Trellis Display: Questions and Answers

Richard A. Becker
William S. Cleveland
Ming-Jen Shyu

Abstract

The paper is an ”interview” on the topic of Trellis display with Rick Becker, Bill Cleveland, and Ming Shyu. Rick, Bill, and Ming, members of the Statistics Research group at Bell Labs, originated the basic notions of Trellis display, and also designed and implemented the software. The interviewer has asked not to be identified since Rick, Bill, and Ming could not hide their disgust at a few of the questions.

What Is Trellis Display?
What Is Trellis Display?

Broadly, it is a framework for 2D and 3D data display. A Trellis display can consist of just one panel, or many panels. When there are many panels, there is a structure: a three-way rectangular array of panels with columns, rows, and pages that is reminiscent of a garden trelliswork (which consists of crossed wooden strips upon which vines and other plants grow).

WHAT DOES THIS TRELLIS STRUCTURE DO FOR ME?

It provides a mechanism for displaying databases with many observations and many variables that often leads to penetrating views of the structure of the data. Two or three “primary variables” are selected for display on the common axes of the panels (two axes for 2D displays and three axes for 3D displays). “Conditioning variables” are also selected. Each panel shows values of the primary variables for one combination of values of the conditioning variables. For example, suppose there are four variables: blood pressure, weight, sex, and race. Each panel might be a scatterplot of blood pressure against weight for one combination of race and sex.

IN YOUR EXAMPLE, THE TWO CONDITIONING VARIABLES ARE CATEGORICAL. CAN A NUMERICAL VARIABLE BE A CONDITIONING VARIABLE?

Yes, we use a new data structure called a “shingle”. It consists of the numerical values together with a set of intervals. When the shingle is used as a conditioning variable, then we see values of the primary variables in a single panel that correspond to numerical values of the shingle that lie in one interval. So the shingle is a mechanism for turning a numerical variable into a categorical variable of sorts, but one in which the intervals are allowed to overlap. This is a bit strange at first since people are so used to dividing data up into disjoint subsets, and we often get skeptical queries. But overlapping intervals can greatly increase the resolution of our visualization of interrelationships.

I AM BEGINNING TO GET THE IDEA THAT TRELLIS DISPLAYS OF REALLY LARGE DATABASES CAN EASILY RESULT IN HUNDREDS OF PANELS ON MANY PAGES. CAN I ABSORB ALL OF THIS INFORMATION?

Yes. It's hard work at times, but you have to be realistic. If you have a large database with many variables and your goal is to get a good understanding of the interrelationships, then, unless you get lucky, this complex structure is bound to require some hard work to understand. The trellis structure helps you to absorb all of this, because of its regularity. However, no method is going to let you see the complexity without your expending a fair amount of effort. The one thing that now seems pretty clear is that multipage displays are necessary for many large database visualizations; it is just not feasible in many cases to cram all of the information about a large database into a single page or window. You just do not have the real estate.

BUT THE TRELLIS DISPLAYS I HAVE SEEN HAVE A LOT OF INFORMATION ON EACH PAGE. EACH PAGE SEEMS QUITE BUSY ACTUALLY.

Yes, in our visual design of Trellis display we worked hard to allow dense packing of information on a single page. For example, color becomes very important to such data packing. But one page is often still not enough.

OK, BUT THERE IS ANOTHER WAY TO APPROACH THE STUDY OF A LARGE DATABASE. DEVELOP A STATISTICAL MODEL AND SEE IF IT FITS THE DATA. IF IT DOES FIT, USE THE MODEL TO LEARN ABOUT THE STRUCTURE OF THE DATA.

Yes, and Trellis display is a big help in doing this because it allows you to make a good guess about an initial model to fit and then to diagnose how well it fits the data. This is one of the important themes in the book *Visualizing Data*.

BUT INSTEAD OF AGONIZING OVER ALL THOSE PANELS I COULD DO A BUNCH OF CHI-SQUARED TESTS FOR GOODNESS FIT.

You're joking, right? If not, we're leaving.

OK, I GUESS I'M JOKING. LET'S PUSH ON. IT ALSO SOUNDS LIKE YOU HAVE A LOT TO SPECIFY WHEN YOU MAKE A SINGLE TRELLIS DISPLAY. DOESN'T THIS MAKE HIGH-LEVEL SOFTWARE MORE COMPLICATED FOR THE USER?

It is true that you have a number of things to think about. For the trellising you need to specify the primary variables, the conditioning variables and their order, the levels of the conditioning variables and their orders, and the dimensions of the trellis. But we think that our high-level design for Trellis software leads to a control mechanism that, if not trivial, is at least sufficiently simple that it will not deter users. There are a number of key ideas that contribute to this.

WHAT ARE SOME OF THOSE IDEAS THAT HELP MAKE TRELLIS DISPLAYS EASY TO SPECIFY?

One involves labeling. The labels that convey the values of the conditioning variables for a single panel are plotted right on the panel and not in margins. This, for reasons we will not try to convey here, was a critical idea. It took us a long time to figure it out. Either we are geniuses for having thought of it at all or dummkopfs for not having figured out the obvious immediately; but whatever we are, the idea seems to work quite well. We also allow the data structures to carry the labels and avoid complicated labeling specifications.

Another idea is to use a little language to specify the primary and conditioning variables. There are other things like this that contribute to straightforward control that we will set out in a technical paper. By the way, this high-level design is language independent; it could be implemented in any system that has basic capabilities for graphics – that is, capabilities for scaling, drawing plotting symbols and lines, and so forth. The S-Plus Trellis Module shows that implementation is feasible.

WHAT ABOUT SMALL DATA SETS?

Well, a Trellis display can be very simple with just a few panels or even one. So if you have two sets of 15 measurements of two variables, say one set taken at night and one during the day, you can make a Trellis display with two scatterplots.

BUT THEN I AM NOT REALLY USING THE TRELLISING IN A VERY PROFOUND WAY.

Right, the trellising is not important, but there are many other features that we have included in the Trellis design that are vital to data display. One is aspect ratio. (The aspect ratio of a graph is its height divided by its width.) Another is the coordination of scales on the different panels of a Trellis display; they can be the same, have the same number of units per cm, or be completely uncoordinated. This provides by far the most advanced high-level control of graph scaling of any graphics software around today. We also have provided defaults for color schemes, for different types of plotting symbols, and for different line types when the display needs to provide good visual assembly of different sets of graphical elements that need to be visually distinguished; these defaults change with the graphics devices you are using.

The assembly problem is particularly acute for Trellis display because we are trying to pack in a lot of data per square cm. We had to work very hard to find decent solutions. Because the above items are mundane issues, people tend to think they are simple. But they actually bring you rather quickly to deeply intellectual matters about how people carry out visual information processing; we have drawn on a lot of work in visual perception to find solutions.

ASPECT RATIO IS DEEPLY INTELLECTUAL? I THOUGHT DARRELL HUFF WORKED IT ALL OUT IN HIS BOOK *HOW TO LIE WITH STATISTICS*. JUST BE SURE TO INCLUDE ZERO ON YOUR SCALE.

Good grief, not you too. Huff had nothing at all to say about aspect ratio except a red herring that those who cannot be bothered to look at the scales on a graph might be fooled. The aspect ratio is vital because it has a large impact on our ability to judge rate of change. A number of studies in visual perception have shown that our ability to judge the relative slopes of line segments on a graph is maximized when the absolute values of the orientations of the segments are centered on 45 degrees. Trellis display allows you to automatically carry out this "banking to 45 degrees". You just need to specify the segments to be banked.

OK, I'LL TRY TO PURGE HUFF'S REMARK FROM MY MIND. BUT BACK TO THE ISSUE OF SMALL DATA SETS, SUPPOSE I HAVE A SMALL AMOUNT OF DATA AND AM IN SOME SENSE LOW IN INFORMATION. FOR EXAMPLE, SUPPOSE I RAN A NEARLY SATURATED EXPERIMENT WITH 32 RUNS, WITH FIVE VARIABLES EACH AT TWO OR THREE LEVELS, AND WITH ORTHOGONAL OR NEARLY ORTHOGONAL ARRAYS. COULD TRELLIS DISPLAY SHOW ME ANYTHING?

We do not yet know. We have actually had a look at two data sets from such experiments and were quite amazed at how much we saw. In both cases the amount of noise was small and in both cases Trellis helped to show how to approach the modeling of such data. At first we reasoned that we could not expect to see much in experiments with more noise. But then it dawned on us that it might be that nearly saturated experiments succeed as often as they do because the variables and their levels are chosen in such a way that noise is low; people need to proceed very conservatively in such experiments, and if there is much noise at all, no method will reveal effects. But two data sets are far too few to conclude anything at this stage.

YOU HAVE NOT MENTIONED ANYTHING ABOUT DIRECT MANIPULATION METHODS FOR TRELLIS. THAT IS, METHODS IN WHICH THE MOUSE IS USED TO ADDRESS GRAPHICAL ELEMENTS ON THE SCREEN TO MAKE ASPECTS CHANGE IN REAL TIME — FOR EXAMPLE, BRUSHING.

Trellis display as we have defined it is a static framework for graphics. As such it provides a visual environment within which direct manipulation methods can be added. In fact, because the Trellis environment is so rich in information, there are many opportunities for developing direct manipulation methods. Since Trellis is new, we have not yet tried anything out, but our first experiment is about to begin. We are going to explore direct manipulation methods for managing the Trellis pages on the screen. There are some interesting 3D display and manipulation tools that might prove useful. We hope others will get ideas and experiment as well.

COULD YOU TELL US HOW TRELLIS DISPLAY WAS DEVELOPED AND HOW IT LED TO THE S-PLUS TRELLIS MODULE?

As work on the book *Visualizing Data* progressed, it began to take more and more energy to write S code to carry out the methods that were emerging. About two years before the book was completed we began a project whose goal was to develop software that would allow simple access to the methods. By the time the book was done we had a pretty good system in place, but there were many inadequacies. We kept right on working and a number of new ideas began to emerge; before long we had something that was far more general and did far more than was encompassed by the original goal. About a year ago, when the visual design of Trellis display and the high-level design of the software were completed, we contacted StatSci and began a project with them to bring the software to S-PLUS users. Stephen Kaluzny and Bill Dunlap at StatSci joined us and contributed a number of important ideas to the implementation. They also worked with us to hunt for bugs. Rich Calaway at StatSci worked on the documentation.

IF I CAN BE BLUNT, HOW DO YOU KNOW TRELLIS WORKS? MOST IDEAS FOR DATA

DISPLAY HAVE NOT REALLY WORKED OUT VERY WELL. I CAN THINK OF MANY METHODS THAT CAUSED A BIG SPLASH BUT THEN NEVER GOT USED. HOW DO WE KNOW TRELLIS WILL NOT SIMPLY SINK AFTER ITS SPLASH?

Well, it is possible that Trellis has properties that we did not anticipate that in the end will lead to its not being used, but to be equally blunt we think that somewhat unlikely for a very simple reason. The ideas of Trellis display have been under development for over four years if you count the time spent experimenting with the methods that are in *Visualizing Data*. During this entire time the evolving methods were used to analyze data — data that were lying around from past data collections and data from ongoing projects where people cared about subject matter conclusions and did not particularly care what methods were used so long as the conclusions were sound. Throughout, the methods of Trellis display seemed to do a good job of revealing structure in data. We think *Visualizing Data*, which describes a part of what is now Trellis display, demonstrates this. During the next months we hope that the writings we will produce on Trellis will also demonstrate this. In summary, unlike many graphical methods that have been suggested, and more generally, many statistical methods that have been suggested, Trellis display was shaped by usage on real data to solve real problems.

THANK YOU FOR YOUR TIME. IF SOMEONE HAS QUESTIONS, WILL YOU TRY TO ANSWER THEM?

Yes. Email is best. Our addresses are rab@research.att.com (Becker) wsc@research.att.com (Cleveland), and shyu@research.att.com (Shyu)