C4.1 Interactive Statistical Graphics

C4.1.1 History

Examining data using graphical tools, such as histograms, bar charts, quantile plots, scatterplots and other statistical graphics, is a necessary part of any serious analysis effort. Statistical graphics are typically most useful at two stages of an analysis. Initially, plots are useful for data cleansing [cross-ref. C2.1] where unusual sets of values can be investigated, attribute distributions can be evaluated and transformations can be suggested using graphical techniques. After a data mining model has been built, similar techniques can be used to validate the results of the model, highlight unusual or atypical results and present the model in a way that facilitates understanding [cross-ref. C8.1, C9]

With the advent of cheap graphics-capable desktop computing, these tools have become widely available. But the use of computers enables more than simply reproducing static plots on a display; it allows users to interact with plots, changing parameters, querying, zooming and linking plots together so that interesting features of one plot can be seen in the light of the others. Interactive statistical graphics explore data by manipulating a graphical representation of the data. Becker et al. (1987) define the process thus: "The data analyst takes an action through manual manipulation of an input device and something happens on the screen. These computing capabilities provide a new medium for the invention of graphical methods for data analysis".

The roots of interactive statistical graphics lie in Exploratory Data Analysis (EDA) as described by Tukey (1977): a set of techniques for investigating data to spot trends, patterns, errors and features. A key feature of EDA is the use of simple, robust plots to show characteristics of the data. According to Tukey, EDA is about "looking at data to see what it seems to say" (ibid., p. v). "It is detective work - numerical detective work - or counting detective work - or graphical detective work". (p. 1) and "Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider" (p. 3).

Interactive graphics (also known as Direct Manipulation Graphics or Dynamic Graphics) provide a satisfying extension of the principles of EDA to the computing environment. Many of the graphic facilities developed for exploratory data analysis are present in interactive graphic analysis, but in an enhanced form. Compared to the centuries old history of static graphics for data display, dynamic display and high-interaction techniques for the analysis of statistical data are a relatively recent phenomena. One of the earliest pieces of work in this area was PRIM-9, a system which enabled the user to view a multi-dimensional cloud of data and rotate it on the computer screen in real time (Fisherkeller, Friedman and Tukey, 1974; reprinted in Cleveland and McGill, 1988).. The system projected a multivariate data set into a three-dimensional point cloud and allowed the user to spin the cloud, manipulate the data visibility and alter the projection. PRIM stands for Picturing, Rotation, Isolation and Masking, the four major components of its operation In this early work, interaction was necessarily somewhat crude, being by means of an alphanumeric keyboard, light-pen and a function keyboard with 32 buttons, but the principle was of lasting importance.

In more recent years other interactive techniques such as dynamic scatterplot brushing have been introduced into statistical thought. Dynamic scatterplot brushing is a technique for visualizing several numerical variables by the creation of a matrix of scatterplots, one for each pair of variables in a data set. Then, using a pointing device, a small region (denoted a brush) is moved over one of the scatterplots, with the result that points that lie within this region are highlighted both in that scatterplot and in the other scatterplots which comprise the matrix. This early work in the field of interactive graphics focused on individual plots. The idea was to create unique views with which the user would interact to gain a more thorough understanding of the data. More recently, the emphasis has been on making interaction pervasive throughout the analysis process. Environments such as *Lisp-Stat* (Tierney 1990), *Data Desk* (Velleman 1988) and *XGOBI* (Swayne et al. 1991) have provided statistical environments where interaction is an integral part of the analysis. Furthermore, there has been much research into designing specific interactive systems to solve more narrowly defined problems. Examples of these systems include *REGARD* (Haslett et al., 1991) which applies these techniques to Geographic Information Systems, and *SeeSoft* (Eick et al., 1992) which applies them to the analysis of software change management data.

C4.1.2 General features of statistical graphical views

Since interactive graphics methods are mainly exploratory and diagnostic in intent, the graphical displays must be easy to interpret. Preferably, the results of an analysis should "jump out", as in the case of a bivariate outlier in a scatterplot. A good display will

- have an obvious encoding method,
- focus attention on the data, and
- integrate into the environment.

Note that the goals of interactive statistical graphics and presentation graphics are not the same; the former seeks to help the analyst discover, hypothesize and explore. The latter seeks to present a known result in the clearest possible manner. The characteristics listed above play mainly to the former goals. An analyst may look at many views without knowing whether there are any interesting features in them and so it is critical that views can be evaluated rapidly without requiring laborious deciphering. This motivates the desire for an intuitive encoding mechanism.

Achieving an *intuitive encoding* can be quite difficult. Suppose that we wish to encode a third variable on a scatterplot. Should we use a glyph whose size encodes the variable? Or should we use color? Or type of glyph (circle, square)? And if we decide on size, should the area encode the value or should the radius? Psychophysical experiments have provided much guidance in this area (see, e.g., Stevens 1975), and Cleveland (1985) has translated some important principles into the statistical graphics arena, but much care is still needed to ensure a clear, intuitive display.

Focusing attention on the data means that labels, grid lines and other additional view elements should be minimized. In an interactive setting, the user should be able to query the view for labels and values, often rendering the default display of labels and grid lines unnecessary.

Because interactive statistical graphics is an enabling technology, it is necessary that views be *integrated into an environment* in which they can add value to each other and to the overall analysis. The analyst should be able to interrogate views and see the results of either formal or informal queries (e.g. pointing at view elements or brushing data items) in all available views. In the next section we introduce a method of integrating views that has been used in a wide variety of contexts and is becoming a standard feature of modern interactive statistical graphic environments.

C4.1.3 The Linked Views Paradigm

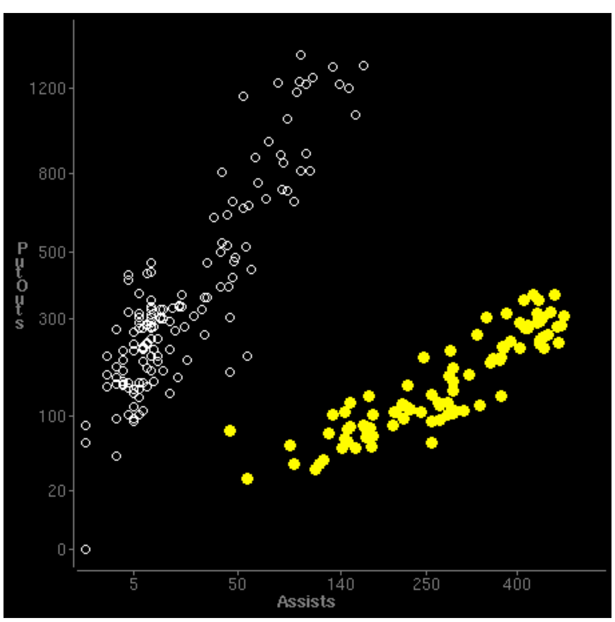
Linking is not a new concept, as can be seen from examination of the relevant literature (Becker et. al., 1987; Tierney, 1990; Velleman, 1988). For static graphics, a number of methods have been employed, possibly the most widespread of which is to split points into a number of classes and to assign a unique symbol or color to each class and use that symbol or color to draw points within that class within each plot that is displayed. An excellent example of this is shown in Chambers et al. (1983, p.172) where a data set of measurements on a set of flowers has been displayed via a scatterplot matrix of petal width, petal length, sepal width and sepal length.

The symbol used to plot each point in each scatterplot is a dot, circle or cross ('.', 'o' or 'x') depending on the variety of iris on which the measurements were made. This coding links the scatterplots together in a way which augments the natural linking caused by the juxtaposition of the scatterplots within the matrix structure. As an aside, note that Cleveland and McGill (1984) point out that plotting symbols of different colors provide the most acceptable coding mechanism, but that limitations on reproduction of color images can require the use of non-optimal encodings in black-and-white media. Linking is an intuitively attractive idea for increasing the information content of a set of data plots. Linking shows visually which parts of one data plot correspond to those of another. This allows the interesting or anomalous areas of one view of the data to be seen in the context of other views of the data in a rapid and intuitive fashion.

For data analysis one common method used to implement view linking is to associate a binary state with each data item, either highlighted or unhighlighted. The subset of data which is marked highlighted is assumed to be the focus of interest to the user. The user should be able to highlight a subset of data dynamically in one plot and see the results of that selection in another plot. The scatterplot brushing technique mentioned in C4.1.1 is a limited example of this method.

C4.1.4 Interactive Statistical Graphics Scenario

In this section we will look at a small data set which describes the career performances and salaries of baseball players in 1987. Our purpose is to use this data set to illustrate the utility of interactive statistical graphics at several stages of analysis.



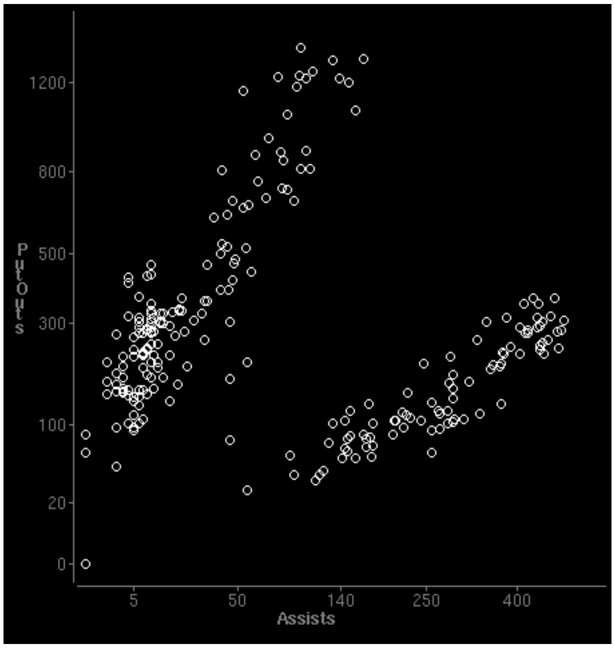


Figure C4.1.1 Scatterplot of Player's main fielding statistics.

If the analyst wanted to examine players' fielding performances, they might create a scatterplot such as figure C4.1.1. This figure shows the joint distribution of *PutOuts* (an opposition hitter is retired by the player) and *Assists* (where the player throws the ball to another player who performs a *PutOut*). The analyst has examined the distribution of both these variables in histograms or normal score plots and has determined that a square root transform should be applied to each variable (this makes sense statistically as both variables are counts). The resulting figure shows two distinct clusters, each with a strong linear pattern. The user then users the mouse to select the lower right cluster to highlight those items in it, and the resulting highlighting (figure C4.1.2) is examined in a bar chart of fielding position (figure C4.1.3). It is immediately apparent, without requiring any knowledge of baseball, that the relationship between *PutOuts* and *Assists* is conditional on the fielding position, with the positions *2B*, *3B*, *SS* having more assists than *PutOuts.* Note that *UT* denotes a utility fielder, one who might play anywhere, which explains why that bar is partially selected.

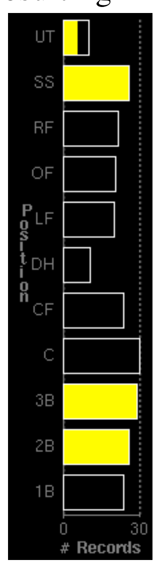


Figure C4.1.3 Linked fielding positions.

Following this analysis further, selecting one of the fielding positions from the bar chart, the analyst can see that the clusters themselves are composed of unions of even more linear features associated with each position. In figure C4.1.4 the analyst has selected two positions – one in each cluster – using the bar chart and has coded the number of *AtBats* (a measure of how much playing time each player had in the year) as color for the highlighted points. The color scale runs through shades of gray from black (few *AtBat*s) to white (many *AtBat*s). The unhighlighted points have been displayed as points to make them less noticeable in the scatterplot view. The analyst is now in a good position to propose a model; *PutOuts* and *Assists* have a linear relationship whose slope is determined by the fielding position, and players who have more playing time have increased counts of both *PutOuts* and *Assists*.

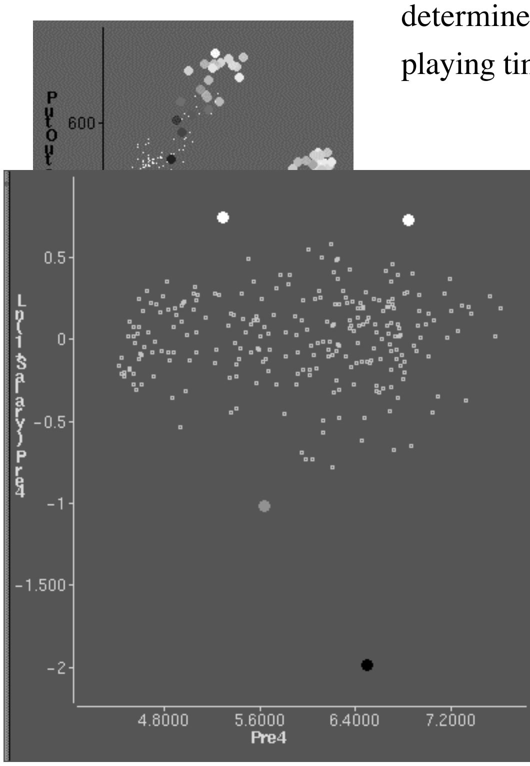


Figure C4.1.5. Prediction errors.

One of the original goals of the analysis of this data set was to explore the relationship between players' salaries and players' performances. Are players paid based on their ability? There are many measures of hitting performance which could be used in a model. For this data set, career averages of seven such measures (*Runs*, *Hits*, *Runs Batted In*, *Walks*, *Home Runs*, *AtBats*, *Years*) were used. These were used as inputs into a feed-forward neural net with one hidden layer containing four nodes and trained by back-propagation to predict a single output, a logarithmic transform of the salary. The resulting predictions for the training data were very good; probably due to over-training on this small data set of only 262 players. A scatterplot of prediction error against predicted value (figure C4.1.5) shows generally good shape with a few poorly predicted points and one extremely poorly predicted point towards the bottom of the plot. We have colored the highlighted values by the error term *Log(Salary)-Predicted*. In figure C4.1.6 we show a diagnostic plot to help us explore these outliers.

Figure C4.1.6 shows a vertical strip for each variable used in the analysis. The *Year* variable, shown on the far right, has been shown as a form of bar chart where circles are used instead of rectangles. The other variables are continuous and have been shown via a modified form of boxplot (as defined in Tukey, 1977). The dark rectangles show where 95% of the data values are expected to lie and the inner gray boxes contain exactly the middle 50% of the data, with the central line being the median. Note that these statistics are relatively robust. Superimposed on these plots we draw a line for each of our unusual players which links the player's statistics on each variable. By comparing the trajectories these take, we can immediately compare individuals' performances. When we select the outliers in figure C4.1.5, the corresponding players trajectories are shown on this linked display.

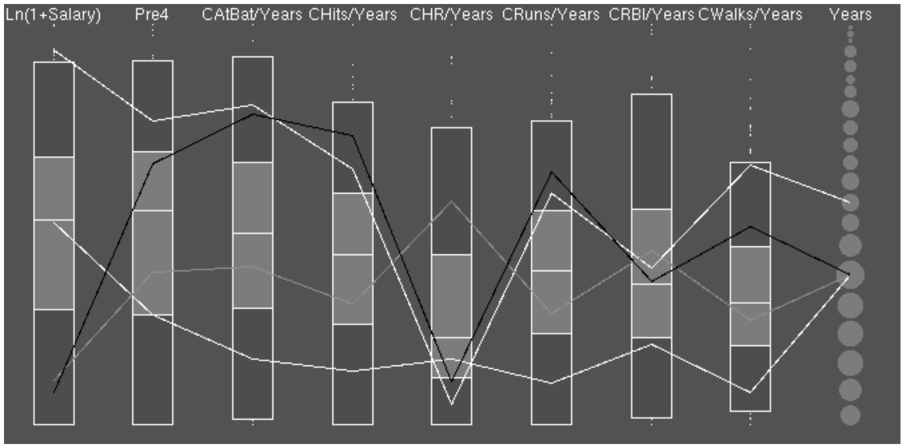


Figure C4.1.6 Parallel Axes superimposed on Boxplots and a bubble plot showing outliers in relation to the overall data trends.

In this example, if we examine the outlier with the highest salary amongst the outliers, shown in white, with the outlier with the least, shown in black, we see that they follow each other very closely except for *Years* (in which they differ by three) and in *Log(Salary)* (in which they are very different). These are players who have very similar hitting ability, have both been playing for a while and therefore our model predicts they have very similar salaries, whereas in fact one of them is paid $90,000 and the other $1,940,000. Clearly the model is not at fault; there must be other factors at work which may or may not be present in the rest of the data. The analyst can either accept the current model, noting that there is at least one exception to the model, or they can examine the data and maybe collect more data in an effort to expand the model and improve it for all cases.

C4.1.5 Challenges

Interactive statistical graphics are sets of general purpose views that allow users to look at data in different ways to explore features and focus attention on unusual items, both pre- and post-modeling. The roots in exploratory data analysis have given the analyst a set of standard plots such as histograms, boxplots, scatterplots and bar charts. These have been augmented and adapted for interactive environments so as to add to their utility, and a number of new views are making their way into more common usage. Current research challenges lie in two main directions. The first of these is in dealing with large data sets, both in terms of numbers of data items and numbers of variables. In gigabyte and terrabyte sized data sets, statistical significance becomes less and less useful a concept, and robust summaries are important. Moreover, small percentages of a data set represent sizable amounts of data and cannot be ignored by a technique or graphic display. When there are numerous variables, standard views can become overly complex, and views which summarize multiple relationships become necessary.

The second challenge is to build interactive graphical views for specific types of analysis; for geographic and spatial analysis, for time series, for neural net and decision tree analyses, for clustering, for document and text analysis, for real-time analysis, network analysis and many other emerging areas. The use of interactive statistical graphics techniques for database discovery will continue to grow in importance over time as these challenges are addressed.

References

Becker, R.A., Cleveland, W.S. and Wilks, A.R. (1987) Dynamic Graphics For Data Analysis *Statistical Science* **2** 355-395

Chambers, J., Cleveland, W.S., Kleiner, B., and Tukey, P. (1983) Graphical Methods for Data Analysis (Wadsworth, California)

Cleveland W.S. (1985) The Elements of Graphing Data (Wadsworth, California)

Cleveland, W.S. and McGill R. (1984) Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods *Journal of the American Statistical Association* **79** 531-554

Cleveland, W.S. and McGill R. eds. (1988) Dynamic Graphics for Statistics (Wadsworth and Brooks, California)

Eick, S., Steffan, J. and Sumner E. (1992) Seesoft - a Tool for Visualizing Line Oriented Software Statistics *IEEE Transactions on Software Engineering* **18/11** 957-968

Fisherkeller, M.A., Friedman, J.H., and Tukey, J.W. (1974) Prim-9: An Interactive Multidimensional Data Display And Analysis System *SLAC-Pub-1408* (SLAC Publications Office, Stanford, California)

Haslett, J., Bradley, R., Craig, P.S., Wills, G.; and Unwin, A.R. (1991) Dynamic Graphics For Exploring Spatial Data, With Application To Locating Global And Local Anomalies *American Statistician* **45** 234-242

Stevens, S. (1975) Psychophysics (Wiley, New York NY)

Swayne, D.F., Cook, D. and Buja, A. (1991) Xgobi: Interactive Dynamic Data Visualization In The X Window System *Journal of Computational and Graphical Statistics* **7** 113-130

Tierney, L. (1990) LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics (Wiley, NewYork, NY)

Tukey, J.W. (1977) Exploratory Data Analysis (Addison-Wesley, Reading, Massachusetts)

Velleman, P.F. (1988) The Datadesk Handbook (Odesta Corporation)