Print

Using Trellis Plots and Overlay Variables

Visual tools for exploratory data analysis are particularly useful when you have complex or multidimensional data.

You might have many observations, many variables, many categorical variables with many levels, many measurements taken over time, or data with a geographic dimension.

In this video, you learn how to explore time-ordered data using run charts, trellis plots, and overlay variables.

As an example, we use the data set Mobile Cellular.jmp.

These are data on mobile cell phone subscriptions per 100 people, from 1990 to 2017 (a 28 year period). The data were compiled from the World Bank and the US Census.

The data set includes information about 217 countries grouped into seven regions and four income groups.

The file also contains information on the fixed (wired) telephone subscription per 100, the population, and the land area.

Let's say that you are interested in exploring how mobile phone subscriptions have changed over the past 28 years. What's the best way to visualize this information?

Year and Mobile (per 100) are both continuous variables, so you could use a scatterplot.

It looks like mobile phone subscriptions started to increase in the mid-1990s, and it looks like the overall trend is increasing. However, there's a lot of noise and it's hard to see the overall pattern over time.

You can see this noise when you add box plots for each year.

Can you see how wide the box plots are?

The center of each box plot is the median. You can see an increasing trend in the medians, starting around the year 2000.

When you have time-ordered data, you can use a run chart (or a line graph). This run chart shows the average mobile subscriptions per year across all of the countries. Gridlines were added to help with interpretation.

Can you see the story in this graph? It looks like the line was fairly flat until the mid-1990s, and then average mobile subscriptions increased dramatically in the late 1990s. It also looks like the line leveled out somewhat around 2013

or 2014. Let's take another look at the scatterplot. What's driving all of the noise, or scatter, within each year?

Remember that there are 217 countries, grouped into seven regions and four income levels. Each of these variables is a source of variation in mobile phone subscriptions.

What would the graph look like if you overlay a separate line or curve for each income group on the same graph?

In this graph, you can see that the curves, or profiles, are very different for the four income groups.

Countries designated as high Income were among the first to see an increase in mobile phone subscriptions, in the early 1990s, and this trend leveled off in 2012.

The increase in subscriptions was much later for other income groups.

Now, we use Region as an overlay variable.

Does it look like there is a difference in the profiles of the lines for the different regions?

Notice that North America was among the first regions to see an increase in cell phones but was quickly passed by Europe and Central Asia.

There are also a couple of dips in subscriptions for North America.

How can you look at both the region and income group on the same graph?

You can use a trellis plot.

In a trellis plot, you create a matrix of graphs, where each graph shows a subset of the data.

In this trellis plot, each panel is a region, and the lines within the panel are for the different income groups.

The X and Y axes are the same for each panel, so you can compare the profiles for the income groups across the different regions.

For example, look at South Asia. None of the countries in South Asia are designated as high income, and in 2017 the upper middle income group in South Asia has the highest average mobile subscriptions per 100 people of all groups across all regions.

Trellis plots are not limited to run charts. You can use them any time you want to create a series of plots for multivariate data that is, when you want to graph many variables at one time.

For example, let's say that you're interested in comparing the mobile cellular subscriptions for each of the regions across four years.

You can create a trellis plot with box plots and use a data filter to select the years you're interested in comparing.

This graph shows, in general, how subscriptions changed for the regions over time. Some regions saw a big increase in 2005, others caught up in 2010, and in 2017 most of the regions are similar in terms of the average number of mobile subscriptions.

In this video, you learned about using run charts with overlay variables and trellis plots to explore multidimensional data collected over time.

There are other effective ways of visualizing time-ordered data. In an upcoming video, you learn about two additional methods: bubble plots and heat maps.

Statistical Thinking for Industrial Problem Solving

Close