Slide 1

Hello everyone! It is very interesting that this part of the work that I am going to present today actually originated from the questions and discussions during my Mid candidature presentation. So thank you so much all of you for your questions and suggestions that day. In the next 20 minutes, I will share with you a quick recap of my research work so that we all know how the current work fits in that framework. I really hope you enjoy the talk today and the questions and discussions along the way lead to expanding and improving the existing ideas. There is a fun live quiz for you along the way, which is basically for testing myself on how well I have been able to formalize the reasonings. This is a joint work with my thesis supervisors Di and Rob.

Slide 2

So I have been working on visualizing probability distributions across bivarite cyclic granularities. All the ideas presented have been implemented in the open-source R package `gravitas`.

Slide 3

So what are time granularities? They are different time deconstructions that can assist in exploration of temporal data sets. Linear time granularities respect the linear progression of time like hours, days and weeks and generally represented through a line plot going from past to future. Cyclic granularities like hour of the day or day of the week could be used to explore periodicities in the data. These granularities can be considered to be categorical variables (ordered or unordered) which induces a grouping of the observations, that is, going from linear to cyclic we have a scenario where we have multiple observations for each category. Hence, there is a need to summarize them.

Slide 4

In these cases, it is common to see aggregates of a numeric response variable or just one particular summary statistic. But studying aggregates hides the distributions of the response variable at finer temporal scales. Hence we decided to look at the probability distribution to summarize these multiple observations.

There are several possibilities at your disposal for visualizing statistical distributions. Traditional methods of plotting distributions include box, violin or ridge plots, whereas more recent forms of visuslizing distributions include Letter values, quantiles or highest density region plots.

Slide 5

The data structure considered for our visualisation is a tsibble, a data structure developed by the former PhD student Earo Wang. A tsibble consists of an index, key and measured variables. An index is a variable with inherent ordering from past to present and a key is a set of variables that define observational units over time. In a tsibble, each observation (row) is uniquely identified by index and key. Since, any cyclic granularity is a function of the index set, we extend the tsibble to obtain columns corresponding to cyclic granularities. And why are we interested to look at multiple cyclic granularities? Exploratory Data analysis developed by **John Tukey** encourages us to look at data from multiple perspectives and looking at multiple cyclic granularities would help us do that. More than 2 cyclic granularities could be visualised, but we focus on visualizing two cyclic granularities (C_i and C_j) by representating it in a 2D space, where C_i maps to categories $\{A_1, A_2, \ldots, A_K\}$ and C_j maps to categories $\{B_1, B_2, \ldots, B_L\}$

For relating data space to the graphic space through layered grammar of graphics using one numeric response variable and two cyclic granularities, we employ the facet-ing approach. is a mechanism for splitting the

data into subsets, then plotting each subset in a different panel. We map C_i to facets, so each C_i label is a facet and for each facet, C_j is mapped on the x-axis and the response variable is mapped on the y-axis.

Now suppose we have N_C cyclic granularities, the number of displays that you can make becomes too large to conume too soon. Just with 10 cyclic granularities, we will have 10 choose 2, which is 45 displays to look and analyze. Without a systematic framework, it is difficult for an analyst to do that. So we will essentially try to do that and employ a strategy which will help us to screen only those visualizations which are interesting.

Slide 6

Now, to start with not all pairs are compatible to bring out the best of exploration. For example, take the first one as an example,here facets show month of the year and x-axis show day of the year - we are unable to compare the distribution across facets because many of their combinations are missing. This is also intuitive because the first day of the month can never be the 2nd or 3rd day of the year. These pairs which lead to structurally empty combinations are called clashes. The pairs that are compatible with each other are called harmonies. For the second plot, every day of the week corresponds to every month of the year and vice versa. Hence there are no empty combinations. Still for large N_C, there could possibly be many harmonies. Can we rank them in order of importance? Can we scrape some harmonies for which variations in the response variable is not significant enough?

Slide 7

So let's see if we can do that. So all these four graphs displays harmonies. Can I have you to rank them from most to least interesting? I know here the interesting word is not well defined, but I would want you to rank them according to your definition of interestng. Take a couple of minutes and fill up the poll.

Slide 8

Time for some interaction! Can I have all of you think for a couple of minutes about their ranking rationale here? What all did you think while ranking them?

Slide 9

This is what we thought while ranking them!

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, given our data structure, it is easier for our eyes to capture differences between categories within a group rather than categories across groups. Displays that capture more variation within different categories in the same group would deem to be important to our eyes. Hence, the idea here is to rank a harmony pair higher if this variation between different levels of the x-axis variable is higher on an average across all levels of facet variables. One potential way is to compute pairwise differences between distributions across different x-axis categories within a group. Also, since we are looking at time here, which is ordered, even within a group we aim to look at ordered pairwise differences. Did you think with more facet levels, chances are more that you will find them interesting? You are right. So there is a need to normalize for the number of levels. Would you even consider looking at all of them given a choice? May be you won't if there is no significant within or between group variation.

Slide 10 (not yet there)

Considering all of this in our metric MMPD(Median Maximum Pairwise distance)

$\frac{median_k(\max_{s,t}(JSD(p_s,p_t)-a_k)/b_k}{log(K)}$,

where JSD is the Jensen Shannon distances between probability distribution p_s and p_t of the s^{th} and t^{th} x-axis category and a_k and b_k are the normalizing constants for the k^{th} facet category.

The thresholds are chosen using ordered permutation tests.

Slide 11

Here I show the harmonies from highest to lowest MMPD and reveal the ranks that we obtain using our metric. How many of you got something else and want to argue if their ranking is better? Bring it on :)

Slide 12

We can further reduce the number of harmonies to visualise using a thresholds computed using permutation tests and as we can see we are left with few harmonies from the long list of cyclic granularities we might have.

Slide 13

Furthermore, we can see here that the number of levels do not have a role to play in the value of MMPD, which essentially implies normalization has been done to eliminate the impact of different levels.

Slide 14

So we started with 6 cyclic granularities and hence $\binom{6}{2} = 30$ displays, reduced to 16 displays using the concept of harmonies and clashes and then reduce it futher to 6 using the threshold - moreover ranked them in order of importance. Now 6 displays are much easier to analyze than 30, isn't it?

Slide 15

Thank you all for listening.