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, 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

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.

Slide 4

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. Just to elaborate it further, i took the example of Anscombe's Quartet which is a set of four datasets, where each produces the same summary statistics which could lead one to believe the datasets are quite similar. However, after visualizingthe data, it becomes clear that the datasets are very different.

Slide 4

Well, so what is my problem?

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

I have been talking about time granularities from my first slide, but there is a need to define them formally. 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. Granularilities can be circular when they repeat at regular intervals like day-of-week, or quasi circular like day-of-month or completely aperiodic like public holidays.

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.

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. Also, any cyclic granularity can be computed using linear granularities and their relationships.

Slide 9

Next we see what is the data structure that we consider for solving the problem.

Tsibble is an explicit data- and model-oriented object. 1. Index is a variable with inherent ordering from past to present. 2. Key is a set of variables that define observational units over time. 3. Each observation should be uniquely identified by **index** and **key**. 4. the rest of the columns are measurements.

We extend a tsibble object with additional columns containing cyclic granularities. Now, following the grammar of graphics which is a framwork to relate data space to the graphic space, we take one of the cyclic granularities as a aesthetic (x-axis), the measurement is considered on the y-axis and the other cyclic granularity on the facet for the purpose of visualisation.

Slide 10

However, we need to be cautious about how two granularities interact using this data structure and mapping. 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

Again, we have several possibilities at your disposal for summarising a probability distribution. 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.

Slide 12

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

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 6. So if we choose any two from them, we can have a total of 6 permutation 2, that is, 30 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 30 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,

gran_advice gives you advice based on the data structure of teh cyclic granularities.

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 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,