

Trellis Displays

The "curse of dimensionality" as described by Huber [6] is not restricted to mathematical statistical problems, but can be found in graphic-based data analysis as well. Most plots like histograms or boxplots can only handle one single variable. Scatterplots can cope with two continuous variables, rotating plots with three. Mosaic plots can deal with several categorical variables.

Trellis Displays offer the possibility of combining up to eight variables in one plot panel in a matrix like manner.

The name of Trellis Displays derives from the trellis like (Lat. tri-liceum; a frame of latticework used for climbing plants) arrangement of the single plots.

Definition

Historical Development

Historically Trellis Displays are based upon the so-called *Co-Plots*. They were first mentioned by Chambers [4]. "Co" stands for "conditioning", which means that a specific plot is drawn for different subsets of a conditioning variable.

The generalisation to Trellis Displays is presented by William S. Cleveland [5], although he does not use the term there. Becker, Cleveland and Shyu [2], [3] describe the setup of Trellis Displays by means of an example.

A commercial implementation of Trellis Displays is available in S-Plus (version 3.3 and higher). An interactive version of Trellis Displays can be found in the MANET software (c.f. Unwin, [9]).

Technical Definition

The core of a Trellis Display is the one-, two- or three-dimensional graphics of the so-called *axis variables*. Those plots are designated as *panel plots*. The kind of plot is not limited any further; they can be dot-plots, scatter-plots, box-plots, surface-plots etc.

Besides the *axis variables*, there can be up to three *conditioning variables* chosen, to build the trellis. These are either categorical, or they have to be subdivided into several overlapping intervals. Those variables, called *shingles*, are then interpreted as categorical. For each category (or interval) or each combination of categories of the conditioning variables, the *panel plot* is drawn.

The plots are arranged in a vector (one conditioning variable) a matrix (two conditioning variables) or

a set/stack of parallel matrices (three conditioning variables).

Finally two more categorical or *shingle* variables, called *adjunct variables*, can be displayed. The different levels of adjunct variables are coded by different colours and markers, since the three natural dimensions might be already used by the conditioning variables.

The number of categories of the *conditioning variables* should not exceed eight, if the display is printed on an US-letter-size paper, to ensure a sensible size for each single plot panel. The limitation of categories with the *adjunct variables* is much tighter. Using more than three to four colours or markers would overtax the viewer of the plots, because he would not be able to discriminate between so many levels.

A critical point in the design of Trellis Displays is the scaling of each *panel plot*. Trellis Displays guarantee that all plots have the same scale, thus facilitating comparison of all plots.

Example

To explain the specific elements of a Trellis Display, the example of the Cars dataset, taken from the DataDesk datasets (Velleman, [10]), shall be shown here. Figure 1 shows a Trellis Display of scatterplots for Miles per Gallon (MPG) vs. Weight. The conditioning variables are chosen as Continent (derived from Country) and Cylinder (the only car with 5 cylinders has been added to a group of less than 6 cylinders). There are no *adjunct variables* present.

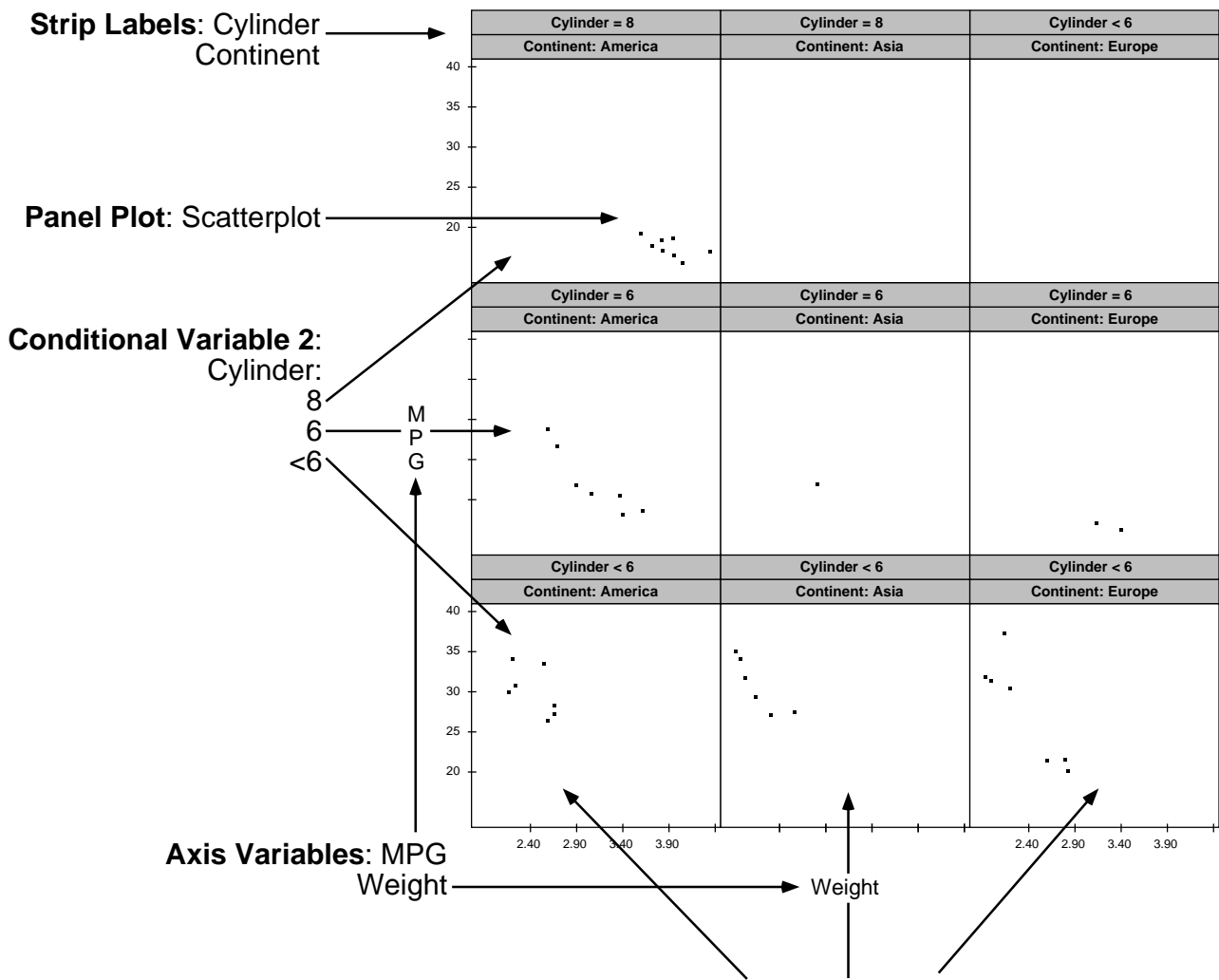
To optimise the Trellis Display the categories of the variable `Cylinder`, have been sorted by size. If a *conditioning variable* is not ordered, the categories should be ordered according to a statistic (mean, median, etc.) of one of the *axis variables*, to achieve a meaningful order of the *conditioning variable*.

Trellising: Which plot is best?

Trellising means the assignment of the different variables of a dataset to the different components of a Trellis Display. In particular this means to:

1. Specify the dimension of the trellis: columns, rows and pages
2. Specify the order of the conditioning variables
3. Specify the exact layout of the panels, i.e.
 - assignment of levels to rows, columns and pages
 - skipping of panels

In the above example it was easy to assign the roles of the four variables, the order of the levels and the layout of the panels. The results are not very



Conditional Variable 2: Continent: America, Asia, Europe

Figure 1: A sample Trellis Display with all elements

surprising even if the roles of `Continent` and `Cylinder` were exchanged. But this correct assignment was supported by a preceding knowledge of the data, and an insight into the underlying technical background.

A complete examination of the dataset would include all 7 variables, namely `Continent`, `MPG`, `Weight`, `Drive Ratio`, `Horsepower`, `Displacement` and `Cylinders`. Two of them are categorical the rest –5– are continuous. Not including shingle variables, this would lead us to ten (two chosen out of five) Trellis Displays, using scatterplots of the axis variables. Permuting the x- and y-role, as well as the assignment of the two conditioning variables, yields four times as many plots, giving 40.

Using shingling, there is no natural assignment of the variables any longer. This means, that we have $n!$ potential arrangements. For the Car dataset we get $7!=5040$ permutations resp. plots – which is more than a data analyst would like to consider.

Typically only a few assignments offer a relevant view of the data. But changing and understanding a specific assignment can be very hard.

Advantages: Modelling

Trellis Displays offer a systematic view of the different groups inside a dataset. This can be an excellent support for judging a statistical model. For this, the display can be extended with the model function, or the residuals can be plotted instead of the raw data. For an illustration we refer back to the car-data example. We extend the definition of the *panel plot* to a scatterplot with an additional superimposed scatterplot-smoother for the data in the panel. The resulting Trellis Display is shown in Figure 2. Note, that plot panels with too little data for fitting the model for that group do not show the superimposed model function.

Figure 3 shows the Barley data cf. [5]. Parallel boxplots are used to visualise the differences between the two years and the six locations. This example shows, how interactions can be visualised using Trellis Displays. Both factors, `Year` and `Site`, seem to be significant. The change in pattern is due to the location `Morris`, which points to an interaction of `site` and `year`.

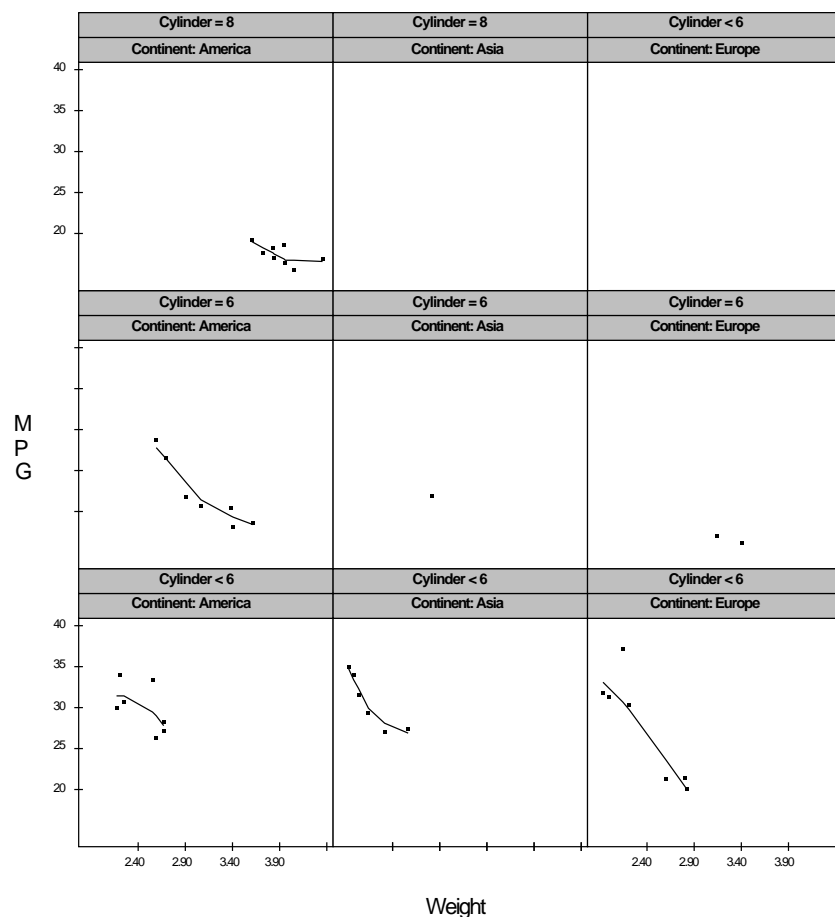


Figure 2: A scatterplot smoother superposed on all panels

Limitations: Categorical Data

A limitation with Trellis Displays is the fact, that all variables besides the axis variables must be categorical. To bypass this, shingling is recommended by Becker and Cleveland. As mentioned before, shingling means to subdivide a continuous variable into several overlapping intervals. This is usually done automatically with some control options.

But shingling is equivalent to the slicing process in an interactive graphical environment, when the user brushes continuously over the range of a variable, cf. Becker et al. [1]. In this situation plot panels correspond to snapshots taken during the brushing process. But in contrast to the interactive approach, the user has no visual control about an automatic shingling mechanism. This can be very misleading, because the individual structure of the shingled variable – e.g. gaps, ties etc. – are usually not considered by the automatic process.

There is another problem with shingling. How many observations are inside each plotpanel? In scatterplots e.g., there is the possibility of comparing the number of observations falling into the different intersections visually. But all plots that summarise the data, e.g. boxplots, hide the

actual amount of data being used to form the plot. This can be dangerous when judging optional model-fits in a Trellis Display. Comparing figure 2, an unusual smoothing line would be only interesting, if based upon a large amount of datapoints. For this it is necessary to have a clear knowledge of the underlying number of observations in each plotpanel.

Plotting categorical data as axis variables does usually not deliver sensible results. This is due to the great differences of the counts of different classes occurring in a categorical dataset. In this case Mosaic Plots seem to be the right choice for displaying purely categorical data.

Competitors

As mentioned above, interactive graphical methods offer the possibility of setting up single plot panels dynamically.

Selecting a particular subset in the conditioning variables delivers a highlighting of exactly this subgroup in the corresponding panelplot.

Using hot-selectors as described in [10], enables the analyst to set up exactly the panelplot shown in the corresponding Trellis Display, since only selected points are plotted.

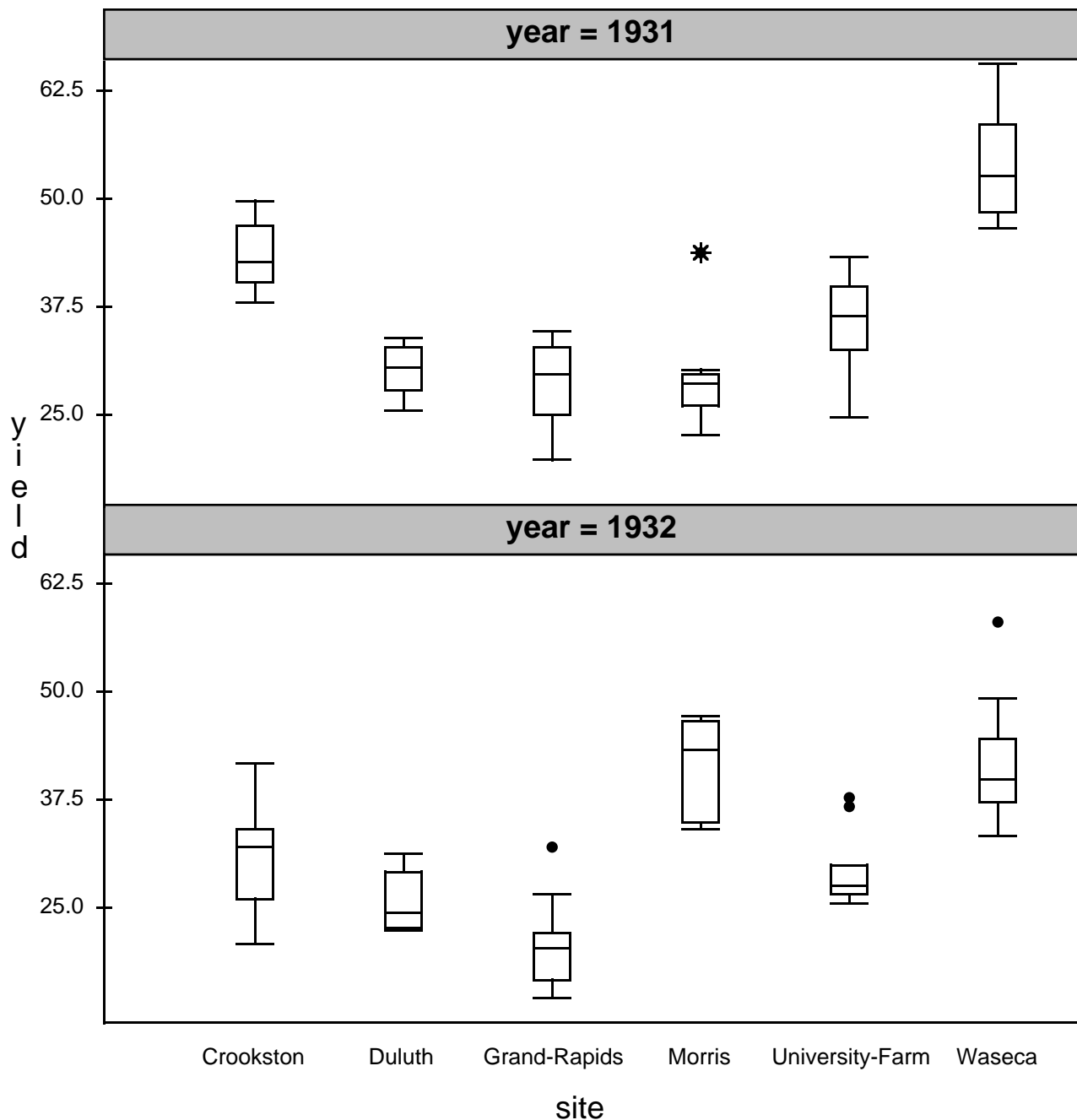


Figure 3: Parallel boxplots serve for judging ANOVA-models

Although these interactive approaches give the user much more control over the actual selected data, these techniques do not support a systematic plotting of all subsets.

An extension to standard selection techniques, the *selection sequences*, which support a systematic and hierarchical selection of data is discussed in [8] and implemented in the MANET software.

Conclusions

Trellis Displays offer a lot of useful extensions to standard static graphics. No other technique offers the possibility of setting up plots in a systematic way like Trellis Displays do. The Trellis-library inside S-Plus is the most complete and promising

implementation yet. But trellising and shingling should only be performed by an experienced data analyst, or at least need a solid knowledge of the dataset analysed.

This leads to the following use of Trellis Displays in statistics: first analyse the data and look for a suitable model with exploratory and interactive techniques, second present the model graphically with Trellis Displays, using the full range of plots and plotting options to achieve the best representation of the model.

References

- [1] BECKER, R. A., CLEVELAND, W.S., WILKS, A.R. (1987). Dynamic Graphics for Data Analysis. *Statistical Science*, 2, 355-395.
- [2] BECKER, R. A., CLEVELAND, W. S., SHYU, Ming-Jen, KALUZNY, St. P. (1994a) *Trellis Display: A Framework for Visualizing 2D and 3D Data* AT&T Bell Laboratories Statistics Research Report No. 8
- [3] BECKER, R. A., CLEVELAND, W. S., SHYU, Ming-Jen, KALUZNY, St. P. (1994c) *Trellis Display: User's Guide* AT&T Bell Laboratories Statistics Research Report No. 10
- [4] CHAMBERS, John M., HASTIE, Trevor J. eds. (1992), *Statistical Models in S* Wadsworth & Brooks/Cole, Pacific Grove CA.
- [5] CLEVELAND, William S. (1993) *Visualizing Data* Hobart Press, Summit NJ.
- [6] HUBER, Peter J. (1985) Projection Pursuit *The Annals of Statistics*, Vol. 13, pp. 435-475
- [7] THEUS, Martin (1995) Trellis Displays vs. Interactive Graphics. *Computational Statistics*, Vol. 10, No. 2, 112-127.
- [8] THEUS, Martin (1996) *Theorie und Anwendung Interaktiver Statistischer Graphik*. Wißner, Augsburg.
- [9] UNWIN, Antony R. (1996) *Interactive Graphics for Data Sets with missing values – MANET* Journal of Computational and Graphical Statistics, Vol. 4, No.3
- [10] VELLEMAN, Paul F. (1995) *Data Desk 5.0* Data Description, Ithaca, New York.