My presentation script

Sayani Gupta

29/09/2019

Slide 1

Hi everyone, today I'm going to present to you exploring probability distributions for bi variate temporal angularities. Before I jump into the problem, I want to briefly go through why I got interested in this problem. A smart meter is a device that digitally measures our energy use. I had the smart meter data for every 30 minutes available from 2012 to 2014. Hence, the finest temporal scale available to me was half an hour. This was available for 14K households across different local government areas in New Castle, NSW and parts of Sydney. The volume was of the order 40 billion observations that I wanted to visualize.

Slide 2

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 difficult to get insights of their behavior.

Slide 3

Moreover, things become more hard when we try to visualize periodicity in the data for all households. For example, if we plot the energy consumption for a household across each hour of the day, the structure of the data suggest that there are several data points for each hour of the day for just one household. Hence, for several households, there will be a blob of points for each hour of the day.

In such cases, 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, making it difficult to examine the distribution of energy behaviors of all households.

Slide 4

Well, so how can you visualize this data, given the huge volume and spread?

There can be numerous ways to draw insights from this data!

Slide 5

But what seemed most exciting to me was a way to explore systematically multiple perspectives of this temporal data across deconstructed time through visualization of probability distributions to find regular patterns or anomalies in behaviors. 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.

I would discuss these two key terms in the next slides.

Slide 6

If we call any abstraction of time as a granularity, granualrities can be defined from different standpoints. The first one being arrangement, where angularities are said to be linear when they are defined unidirectionally from past to future. So we have entire set of time points and they we divide these into blocks of hours, days, weeks and so on. Granularilities can be circular when they repeat at regular intervals like day-of-week, or nearly circular like day-of-month.

Slide 7

The next standpoint comes from the idea of how a calendar or hierarchy is arranged. The hierarchical structure of many angularities creates a natural nested ordering. For example, hours are nested within days, days within weeks, weeks within months, and so on. We refer to angularities which are nested within multiple levels as "multiple-order-up" angularities. 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 8

The challenges when it comes to deconstructing time is to understand how to compute these different circular, aperiodic, single order or multiple order up angularities and be mindful of what are the total number of granularilities that we can look at depending on the level of details of temporal analysis we are looking for.

Slide 9 - 14

The next key point is visualizing probability distributions. We have several possibilities at your disposal for visualizing statistical distributions like boxplots, violin plots, decile plots, letter value plots, ridge plots or many variations of them.

Slide 15

So the challenge here is to decide which distribution plots to choose when are trying to explore patterns or behaviors and be mindful that is the number of observations are sufficient enough to draw a distribution plot.

Slide 16

One can choose to visualize many angularities at once, however, we break the problem and try to visualize only two at a time. Then we get into this problem of how two angularities interact with each other. Not all pairs are compatible to bring out the best of exploration. For example, take these examples. Talking about the forth one here where 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. We call these pairs of angularities as clashes. The pairs that are compatible with each other to bring out the best of exploration are called harmonies.

Slide 17

Hence, challenges can consist of knowing if two angularities are harmonies, before we plot them or having a handy list of harmonies to look at for a given data set.

Slide 18

So my package gravitas has functions which try to address each of these challenges.

It uses the nested nature of time to compute different single order up angularities and then compute multiple orders from them using a recursive function.

The idea in harmony is to check if any categorizations of angularities lead to empty sets and tagging them as clashes.

The idea in granplot is to have a recommended list of distribution plots depending on the levels of angularities to be plotted across facets or x-axis and gran_obs gives us a medium to see the number of observations before we proceed on to draw a distribution plot.

Slide 19

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

Slide 20

Set of angularities 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. Wait, what???

Slide 21

Good news! Thanks to the idea of harmonies, we only have 13 out of 156 to visualize.

Slide 22

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. What we see from the plot is the distribution is extremely skewed to the left. There is less difference between the morning behavior of the top 1% and 10% of the people compared to their evening behaviors for all months. 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. While, the top 5 energy users consume significantly more energy through out the day.

Slide 23

Or see the hourly usage across weekdays and weekends. Few things we can look from the animation is - the energy spikes up pretty steeply for the top 10% households after they get up in the morning. The transition is more gradual for 75% of the households

Also, there are two peaks in the morning for weekends unlike that in weekdays implying they have different times of waking up.

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 24

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 25

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 26

Thank you slide and more information,