Abstract

Trellis display is a framework for the visualization of data. Its most prominent aspect is an overall visual design, reminiscent of a garden trelliswork, in which panels are laid out into rows, columns, and pages. On each panel of the trellis, a subset of the data is graphed by a display method such as a scatterplot, curve plot, boxplot, 3-D wireframe, normal quantile plot, or dot plot. Each panel shows the relationship of certain variables conditional on the values of other variables.

A number of display methods employed in the visual design of Trellis display enable it to succeed in uncovering the structure of data even when the structure is quite complicated. For example, Trellis display provides a powerful mechanism for understanding interactions in studies of how a response depends on explanatory variables. Three examples demonstrate this; in each case, we make important discoveries not appreciated in the original analyses.

Several control methods are also essential to Trellis display. A control method is a technique for specifying information so that a display can be drawn. The control methods of Trellis display form a basic conceptual framework that can be used in designing software. We have demonstrated the viability of the control methods by implementing them in the S/S-PLUS system for graphics and data analysis, but they can be implemented in any software system with a basic capability for drawing graphs.

1 Introduction

1.1 Barley Data: The Detection of A Probable Error

In the 1930s an experiment was run in the state of Minnesota in the United States. At six sites, ten varieties of barley were grown in each of two years. The data collected for the experiment are the yields for all combinations of site, variety, and year, so there are $6 \times 10 \times 2 = 120$ observations. The experiment is of historical interest because it is one of the early field trials that incorporated R. A. Fisher's ideas on randomization and the analysis of variance. The agronomists published the data and an analysis of them in a 1934 paper [11]. Fisher published the data in his classic book, *The Design of Experiments* [10], but he did not present an analysis. Fisher's publication gave the data a large exposure, and many others tried their hands at analyzing them to illustrate new statistical methods [1, 2, 6]. We will do the same here, using the data to illustrate Trellis display. The visualization using Trellis reveals an important happening in the data — there appears to be a major error, one that survived undetected for six decades [4].

1.2 Trellis Display of the Barley Data

Figure 2 is a Trellis display of the barley data. Each panel displays the yields of the ten varieties for one year at one site.

Figure 2 uses an important display method: *main-effects ordering of category levels*. For these barley data, the explanatory variables are categorical. (Since there are only two years, the year variable is also treated as categorical.) The unique values of each categorical variable will be referred to as *levels*. For example, the levels for the site variable are Grand

Figure 2: A dotplot of the barley data showing yield against variety given year and site.

Rapids, Duluth, and so forth. The level medians are a measure of the main effects, and we have arranged that the levels for each variable are ordered based on level medians. On each panel the varieties are ordered from bottom to top by the variety medians; Svansota has the smallest median and Trebi has the largest. The site panels have been ordered from bottom to top by the site medians; Grand Rapids has the smallest median and Waseca has the largest. Finally, the year panels are ordered from left to right by the year medians; 1932 has the smaller median and 1931 has the larger. Later, we will discuss why main-effects ordering is important.

Visually scanning up each column of Figure 2, we can see an anomaly: for each year, the values for Morris appear out of place. Because of the main-effects ordering, the site medians increase from bottom to top. The ordering is preserved in each year separately except for Morris. But the visual impression is that if we were to interchange the years at Morris, the site would then fit into the patterns formed by the other sites.

This suggests another display. In Figure 3, the data are graphed again, but this time the 20 values for each site are graphed on a single panel with the year variable encoded by the plotting symbol. Now we can see clearly that at each site except Morris, the overall yield for 1931 is greater than 1932, but the reverse is the case for Morris. However, something else quite critical is also apparent. At Morris, the overall level of the absolute differences between the years has a value that is commensurate with the corresponding values at the other sites. (This is actually the same observation from Figure 2, that Morris would fit the pattern were we to interchange its years.) This suggests that there might be an error in the data at Morris, a reversal of the years. Either there is an error, or nature just happened to reverse effects at Morris in such a way that 1932 exceeds 1931 by an amount similar to the amounts that 1931 exceeds 1932 at the other sites. We will probe this issue later with other Trellis displays.

1.3 Trellis Basics

The salient visual aspect of Trellis display is a three-way rectangular array of panels with *columns*, *rows*, and *pages*. In Figure 2 there are 12 panels, 2 columns, 6 rows, and 1 page. In Figure 3 there are 6 panels, 1 column, 6 rows, and 1 page. Later, we will show a Trellis display with more than one page. We refer to the rectangular array as the *trellis* because it is reminiscent of a garden trelliswork.

Each panel of a trellis display shows a subset of the values of *panel variables*; these values, are formed by conditioning on the values of *conditioning variables*. In Figure 2 the panel variables are variety and yield, and the conditioning variables are site and year. On each panel, values of yield and variety are displayed for one combination of year and site. For example, the lower left panel displays values of variety and yield for Grand Rapids in 1932. In Figure 3 the panel variables are variety, year, and yield and there is one conditioning variable, site.

In Figure 2 the descriptions of the values of the year and site for a panel are given in *strip labels* at the top of the panel. The strip labels for each variable have a dark bar that indicates the value of the variable. This conveys in a graphical way how the values of the conditioning variables are changing over the trellis.