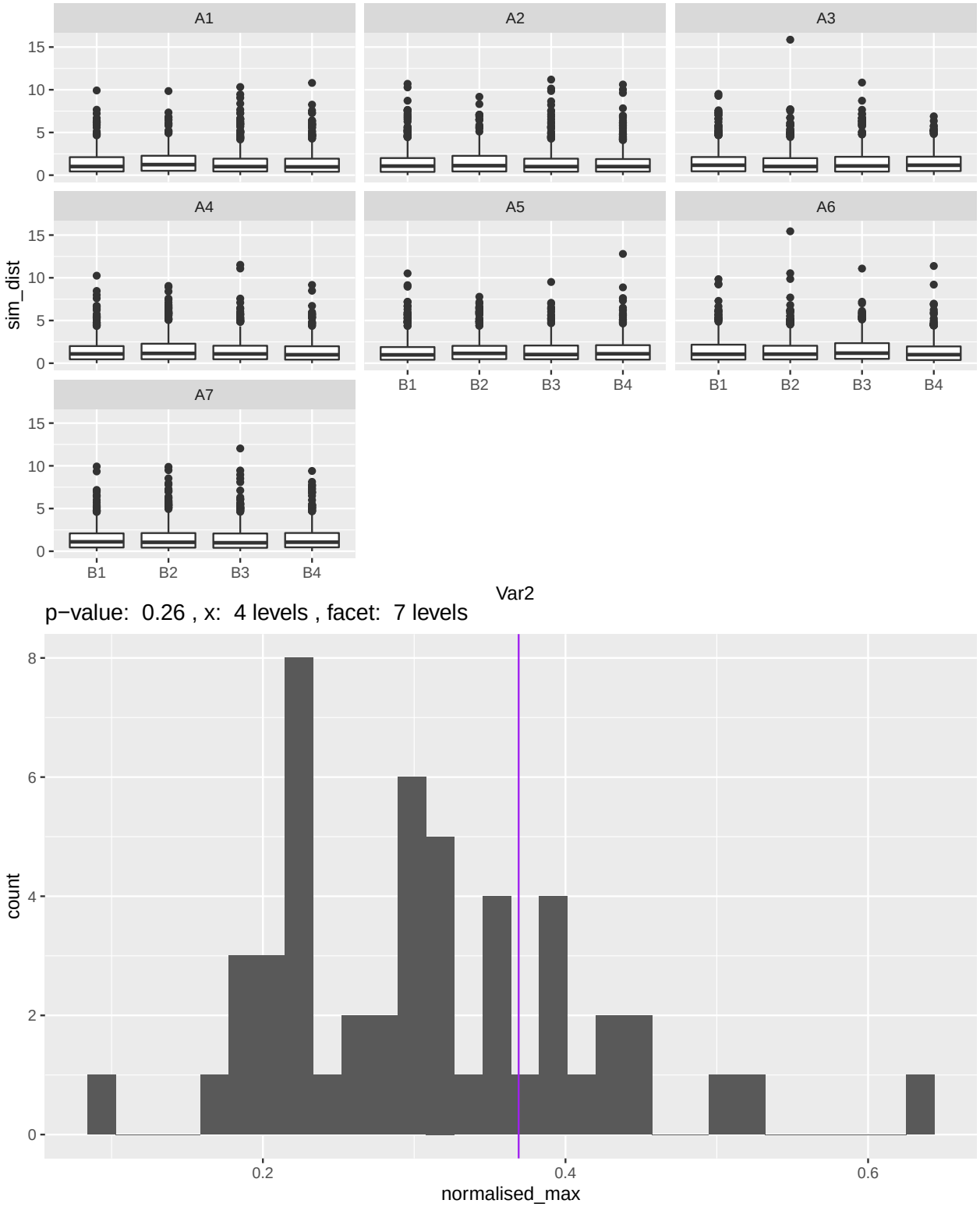
- To check if different distributions impact (simulate with different distributions but same for all levels)
- To check if x-levels and facets are normalized (simulate with different distributions) (simulate with different x and facet levels)

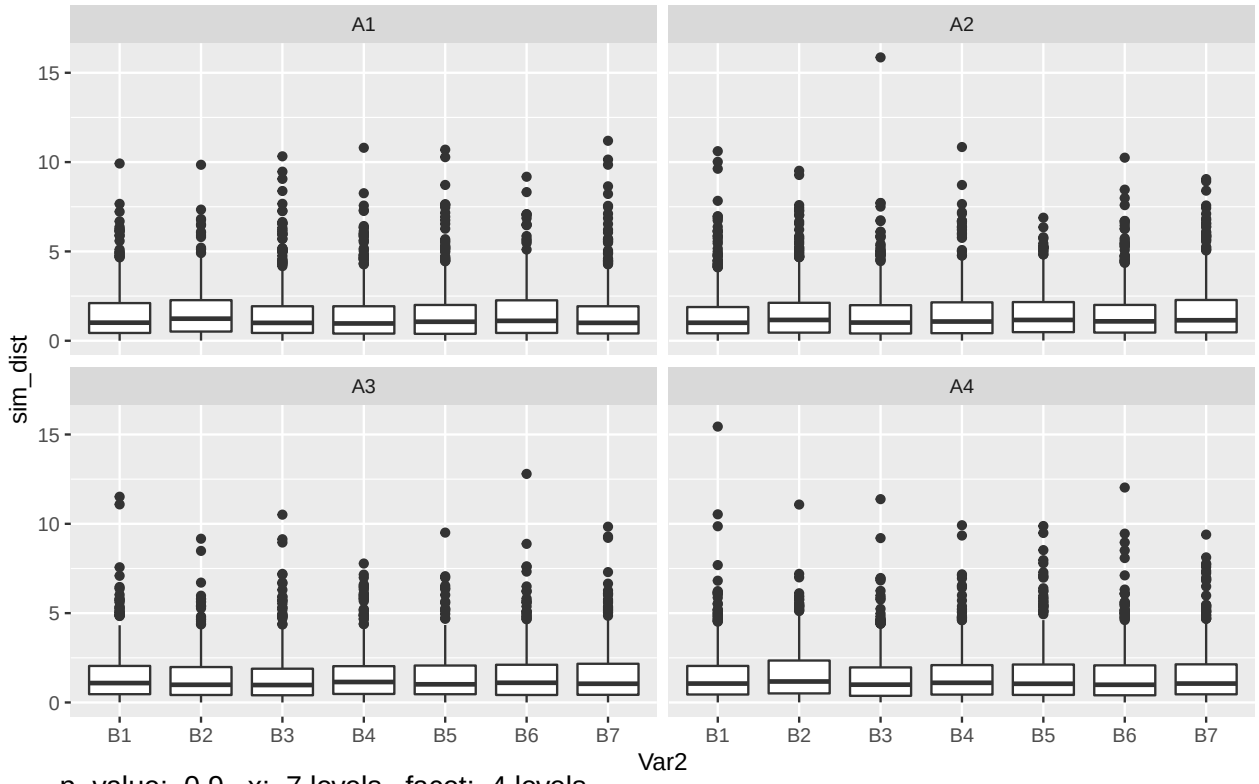
3.4 Size and power

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

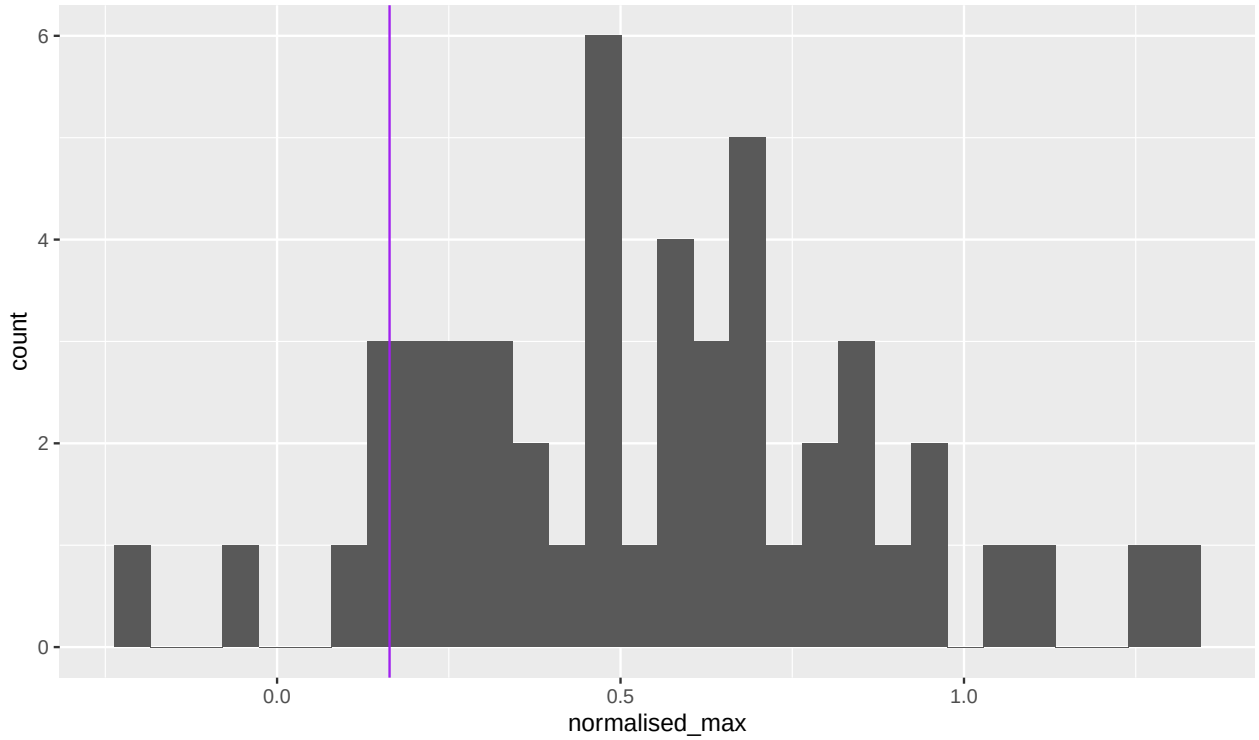
3.4.1 Size: Simulated same distribution for all combinations of categories for all harmony pairs.

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

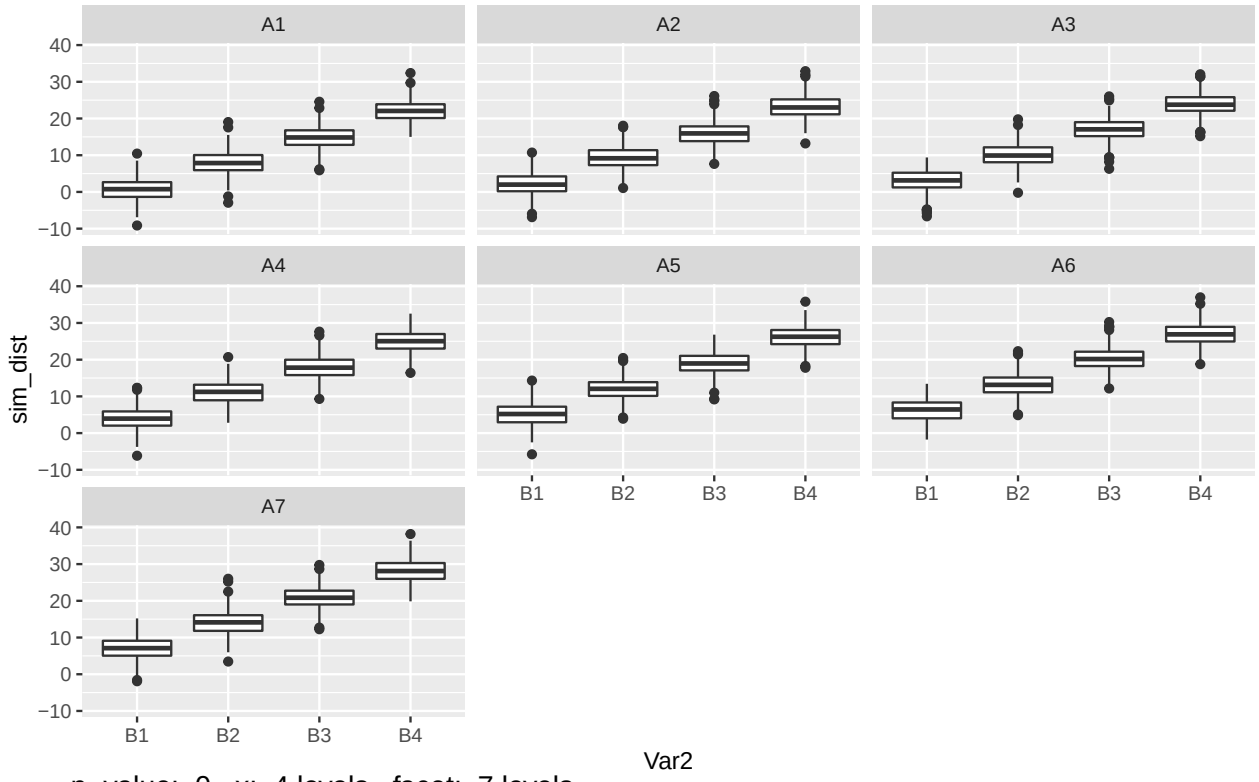


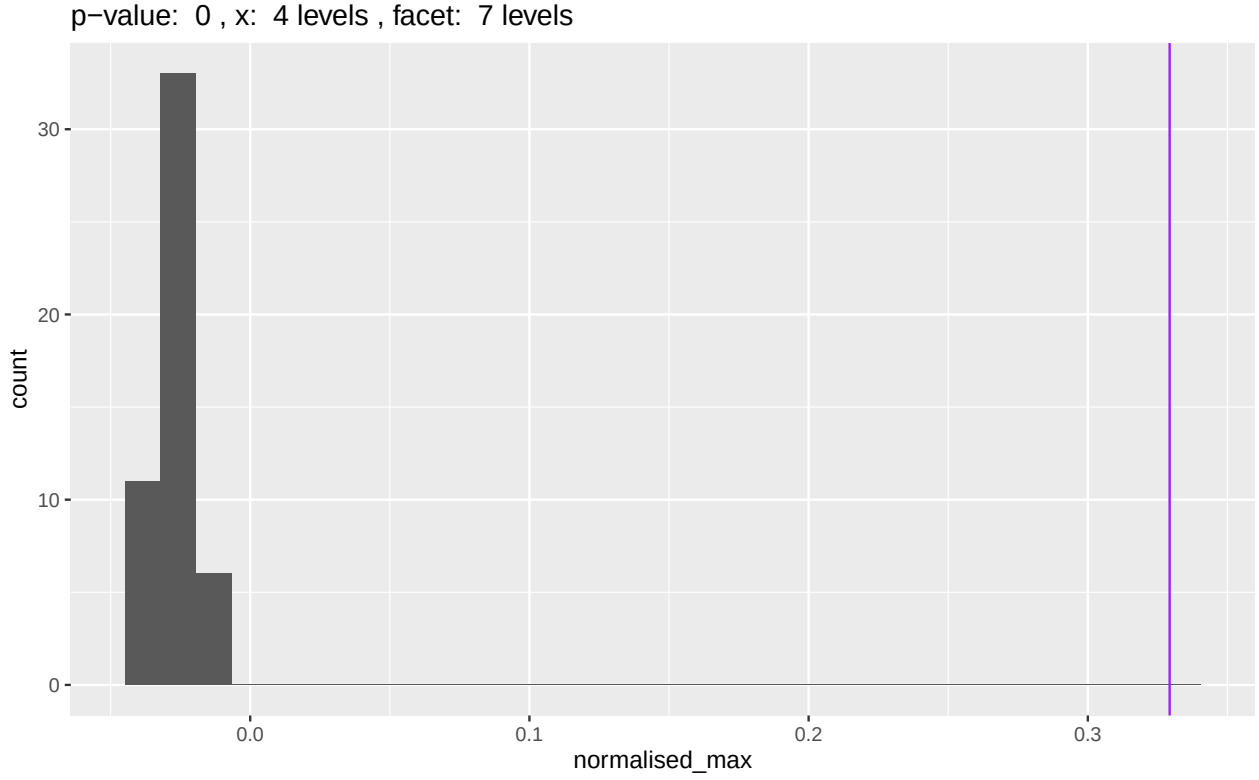


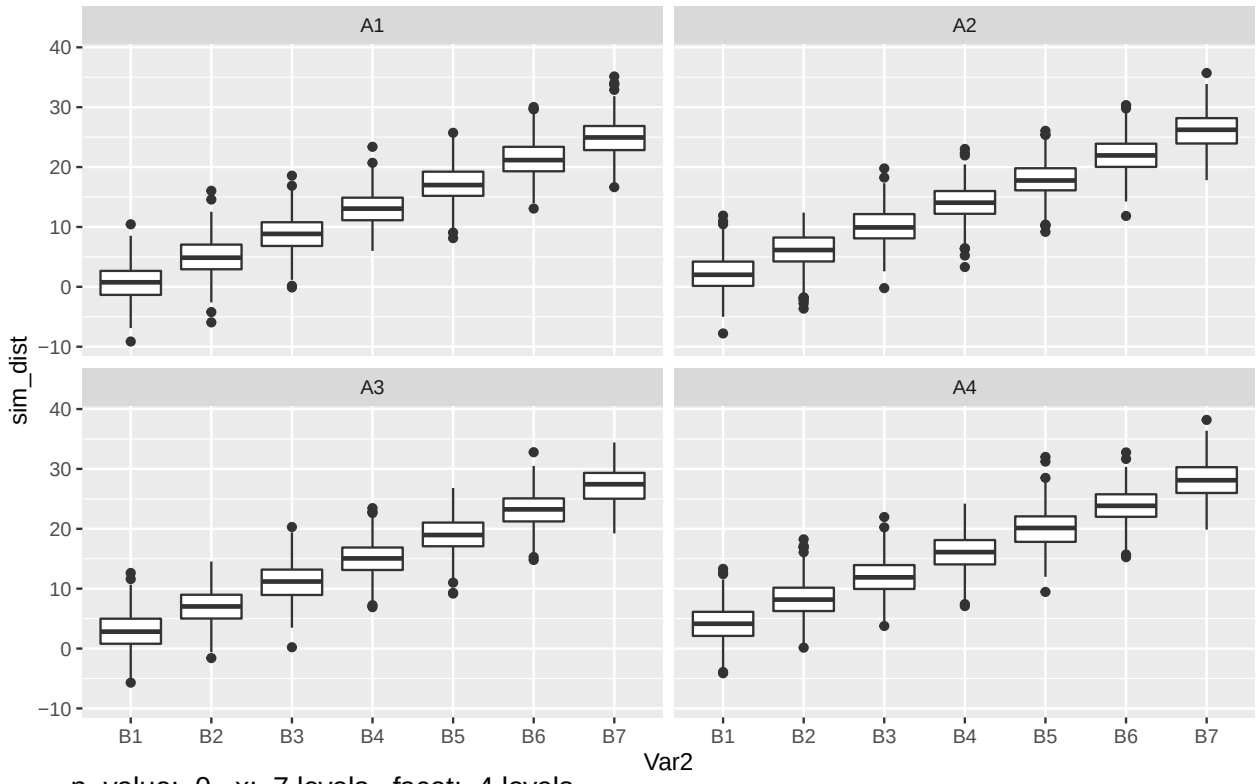
p−value: 0.9 , x: 7 levels , facet: 4 levels

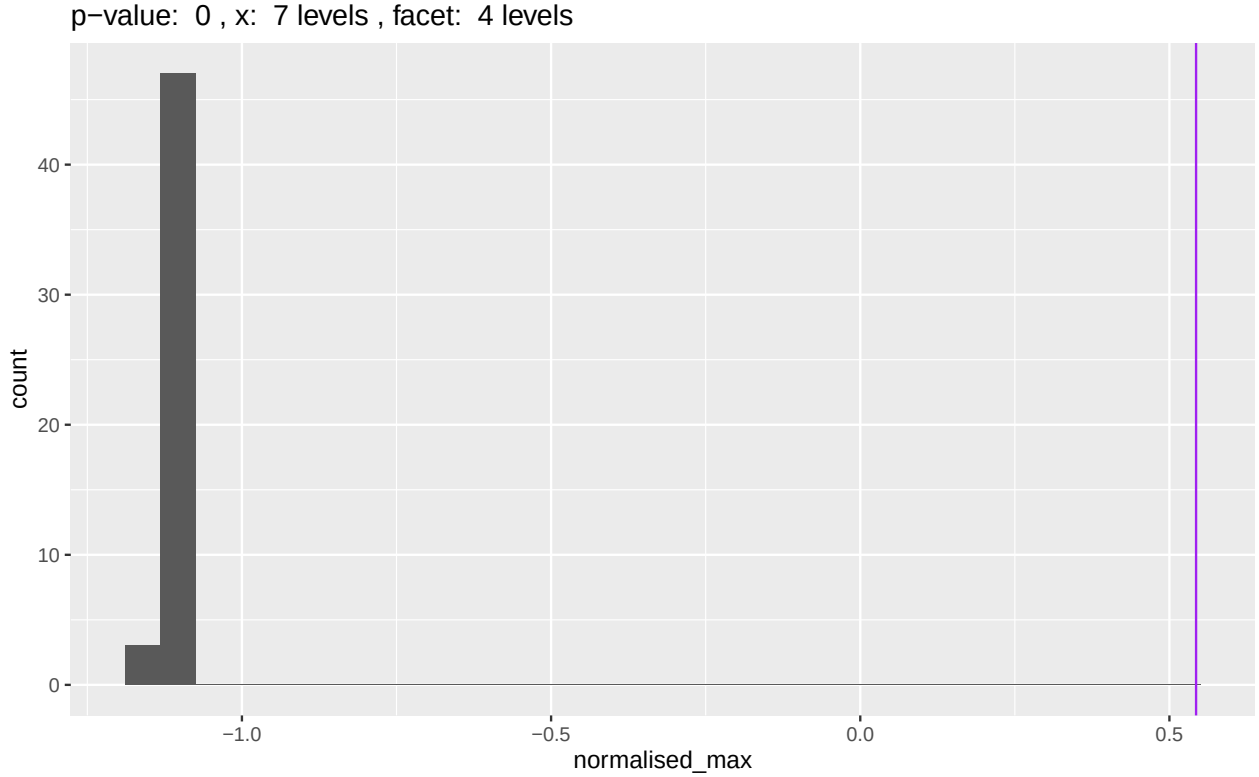


3.4.2 Power: Simulated same distribution for all combinations of categories for all harmony pairs.









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

3.4.3 Scenario 2: Simulated different distributions for all combinations of categories for harmony pairs for few levels.

Conclusion: The test select the harmony pair for which distribution of x-axis categories are significantly different

3.4.4 Scenario 3: Simulated different distributions for all combinations of categories for all harmony pairs with many levels.

Conclusion: The test indicates that both harmony pairs do not have significant variation.

3.4.5 Scenario 4: Simulated different distributions for all combinations of categories for all harmony pairs with many levels - very different distribution across x-axis

Conclusion: The test indicates that only the first harmony pair has significant variation.

3.4.6 Scenario 5: Simulated different distributions for all combinations of categories for all harmony pairs with many levels - very different distribution across facets

–>

–>

4 Applications

4.1 Smart meter data of Australia

Smart meters provide large quantities of measurements on energy usage for households across Australia. One of the customer trials (Department of the Environment and Energy 2018) conducted as part of the Smart Grid Smart City project in Newcastle, New South Wales and some parts of Sydney provides customer wise data on energy consumption for every half hour from February 2012 to March 2014. The idea here is to show how to visualize the distribution of the energy consumption across different cyclic granularities in a systematic way to identify different behavioral patterns.

4.1.1 Cyclic granularities search and computation:

The tsibble object `smart_meter10` from R package `gravitas` (Gupta et al. 2019) consisting of `reading_datetime`, `customer_id` and `general_supply_kwh` denoting the index, key and measured variable of the tsibble is used to facilitate the systematic exploration. While trying to explore the energy behavior of these customers systematically across cyclic time granularities, the first thing to consider is which cyclic time granularities we can look at exhaustively. Let us consider conventional time deconstructions for a Gregorian calendar (second, minute, half-hour, hour, day, week, month, year). Since the interval of this tsibble is 30 minutes, the temporal granularities may range from half-hour to year. Considering 6 linear granularities half-hour, hour, day, week, month and year in the hierarchy table, $N_C = (6 * 5/2) = 15$. If N_C seem too large, the smallest and largest linear granularities could be considered to be removed from the hierarchy table. We remove half-year and year to have $N_C = (4 * 3/2) = 6$ and 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. Now that we have a list of cyclic granularities to look at, we should be able to compute the multiple-order-up granularities using Section **??**.

4.1.2 Screening and visualizing harmonies

From the search list, $N_C = 7$ cyclic granularities are chosen for which we would like to derive insights of energy behavior. Recalling the data structure $<C_i, C_j,$ `general_supply_kwh`$>$ for exploration $\forall i, j \in \{1, 2, \ldots, 7\}$, each of these 7 cyclic granularities can either be mapped to x-axis or to facet. Choosing 2 of the possible 7 granularities, which is equivalent to having ${}^7P_2 = 42$ candidates for visualization. Fortunately, harmonies can be identified among those 42 possibilities to narrow the search. **??** shows 16 harmony pairs after removing clashes and any cyclic granularities with levels more than 31, as effective exploration becomes difficult with many levels (Section **??**). The MMPD is also shown along with indicator (*) only when variation of measured variable across the harmony pair significant. Starting from 42 possible pairs of cyclic granularities to visualize, we are finally left with only 6, which is a very sizable number of displays for exploration.

Few harmony pairs are displayed in Figure 2 to illustrate the siginificance of MMPD, threshold and the impact of different distribution plots and reverse mapping. For each of Figure 2 (b) and (c), C_i is the circular granularity day-type (weekday/weekend) and C_j is hour of the day. The geometry used for displaying the distribution is chosen as area-quantiles and violins in Figure 2 (b and c respectively). Figure 2 (a) displays reverse mapping of C_i and C_j with C_i denoting hour of the day and C_j denoting day-type with distribution geometrically displayed as boxplots.

In Figure 2 (b), the black line is the median, whereas the purple band covers 25th to 75th percentile, the orange band covers 10th to 90th percentile and the green band covers 1st to 99th percentile. The first facet represents the weekday behavior while the second one displays the weekend behavior and energy consumption across each hours of the day is shown inside each facet. The energy consumption is extremely (positive- or right-) skewed with the 1st, 10th and 25th percentile lying relatively close whereas 75th, 90th and 99th lying further away from each other. This is common across both weekdays and weekends. For the first few hours on weekdays, median energy consumption starts and continues to be higher for longer as compared to weekends.

Consider looking at violin plots instead of quantile plots to look at the same data in Figure 2(c). There is additional information that we can derive looking at the distribution. There is bimodality in the early hours of the day, implying both low and high energy consumption is probable in the early hours of the day both for weekdays and weekends. If we visualize the same data with reverse mapping of the cyclic granularities, then the natural tendency would be to compare weekend and weekday behavior within each hour and not across hours. For example in Figure 2(a), it can be seen that median energy consumption for the early morning hours is extremely high for weekdays compared to weekends. Also, outliers are more prominent in the latter part of the day. All of these indicate that looking at different distribution geometry or changing the mapping might shed lights on different aspect of the energy behavior for the same sample population.

If the data for all keys are visualized together, it might lead to Simpson's paradox, which occurs when one observation shows a particular behavior, but this behavior paradoxically becomes obscured by aggregation. For example in a particular neighborhood one household may have the least daily power consumption for a full week, yet still not be the household with the minimum weekly power consumption. This is an intuitive possibility, because heterogeneous `customer_id`'s with very different occupation or demographics will tend to have very different energy behavior and combining them together will somehow weaken any typical or extreme behavior. A strategy for analyzing multiple keys together could be to first group them basis time series or demographic features and then look at their energy behavior. This is beyond the scope of the current work.

This case study shows systematic exploration of energy behavior for a household to gain exhaustive insights on periodic behavior of the households.

4.2 T20 cricket data of Indian Premiere League

The method is not only restricted to temporal data, and can be generalized to many hierarchical granularities (with continuous and uni-directional nature). We illustrate this with an application to the sport cricket. Although there is no conventional time component in cricket, each ball can be thought to represent an

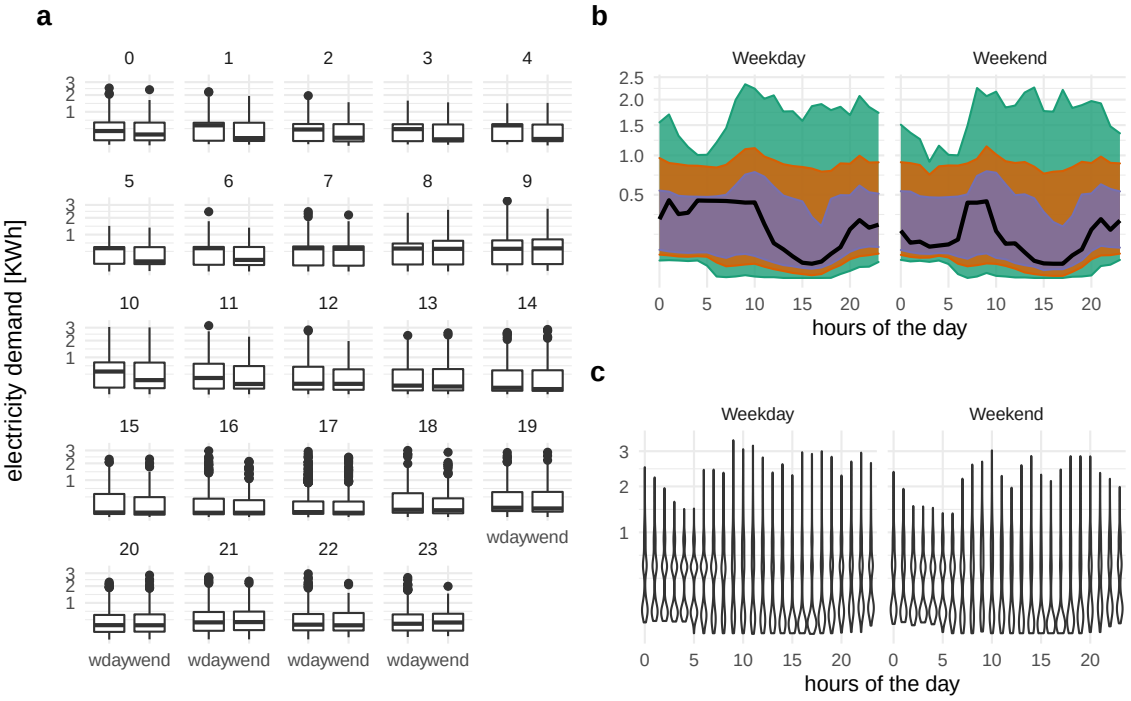


Figure 2: Energy consumption of a single customer shown with different distribution displays, and granularity arrangements. Two granularities are used: hour of the day (I) and weekday/weekend (II). Plot (a) shows granularity I facetted by granularity II, and plots (b), (c) shows the converse mapping. Plot (a) makes a comparison of usage by workday within each hour of the day using side-by-side boxplots. Generally, on a work day there is more consumption early in the day. Plots (b) and (c) examine the temporal trend of consumption over the course of a day, separately for the type of day. Plot (b) uses an area quantile to put the emphasis on the time series, for example, the median consumption over time shows prolonged usage in the morning on weekdays. Plot (c) uses a violin plot to place emphasis on distributional differences across hours. It can be seen that the morning use on weekdays is bimodal, some work days there is low usage, which might indicate the person is working from home and also having a late start.

Table 1: Hierarchy table for cricket where overs are nested within an inning, innings nested within a match and matches within a season.

linear (G)	single-order-up cyclic (C)	period length/conversion operator (K)
over	over-of-inning	20
inning	inning-of-match	2
match	match-of-season	k(match, season)
season	1	1

ordering from past to future with the game progressing forward with each ball. In the Twenty20 format, an over will consist of 6 balls (with some exceptions), an inning is restricted to a maximum of 20 overs, a match will consist of 2 innings and a season consists of several matches. Thus, similar to time, there is a hierarchy where ball is nested within overs, overs nested within innings and innings within matches. The idea of cyclic granularities can be likewise mapped to this hierarchy. Example granularites then include ball of the over, over of the inning and ball of the inning. Although most of these cyclic granularities are circular in design of the hierarchy, in application of the rules some granularities are aperiodic. For example, in most cases an over will consist of 6 balls with some exceptions like wide balls or when an inning finishes before the over finishes. Thus, the cyclic granularity ball-of-over will be circular in most cases and aperiodic in others.

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India contested by eight teams representing eight different cities in India. The ball by ball data for IPL season 2008 to 2016 is fetched from Kaggle. The `cricket` data set in the `gravitas` package summarizes the ball-by-ball data across overs and contains information for a sample of 214 matches spanning 9 seasons (2008 to 2016) such that each over has 6 balls, each inning has 20 overs and each match has 2 innings. This could be useful in a periodic world when we wish to compute any circular/quasi-circular granularity based on a hierarchy table which look like Table 1.

However, even if the situation is not periodic and a similar hierarchy can not be formed, it can be interesting to visualize the distribution of a measured variable across relevant cyclic granularities to shed light on the aperiodic behavior of a non-temporal data set similar to aperiodic events like formal meetings, workshops, conferences, school semesters in a temporal set up. There are many interesting questions that could possibly be answered with such a data set irrespective of the type of cyclic granularities.

First, it would be interesting to see if the distribution of total runs vary depending on if a team bats in the first or second innings. The Mumbai Indians (MI) and Chennai Super kings (CSK) appeared in final playoffs from 2010 to 2015. We take their example in order to dive deeper into this question. From Figure 3(a), it can be observed that for the team batting in the first inning there is an upward trend of runs per over, while there is no clear upward trend in median and quartile deviation of runs for the teams batting in the second inning. This seem to indicate that players feel mounting pressure to score more runs as they approach towards the end of the first inning. Whereas teams batting in the second inning have a set target in mind and are not subjected to such mounting pressure and may adopt a more conservative strategy, to score runs. Thus winning teams like CSK and MI seem to employ different inning strategies when it comes to their batting order.

Another interesting question could be: do runs per over decrease in the subsequent over if fielding (defending) was good in the previous over? For establishing the fielding quality, we apply an indicator function on dismissals (1 if there was at least one wicket in the previous over due to run out or catch, 0 otherwise). Runs in the current over is then the observation variable. Dismissals in the previous over can lead to a batsman adopting a more defensive play style. Figure 3(b) shows that no dismissals in the previous over leads to a higher median and quartile spread of runs per over as compared to the case when there has been at least one dismissal in the previous over.

Wickets per over are considered as an aperiodic cyclic granularity with wickets as an aperiodic linear granularity. These granularities do not appear in the hierarchy table since it is difficult to position them in a hierarchy. These are similar to holidays or special events in temporal data.

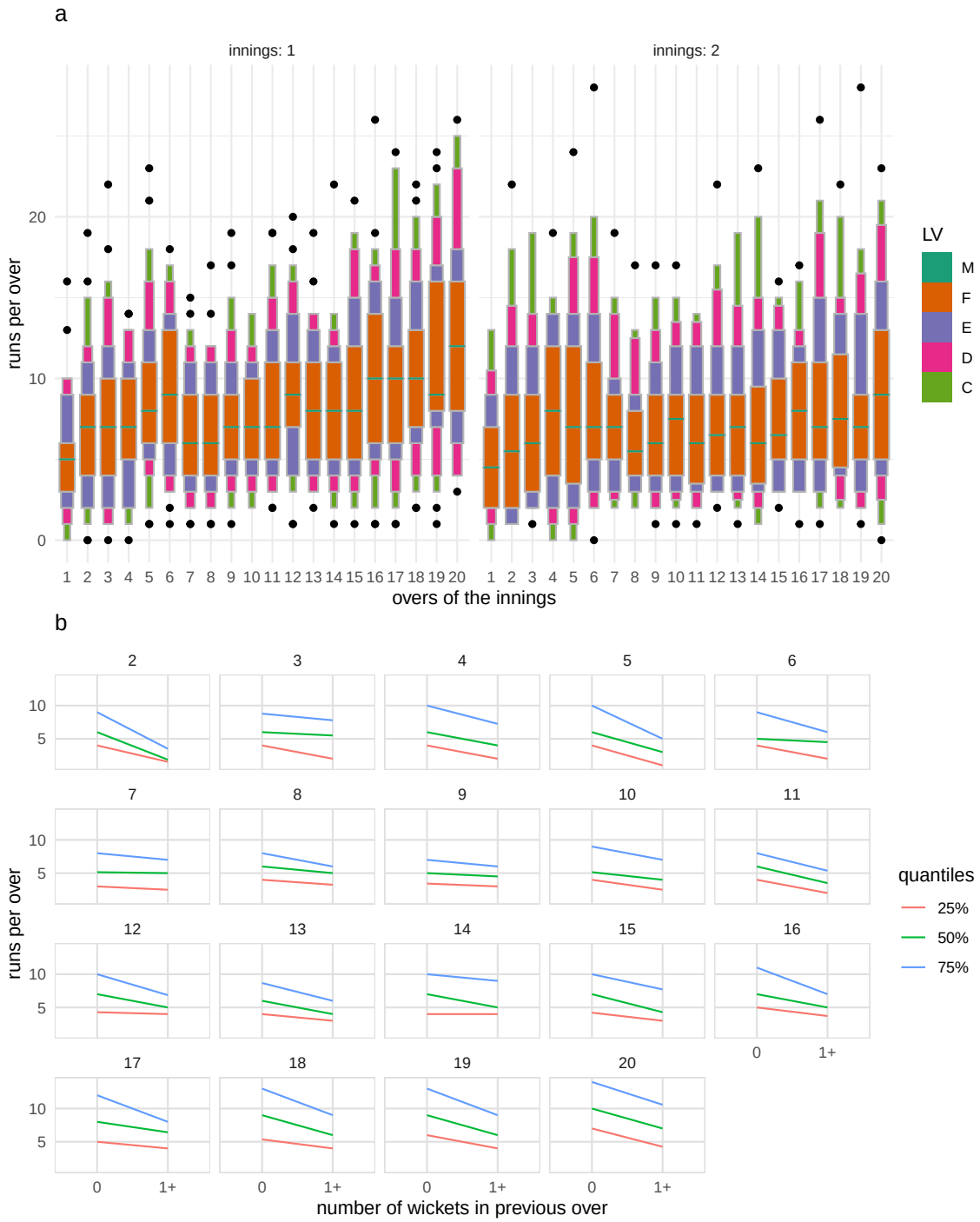


Figure 3: Runs per over shown with different distribution displays, and granularities. Plot (a) shows letter value plot across overs faceted by innings. For the team batting in the first innings there is an upward trend of runs per over, while there is no such pattern of runs for the teams batting in the second innings. Plot (b) shows quantile plot of runs per over across an indicator of wickets in previous over faceted by current over. This indicates that at least one wicket in the previous over leads to lower median run rate and quartile spread in the subsequent over.

Summary and discussion

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

Gupta, Sayani, Rob Hyndman, Di Cook, and Antony Unwin. 2019. *gravitas: Explore Probability Distributions for Bivariate Temporal Granularities*. https://CRAN.R-project.org/package=gravitas.

Haan, Laurens de, and Ana Ferreira. 2007. *Extreme Value Theory: An Introduction*. Springer Science & Business Media.

Kullback, S, and R A Leibler. 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): 145–51.

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

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