My presentation script

Sayani Gupta

01/10/2019

Slide 1

Hi everyone, today I'm going to present to you exploring probability distributions for bivariate temporal granularities. 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 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?

Slide 5 - 6

How can you explore systematically multiple perspectives of this temporal data across deconstructed time?

The solution was to visualize the 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 the two key terms in the next slides.

Slide 7

If we call any abstraction of time as a granularity, granualrities 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. 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 8

The next standpoint comes from the idea of how a calendar or hierarchy is arranged. 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 9

If our time series data is arranged in a tsibble, any single order granularities can be computed using the index of the tsibble which helps to uniquely identify observational units over time, as follows: ... It utilizes the relationship between the index and the linear granulkarity, the hierarchical structure and modular arithmetic to compute single order up granularities.

Multiple order up granularities can be computed from single order up granularities using them recursively. This is possible due to the nested nature of time.

Slide 10

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 11

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.

Standard boxplots display a compact distributional summary with median, quartiles, hinges, whiskers and extreme outliers. These are helpful to get an idea of the distribution at a glance. Disadvantage can be we do not have an idea if the distribution is multimodal and estimates of tail behavior beyond the quartiles are not trustworthy. Also, the number of outliers is large for larger data set.

Violin plots add the information available from local density estimates to summary statistics provided by box plots. Adding two density plots gives a symmetric plot which makes it easier to see the magnitude of the density and compare across categories. The density in violin plots are estimated through kernel density estimation and thus makes assumptions when selecting kernel or bandwidth.

Letter value plots convey detailed information about tails of the distribution, It shows only actual data values and no smoothed summaries.

For decile plots, no distributional assumptions are made about the data since empirical deciles are plotted. It avoids much clutter and just enable us to focus on specific probabilities.

Ridge plots are density plots all aligned to the same horizontal scale and presented with a slight overlap. If there are lot of categories, it is difficult to compare the height of the densities across categories.

For all these plots, we should be vigilant of the number of the number of observations based on which distributions are plotted.

Slide 12

So my package gravitas has functions which try to address each of these three aspects of granularities. You can install it using this link here.

Slide 13

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

Slide 14

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. Wait, what???

Slide 15

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

Slide 16

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.

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 17

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 18

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 19

Thank you slide and more information,