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 large quantities of temporal data. So the format is I will be talking to you for half an hour sharing what I have been working on 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. Smart meters are devices that digitally measure the energy usage for residential and commercial buildings and Electricity smart meter technology facilitates the collection of energy usage data at much finer temporal scales than was possible previously. For example, in Victoria, 97% of the households have a smart meter installed in their home. What I had from CSIRO initally and later made available through 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. In most cases, this data will have multiple seasonal patterns across days of the week when more electricity will be consumed for working days and less for non-working days or hours of the day since more energy will be consumed when members of the household are awake than when they are asleep or across months due to different heating or cooling methods. But we do not learn about the pattern or cycles from this view of linear progression of time. In a cyclic organization of time, where the domain is composed of a set of recurring time values, it would be easier to comprehend patterns.

Slide 4

Moreover deconstructions of time are not restricted to conventional day-of-week or month-of-year and there could be repetitive behavior for different weeks of the month or hours of the week as well. So different unknown periodic behavior might emerge by looking at different cyclic time deconstructions. But when I want to reorganize time from linear to cyclic, the structure of the data gets re-organized such that there are large number of data points for each recurring time values. Exploring probability distributions rather than aggregated information is an useful approach since we have large volume of data. Now, iteratively exploring probability distributions across all possible cyclic time deconstructions for detecting perviously unknown periodic behavior could be overwhelming and thus a systematic way to do so might be helpful.

Visualizing probability distributions across all possible cyclic time deconstructions could be useful to compare the seasonal pattern or anomalies in behavior across few households. But analyzing all of them together can hide key patterns of individual households. Hence, it is useful to club similar households and then examine them for different periodicities. So when we have large number of households to analyze, we need to move on to clustering these households to explore similar pockets of behavior.

Since, the motivation came through the smart meter example, it is then necessary to see if the methods add value in a societal context by combining the findings with external data like weather conditions, socio-economic or other demographic factors of those households.

Slide 8

So my research question is 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?

Slide 9

Now, I have broken down the research aim into three objectives each of which will form the main chapters in my thesis. I will discuss the work done for the first chapter in details and briefly discuss about the next two chapters.

Slide 10

In this work we aim to visualize probability distributions over different time granularities that accomodate for periodicities. The key terms are decontructing time and visualizing distributions.

Slide 11

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 12

How can we compute these time granularities? Bettini et al in 1998 had laid out the foundation for time granularities. He defined these two conditions for defining time granularities, the first of which implies that the granularities are non-overlapping and their index order is same as time order. For example, if we assume that hour is the finest granualarity available such that it is non-decomposable (also called bottom granularities) and we have 100 hours, then we have 100/7 days and 100/7*24 weeks. But all these are uni-directional in nature since the index order is the same as the time order. Ning et al in 2002 latter introduced the calendar algebra to generate new granularities recursively from the older ones. However, these definitions, properties and calendar algebra (latter introduced by Ning et al) are linear and hence inadequate for conceptualizing cyclic time granularities. For example, they didn't express how a finer granularity like

day can be expressed in relation to the coarser granularity week or month or how week could be expressed in relation to month.

Slide 16

Hence, in this work I have conceptualize cyclic time granularities which are additional categorizations of time to express a finer granularity in relation to a coarser granularity.

A circular granularity can be defined using modular arithmetic due to its regular mapping with the bottom granularity (the non-decomposable granularity in the system). So if someone asks what day of the week is day 13. We can say it is 13 mod 7, that is the 6th day of the week. This is possible since each week is composed of exactly 7 days. These labels could be a set of strings that is more descriptive than the index and used to identify a categorization like Sunday, Monday and so on. However, the labels can coincide with indexes when integers are directly used to refer to categorizations of the circular granularity like hours of the week.

Slide 17

A quasi-circular granularity can not be defined using modular arithmetic due to its irregular mapping with the bottom granularity. For example, days in a month can be composed of 28, 29, 30 or 31 days. So if you ask me if what day-of-month is the day 40, then we should be able use this formula to answer that question? The idea here is that if we know the composition of days within each month within one year, we can find how days are distributed within a month beyond any period since the "pattern" repeats itself along the time domain due to the periodic property. We consider a situation without leap years in this case.

Slide 18

Aperiodic time granularities are the ones which can not be specified as a periodic repetition of a pattern of granules. Most public holidays repeat every year, but there is no finite (or reasonably small) period within which their behavior remains constant. An example could be school holidays which comprises of mid semeter break, SWOT vac and semester holidays which do not repeat at regular intervals. If someone asks which day of SWOT VAC is 72nd year of the year, we cannot use the periodic characteristics of time to define it. In that case, using a proxy label for each kind of holidays corresponding to the linear holidays seems to be the best approach.

Now using calendar algebra we can also generate multiple order up singularities from single-order-up granularities and vice versa and this is called cyclic calendar algebra.

Slide 19

Now that we have the ability to compute cyclic granularities, let us look at the data structure that we will consider for exploration. 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. 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. Even more than 2 cyclic granularities could be visualised at the same time, but we focus on visualizing two cyclic granularities by representating it in a 2D space.

Slide 20

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 help us answer different questions and lead to different perspectives which are essential for EDA.

Slide 21

However, can any two cyclic granularities be visualized together for effective exploration? Let us see some examples. The first graph shows the quantile plots across days of the year on x-axis and months on facets. We are unable to compare the distribution across facets because many of their combinations are missing. This is intuitive because the first day of the month can never be the 2nd or 3rd day of the year. These are structurally empty sets because it is impossible for them to have any observations due to the structure of the calendar. Similarly, there can be event based or build based empty combinations. These combinations when plotted together do not aid in exploratory analysis. The pairs that are compatible with no empty combinations 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 .... .