My presentation script

Sayani Gupta

Slide 1

Good afternoon everyone, Thank you all for coming along to listen to my presentation and support me. I am in my 2nd year of PhD and working with Rob and Di to develop methods on exploring probability distributions for bivariate temporal granularities. So the format is I will be talking to you for half an hour sharing what I have been working on these two years and then you are free to ask me any questions that you might have.

Slide 2

Let's start with why I got interested in this problem. Electricity smart meter technology facilitates the collection of energy usage data for residential and commercial buildings at much finer temporal scales than was possible previously. For example, in the state of Victoria, 97% of the households have a smart meter installed in their home. These smart meters are devices that digitally measure the energy use at half hourly intervals, resulting in huge volume of data. What I had from CSIRO initally and later made available by Australian government was the smart meter data for 14K households of Smart-Grid Smart-City project across different local government areas in New Castle, and parts of Sydney from 2012 to 2014, which amounted to 345 millions of observations for households spread across time and space that I wanted to explore.

Slide 3

To have a perspective of how different the energy consumption for each household is, 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.

Slide 4

Things become more hard when we try to explore repetitive behavior in the data. For example, here I show the the energy consumption across each hour of the day for one household. When I want to restructure time from linear to periodic, the structure of the data gets re-organized such that there are several data points for each hour of the day for just one household.

Slide 5

Hence, for several households, there will be a blob of points for each hour of the day.

If we plot the raw data for all of them, it is very difficult (if not impossible) to get useful insights of their behavior. So summarisation of this data is important to analyze regular pattern or anomalies of these households or find similar pockets of behavior to alleviate the problem of volatility in production by capitalizing on the flexibility of these consumers.

Slide 6

When it comes to summarization of this data, it is common to see aggregates of usage across households or just one particular summary statistic. But studying overall energy use hides the distributions of usage

at finer scales and could be misguiding. Here, I show Anscombe's Quartet which displays clearly different and visually distinct datasets are producing the same statistical properties, which not only demonstrate the importance of visualizing data but also demonstrate how excessive aggregation can lead to obscuring essential features of the data.

Slide 7

One of the potential criterion to summarize data in our scenario would be to have the following properties: A summarization which leads to dimension reduction but not at the cost of losing some essential features of the data like shape and uncertainty, alleviates the impact of having households who follow different calendars for holidays. Also, some robustness against incomplete time series is highly important, as data collections may have occurred for different periods of time limiting the chance of analyzing a pool of customers that all have synchronised time series.

Slide 8

Well, so what is my problem?

How can I systematically explore large quantities of these temporal data across deconstructed time (like hour of day, day of week) by analyzing the probability distributions and best exploiting the characteristics of time.

However, the motivation came through the smart meter example, this is a problem which relates to any time series data that needs to be analysed for different periodicities.

Slide 9

Now, I have broken down the research aim into three objectives each of which will form the main chapters in my thesis. The first one is to develop methods to visualize probability distributions over different time granularities. This approach helps in comprehending periodic behavior of one or few households from multiple perspectives. However, when we have large number of households to analyze, we need to move on to clustering these households to explore similar pockets of behavior and then characterize their patterns or irregularities through visualization. In the third chapter, I want to study the Australian smart data thoroughly and provide a preliminary exploratory visualization and summarisation to characterize this data which could be a foundation for any future work with the same data. Further, employ cluster analysis to obtain households showing similar periodic behavior and combine the findings with external data like weather conditions, socio-economic or other demographic factors of those households. Lastly, since earlier studies mostly focussed on clustering raw data across linear time scales we compare how this method compare to clustering probability distributions across periodic scales.

Slide 10 - 12

Briefly I jump on to discuss the details of the first chapter, I want to briefly go through where I stand in my PhD journey. I had the opportunity to do an internship with Google Summer of Code and present my work at two conferences all of which have helped shape the theme of this research. I plan to submit the first work to JCGS by April. I expect to finish the second and third chapter by October and February 2020 respectively. Alongside, as I am learning to reap the benefits of open source softwares and reproducibility, I plan to develop R packages for each chapter and make it available on CRAN. The R package gravitas correponding to the first chapter is already aviable on CRAN.

Slide 13

Quickly coming back to share the details of the 1st chapter now, where we aim to Visualize probability distributions over different time granularities. The key terms are decontructing time and visualizing distributions.

Slide 14

Let's talk about time deconstructions first since time is a data dimension with distinct characteristics. If we call any abstraction of time as a granularity, granularities can be defined from different standpoints. The first one being arrangement, where granularities are said to be linear when they are defined unidirectionally from past to future. Granularilities can be circular when they repeat at regular intervals like day-of-week, or quasi-circular like day-of-month or aperiodic like public or school holidays. Moreover, the hierarchical structure of many granularities creates a natural nested ordering. For example, hours are nested within days, days within weeks, weeks within months, and so on. We refer to granularities which are nested within multiple levels as "multiple-order-up" granularities. For example, hour of the week and second of the hour are both multiple-order-up, while hour of the day and second of the minute are single-order-up.

Slide 15

How can we compute these time granularities? Bettini et al in 1998 had laid out the foundation for time granularities and latter Ning et al in 2002 introduced the calendar algebra for generate new granularities recursively from the older ones. However, these definitions, properties and calendar algebra are inadequate for conceptualizing cyclic time granularities.

Slide 16

Hence, in our paper we re-defined cyclic time granularities and introduced cyclic calendar algebra which could be used to generate new cyclic granularities recursively from older ones. These definitions are new or derived out of the concepts that had been laid out in Bettini et all. Since, circular granularities repeat at regular intervals, their definition could be handled by modular arithmetic. For quasi-circular granularities, the definitions follow from the fact we need to know the pattern of the granules in each period. And the aperiodic cyclic granularities are the ones that either don't repeat or repeat after infinitely long periods of time. Knowing where they belong in the linear scale help us to represent them in the repeating scale.

Slide 17

But 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. For example, the first graph containing cyclic granularities hour-of-day and days-of-week help us to understand if certain hours of the day within a week are different? Whereas, the next one with different cyclic granularities help to know if certain days of the week are different for different months? Different combinations of cyclic granularities lead to different answers and different perspectives which are essential for EDA. Even more than 2 cyclic granularities could be visualised, we focus on visualizing two cyclic granularities by representating it in a 2D space.

Slide 18

The data structure considered for our visualisation is a tsibble. Tsibble is 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 get two more columns corresponding to two cyclic granularities. For relating data space to the graphic space using layered grammar of graphics, we map the two cyclic granularities to x-axis and facets and the measured variable to y-axis throughout our work.

Slide 19

However, we need to be cautious about how two granularities interact with each other. Not all pairs are compatible to bring out the best of exploration. For example, take the forth one as an example, here facets show month of the year and x-axis show day of the month - 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.

Slide 20

The next key point is visualizing probability distributions. We have several possibilities at your disposal for visualizing statistical distributions. Each comes with some pros and cons which we need to consider while choosing the best one for our context.

Traditional methods of plotting distributions include boxplots which display a compact distribution

or violin plots add the information available from local density estimates to summary statistics provided by box plots.

More recent forms of visulizing distributions include Letter value plots which convey detailed information about tails of the distribution or quantile plots which avoids much clutter and just enable us to focus on specific probabilities. Other options can be ridge plots or many variations of these.

For all these plots, we should be vigilant of the number of the number of observations based on which distributions are plotted. Rarely occuring events like 366th day of the year or unequal distribution of event lengths could lead to misleading plots.

Slide 21

Number of levels or categories also has an impact on the choice of visualisation. Space and resolution might become a problem if the number of levels are too high. In this example, I want to know how different hours of the day are different for weekend and weekdays through a ridge plot in the first case and a boxplot in the second case. The first plot is obscuring becuase there is overlap of distribution for two or more categories of the y-axis. Also, with lot of categories, it is difficult to compare the height of the densities across categories.

Slide 22

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. So depending on what we are asking, we need to decide the mapping of the cyclic granularities. For example, in the first case it is easy to compare weekend and weekday for each houe, whereas in the 2nd case, it's more easy to compare hours within weekdays and weekends.

Slide 23

So my package gravitas has functions which try to address each of these aspects of visualization. First, it lets us assess all granularities at our disposal, compute them, screen the harmonies, check if observations are sufficient for plotting distributions and creates a recommended plot based on number of levels and mappings.

Slide 24

We will see an example of the same data set that we initially spoke about.

Slide 25

Set of granularities that we can look at is 15. So if we choose any two from them, we can have a total of 15 combination 2, that is, 156 plots that we have to visualize to have multiple perspectives of the data.

Slide 26

Good news! Thanks to the idea of harmonies, we only have 13 out of 30 to visualize.

Slide 27

For each of these harmonies, gran_advice provides recommendation on the combination of cyclic granularities to be drawn and information on if they are clashes, if number of observations are enough and homogeneity and heterogeneity across facets.

Slide 28

Now that we have 13 harmonies to visualize, we can decide on the distribution plot based on if we want to explore patterns or anomalies.

For example,

We plot the hours of the day on the x-axis and months of the year across facets and energy consumption of 50 households on the y-axis. The narrowest band runs from 25 to 75th percentile, the next one from 10th to 90th and the next from 1st to 99th. What we see from the plot is the distribution is extremely skewed to the left as the lower boundaries of the bands are not visible, whereas the upper boundaries are. The good news is 50% of the households (25) are using energy within the range of 0.1 Kwh. The next 12 households have different behavior only during the peak morning and evening hours in summer. While, the top 5 energy users consume significantly more energy through out the day for all months.

Insights like these can be drawn about the behavior of the households which were not obvious if we plot a summary statistic or see overall usage.

Slide 28

To conclude this, I would quickly want to add that these analysis can also be done for non-temporal data which have a nested hierarchical structure. For example, in cricket, if we hypothesize each ball as an unit of time and think that balls are nested within overs, overs within innings and innings within matches, we can do some behavioral comparisons for teams.

Slide 29

We take two top teams from Indian premiere league and plot their run rate across each over of the innings faceted by innings. We see for one team their run rate is really volatile throughout the innings, be it first or 2nd innings. Whereas, for the other one, which is considered to be a better team, run rates are more consistent with letter values not so distinct in the initial over of the innings and only becoming distinct as they approach the end of the innings.

Slide 30

This is a joint work with Rob Hyndman and Di Cook. The slides are created using .... .