Slide 1 and ## Slide 2

Hi everyone! Today I'm going to talk about visualizing probability distributions across bivariate cyclic temporal granularities. My name is Sayani Gupta and I am currently doing my PhD at Monash university. If you would like to follow along the slides, you can find them at https://sayanigupta-ows2020.netlify.app/. You can follow me on Github @Sayani07 and on Twitter at SayaniGupta07. This is a joint work with my thesis supervisors Di and Rob.

Slide 2

A smart meter is a device that digitally measures your energy use. They are now in place in all homes in Victoria. Which means every year, data for 6.4 million households are recorded for every 30 minutes of an hour, for every 24 hours of the day and for 365 days of the year - which in turn implies that there are more than 100 billion half hourly observations that are collected per year. These households are have different demographic properties such as the existence of solar panels, central heating or air conditioning as well as different behavioral patterns. In this display, I plot the energy consumption along the y-axis against time from past to future. As can be observed from this animation, energy consumption in households vary substantially, which is a reflection of their varied behaviors. If we plot the raw data for all of them, it is very difficult (if not impossible) to get useful insights of their behavior. The motivation of my work comes from the desire to provide methods to better understand these kinds of large quantities of time series data that are observed more than once per year.

Slide 3

My research proposes that deconstructing a time index into time granularities can assist in exploration of periodicity and automated analysis of large temporal data sets. There are different classes of time deconstructions. 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

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 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 map C_i to facets, C_j to the x-axis and the response variable to the y-axis.

Now suppose we have N_C cyclic granularities, the number of displays that you can make becomes too large to consume too soon. Just with 10 cyclic granularities, we will have 10 permutation 2, which is 90 displays to look and analyze. Without a systematic framework, it is difficult for an analyst to do that. So we will essentially try to scrape some displays and keep only those displays which could be interesting.

Slide 6

Now, to start with not all pairs are compatible with each other for exploration. 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 remove some harmonies from the list for which variations in the response variable is not significant enough?

Slide 7

Look at both of these graphs that display harmonies. Clearly (b) captures more variation of the response variable than the other one. 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, 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 efficiently compute a statistical measure that capture the within and between group variation and remove all harmony pairs for which variation is not significant.

Slide 8

Also, we should 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. So here you can see that the variation across different category of the x-axis is higher for (b) than for (a), and hence (b) should be ranked higher. The measure we obtain is called Median Maximum Pairwise distances (MMPD) which used Jenson-Shanon disvergences for measuring distance between distributions and uses the Fisher–Tippett–Gnedenko theorem for normalising for the number of categories.

Slide 9

All the ideas presented have been implemented in the open-source R package `gravitas`. I will post the links at the end again.

Through the package, we provide methods to create all possible granularities for a time index, determine feasibility of examining them together called harmony or clash, refining the exploration by looking at only significantly different pairs and recommending appropriate distribution display.

Slide 10

We will see an example of a sample data set of Smart-Grid Smart-City project from Department of the Environment and Energy, Australia. It has the columns customer_id which serves as the key, the reading_datetime is the index of the tsibble and general_supply_kwh is the measured variable.

Slide 11

So we started with 7 cyclic granularities and hence have ${}^7P_2 = 42$ displays to start with.

Slide 12

The list got reduced to looking at only 16 displays using the concept of harmonies and clashes.

Slide 13

We futher reduce it to 6 by choosing harmonies whose MMPD are above a significant threshold -

Slide 14

Moreover ranked them in order of their ability to capture maximum variation in the response variable through MMPD. Now 6 displays are much easier to analyze than 42, isn't it?

Slide 14

Thank you all listening. Please visit the github repository for more information on the work.