TECHNOMETRICS, MAY 1987, VOL. 29, NO. 2

Brushing Scatterplots

Richard A. Becker and William S. Cleveland

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

A dynamic graphical method is one in which a data analyst interacts in real time with a data display on a computer graphics terminal. Using a screen input device such as a mouse, the analyst can specify, in a visual way, points or regions on the display and cause aspects of the display to change nearly instantaneously. Brushing is a collection of dynamic methods for viewing multidimensional data. It is very effective when used on a scatterplot matrix, a rectangular array of all pairwise scatterplots of the variables. Four brushing operations—highlight, shadow highlight, delete, and label—are carried out by moving a mouse-controlled rectangle, called the brush, over one of the scatterplots. The effect of an operation appears simultaneously on all scatterplots. Three paint modes—transient, lasting, and undo—and the ability to change the shape of the brush allow the analyst to specify collections of points on which the operations are carried out. Brushing can be used in various ways or on certain types of data; these usages are called brush techniques and include the following: single-point and cluster linking, conditioning on a single variable, conditioning on two variables, subsetting with categorical variables, and stationarity probing of a time series.

KEY WORDS: Multidimensional data; Dynamic graphical methods; Scatterplot matrix; Computer graphics.

1. INTRODUCTION

1.1 High-Interaction Graphics

In the late 1960s and early 1970s, most statistical graphics was through systems that produced printer plots in batch output that took from 15 minutes to days to come back. The time and effort that it took for a data analyst to look at a graph, react to it, and change the graph to see information in a new improved way were extensive. This was indeed *low-interaction* graphics, or in many cases, *no-interaction* graphics. Many current statistical graphics systems have *medium-interaction* capabilities. A graph can be studied and a new one made within seconds or minutes. For example, in many current systems a data analyst can make a scatterplot of y_i against x_i, for $i = 1$ through n, decide that it is hard to see the form of the dependence of y on x, and issue a command to add to the graph a nonparametric regression curve of y on x to better see the dependence. Typically, the execution of the commands takes less than a minute.

The newest generation of computer graphics hardware and software has given us, at moderate cost, the capability to develop *high-interaction* graphical methods for data analysis. By using a screen input device such as a mouse, the data analyst can point to elements of a graphical display and manipulate them and can cause aspects of the display to change continuously in real time. It would be hard to overemphasize the importance of this new medium for data display, one which differs significantly from the traditional medium of static graphical displays that have been employed for statistical graphics in the past. The goal now is to invent graphical methods that exploit this high-interaction, dynamic medium and that are useful for data analysis.

1.2 The Evolution of High-Interaction Methods

Interactive computer graphics applications began to be developed in the early 1960s (Newman and Sproull 1979). The earliest dynamic graphical method in statistics of which we are aware is a probability plotting procedure developed by Fowlkes (1969, 1971). By turning a knob the user could continuously change a shape parameter for the reference distribution on a probability plot displayed on a screen; the knob could be turned to make the pattern of points as nearly straight as possible. The data could be transformed by a power transformation, $(y + c)^p$, with c and p each changed by a knob. Points could be deleted by positioning a cursor on them, and graphs could be saved on files and brought up later on the screen for further dynamic development. The system demonstrated that dynamic graphical methods had the potential to be important tools for data analysis.

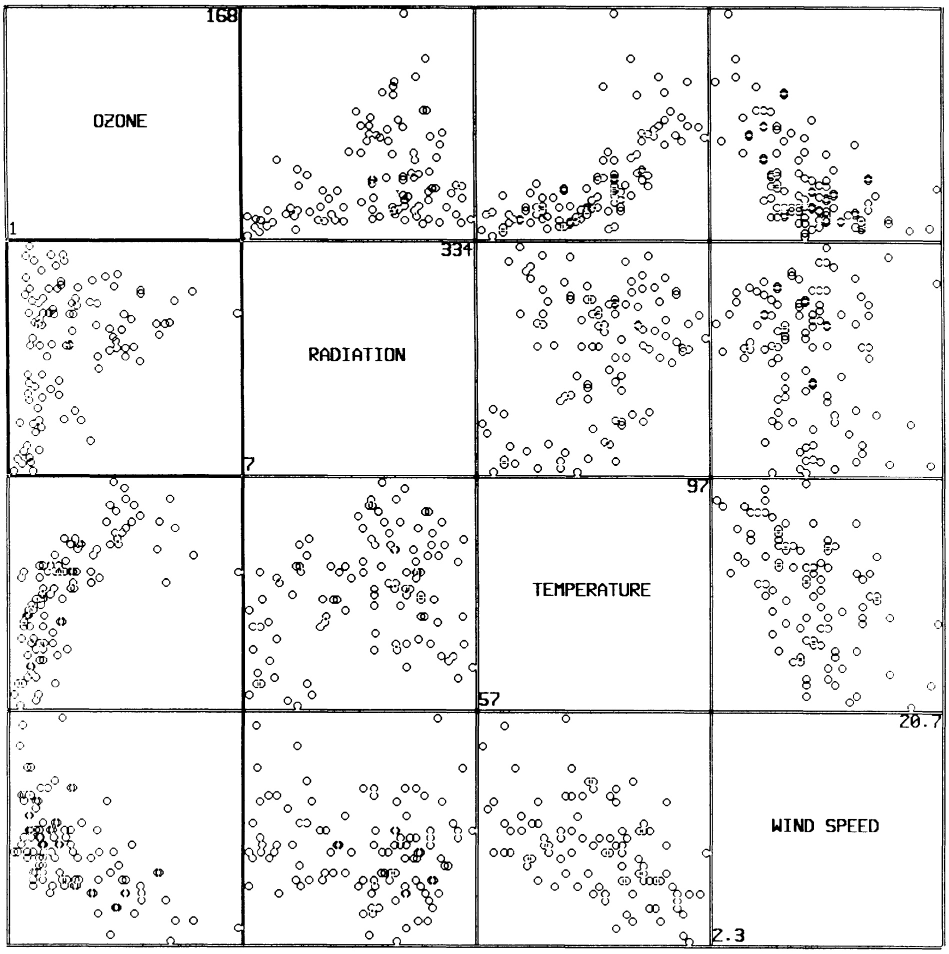

Figure 1. Ozone and Meteorological Data. A scatterplot matrix, all pairwise scatterplots in a rectangular array, shows four-dimensional data.

Another early system was PRIM9 (Fisherkeller, Friedman, and Tukey 1974), a set of dynamic tools for projecting, rotating, isolating, and masking multidimensional data in up to nine dimensions. The central operation was user-controlled rotation, which allows the data analyst to study three-dimensional data by showing the points rotating on the computer screen; this provides one type of three-dimensional scatterplot. PRIM9 also had isolation and masking, features that allowed point deletion in a lasting or in a transient way.

Dynamic methods were not widely used in the early years of their development, and invention moved ahead slowly, because special-purpose, expensive hardware was needed. Today, however, statistical computing environments have evolved to the point where the hardware and software to carry out dynamic graphics are available at low cost. This will be discussed further in Section 4.

1.3 Brushing

Brushing is a collection of new dynamic graphical methods for analyzing data in two, three, and higher dimensions. The data are displayed by a scatterplot matrix showing all pairwise scatterplots of the variables (Chambers, Cleveland, Kleiner, and Tukey 1983; Cleveland 1985). Figure 1 has an example; the data, which will be described in Section 2.4, are 111 measurements of four variables. The numbers in the corners of the boxes along the main diagonal in Figure 1 are the limits of the scales of the variables. For example, the scale for temperature goes from 57 to 97.

The central object in brushing is the *brush*, a rectangle that is superimposed on the screen, as illustrated in Figure 2. (The data in the figure, which will be described in Section 2.3, consist of 30 measurements of three variables.) The data analyst moves the

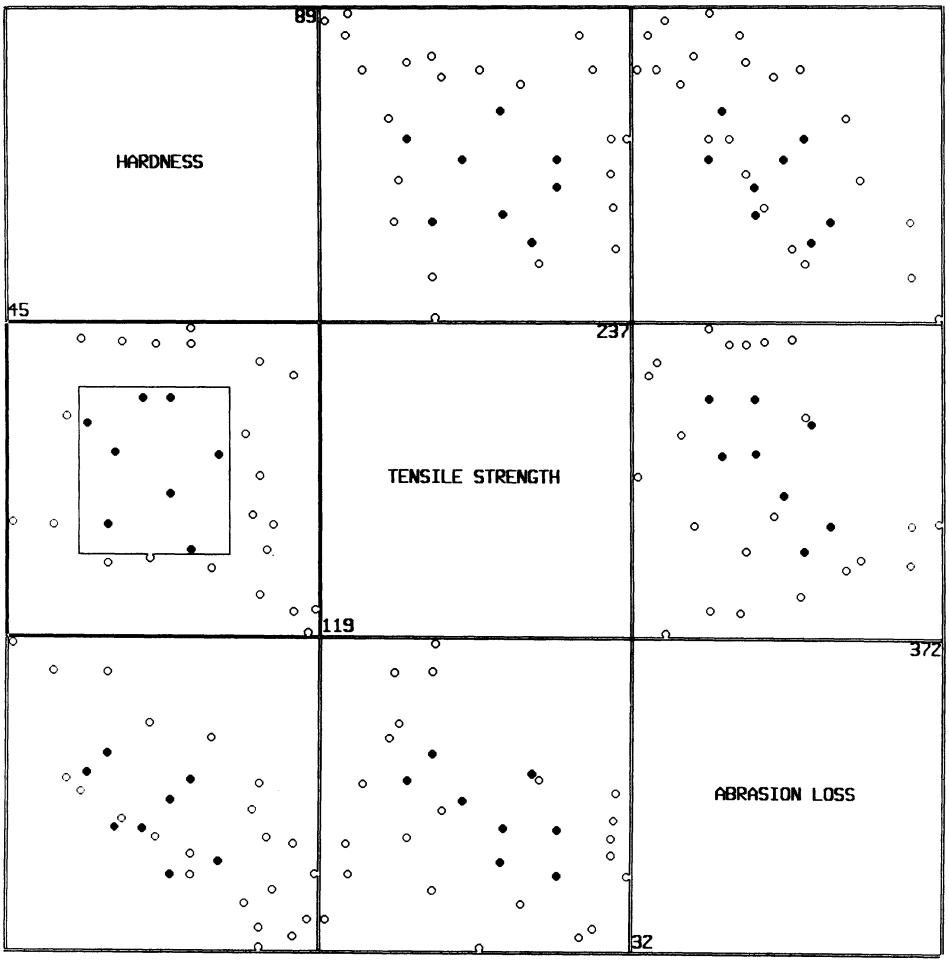

Figure 2. Abrasion Loss Experiment. The central object in brushing is the brush, the rectangle in the (2, 1) panel of the scatterplot matrix.

brush to different positions on the scatterplot matrix by moving a mouse. There are four basic *brushing operations*—highlight, shadow highlight, delete, and label. Each operation is carried out in one of three *paint modes*—transient, lasting, and undo. At any time, the data analyst can stop the brushing and change one or more of the features—the shape of the brush, the operation, or the paint mode—and then resume the brushing. The brushing methodology provides a medium within which a data analyst can invent data analytic methods, which we call *brush techniques*. This framework for invention is analogous to having, say, a Cartesian graph as the medium for inventing static displays such as normal probability plots, plots of residuals against fitted values (Anscombe and Tukey 1963), sharpening (Tukey and Tukey 1981, chaps. 10–12), and many other graphical methods (Chambers et al. 1983).

Thus the salient features of brushing are visual manipulation of the graph by the analyst, instantaneous change, the movement of the brush to different positions on the scatterplot matrix, the ability to change the shape of the brush, four operations, three paint modes, and brush techniques built up from the previous capabilities. In the next two sections, these features are described and illustrated by several applications.

2. THE HIGHLIGHT AND SHADOW HIGHLIGHT OPERATIONS, THE THREE PAINT MODES, AND SEVERAL BRUSH TECHNIQUES

2.1 The Highlight Operation

Figure 3 is a scatterplot matrix of 275 measurements of three variables that will be described shortly. On the (1, 2) panel, we can see that for low values of $\log_2 n$ there are several outliers—values of $\log_2 w$

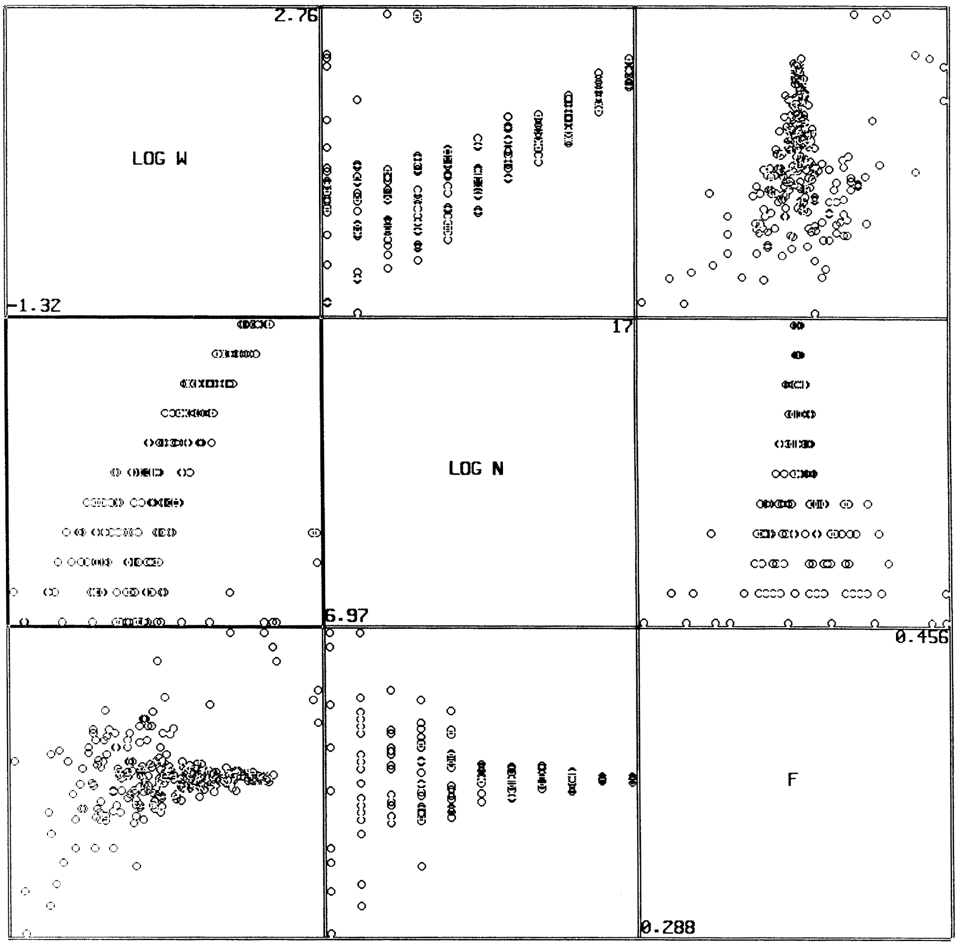

Figure 3. Bin Packing Data. The scatterplot matrix shows three-dimensional data. It is not possible to visually link all of the outliers in the upper left corner of the (1, 2) panel to the corresponding points in the (1, 3) panel.

that are large. Consider the three largest outliers, which occur at the third and fourth values of $\log_2 n$; by scanning across to the (1, 3) panel, we can see that these three outliers have large values of f. Consider, however, the next five largest outliers, which occur at the two lowest values of $\log_2 n$. When we scan across to the (1, 3) panel, we cannot tell whether these five outliers correspond to points with high values of f or to points lying in the dense spire in the middle of the (1, 3) panel. We cannot *link* the points across the two panels.

The operation *highlight* allows us to solve the linking problem and to carry out a number of other brush techniques. Highlight is illustrated in Figure 2. One panel of the scatterplot matrix is selected as the *current panel*. All points that are both inside the brush and inside the current panel are highlighted, as are the points on other panels that correspond to these points.

The data of Figure 3 are measurements of a computer bin packing algorithm (Bentley, Johnson, Leighton, and McGeogh 1983). First, n weights were generated as random numbers, uniform on the interval from 0 to .8. The weights were ordered from largest to smallest and then packed, in order, into bins of size 1. Bin 1 is tried; if there is no room, bin 2 is tried, and so forth until a bin with enough room for the weight, perhaps a completely empty bin, is found. As many bins are available as are needed to pack the weights. The data in the figure are for 275 independent runs of the bin packing; n was set at 11 different values ranging from 125 to 128,000 and there were 25 runs for each value of n.

The variable w (for white space) is the amount of empty space in bins with at least one weight. The (1, 2) panel in Figure 3 shows that $\log_2 w$, not surprisingly, increases as $\log_2 n$ increases. But in addition the variability decreases, as does $\log_2 n$. What causes these outliers? The mathematical theory of bin packing suggested that the weights above .5 might be the cause; for this reason f, the fraction of weights above .5, was computed for each run and

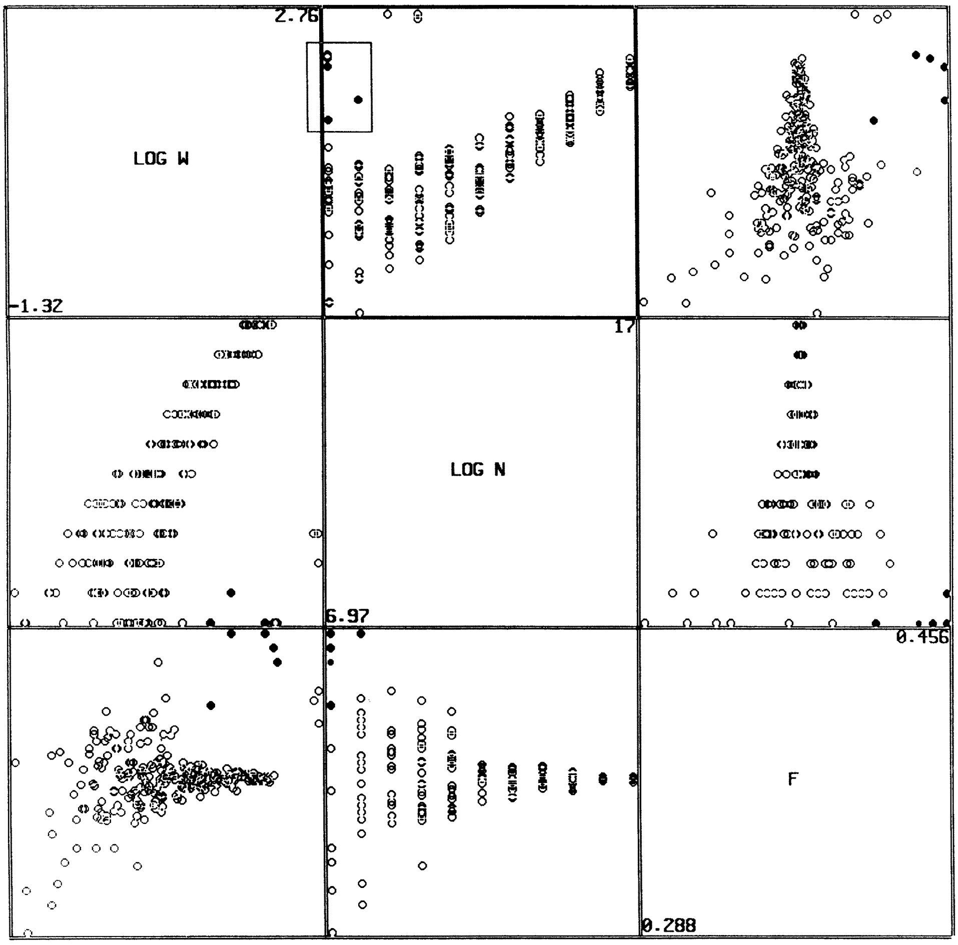

Figure 4. Bin Packing Data. The highlight operation is being used; all points inside the brush are highlighted, as well as points on the other panels that correspond to these points. Now we can see the points of the (1, 3) panel that correspond to the outliers of the (1, 2) panel.

added as an explanatory variable. In Figure 4 the outliers on the (1, 2) panel for the first two levels of $\log_2 n$ have been highlighted by brushing. Now we can see clearly from the (1, 3) panel that these outliers, as well as the three largest, do indeed correspond to high values of f and are not in the dense spire. The (2, 3) panel in Figure 4 provides further insight: The law of large numbers insists that f not vary by much from its expected value of .375 when $\log_2 n$ is large; thus f cannot, with nonnegligible probability, be large and cause outliers in $\log_2 w$.

2.2 The Transient Paint Mode

When the brush is in the *transient* paint mode, the following happens: As the brush is moved, the new points that come inside the brush are highlighted, but points no longer inside the brush are no longer highlighted. This is illustrated in Figures 5–7; the brush starts out on the left of the active panel, moves to the center, and then moves to the right.

2.3 Conditioning on a Single Variable

Suppose the goal is to see how one variable depends on the others. This is the case for the data in Figures 5–7, which are measurements of tensile strength, hardness, and abrasion loss (the amount of material rubbed off by an abrasive material) for 30 rubber specimens (Davies 1958, p. 210). The purpose of the experiment that generated the data was to determine how abrasion loss depends on tensile strength and hardness. In the original analysis, abrasion loss was modeled as a linear function of hardness and tensile strength.

Conditioning on a single variable is a brush technique that provides a simple, direct way to fix the values of one variable to a certain range and see the behavior of the other variables. This technique is achieved by selecting the highlight operation and the transient paint mode and by making the brush long and narrow. This is illustrated in Figure 5, in which

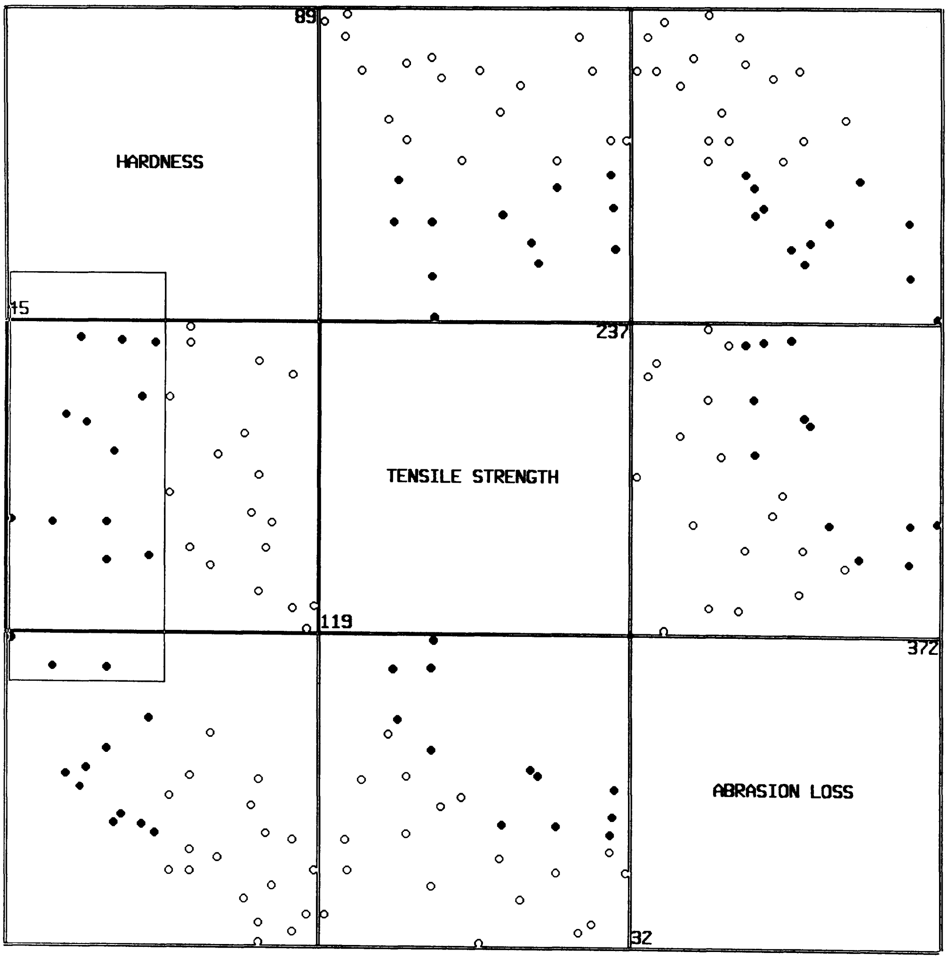

Figure 5. Abrasion Loss Experiment. Figures 5–7 illustrate conditioning on a single variable. In this figure, low values of hardness are highlighted by the brush in the (2, 1) panel. The highlighted points of the (3, 2) panel show the dependence of abrasion loss on tensile strength for low values of hardness.

the brush has been shaped and positioned in the (2, 1) panel so that all points with low values of hardness are highlighted. Brushing only affects the active panel, so the large brush can be used without fear that points on surrounding panels will be affected. The highlighted values on the (3, 2) panel are a plot of abrasion loss against tensile strength for hardness held fixed to low values; the configuration of points looks something like a hockey stick. In Figure 6 the brush has been moved to the right on the (2, 1) panel, so we are now conditioning on middle values of hardness, and the (3, 2) panel again shows a hockey stick effect. In Figure 7 we have conditioned on the high values of hardness; the hockey stick effect on the (3, 2) panel is still apparent, although there appear to be one or two unusual points in the lower left corner. Thus this brushing analysis suggests that there is a nonlinear dependence of abrasion loss on tensile strength and that the linearity of the original model is inadequate. These suggestions have been confirmed through fitting a regression model that is nonlinear in tensile strength.

2.4 Conditioning on Two Variables and the Shadow Highlight Operation

We can condition on two variables by making the brush a square and moving it around an active panel that graphs the two variables against one another. We will illustrate this technique using the variables of Figure 1. The data are 111 daily measurements of ozone, temperature, wind speed, and solar radiation at various sites in the New York City metropolitan region (Bruntz, Cleveland, Kleiner, and Warner 1974). (A table of the data was given by Chambers et al. 1983, pp. 347–349.) The goal of the analysis is to see how ozone depends on the other variables. Solar radiation is certainly an explanatory variable, since it is a requisite for the photochemical reactions that produce ozone. The (1, 3) panel in Figure 1 shows that ozone and temperature measurements are corre-

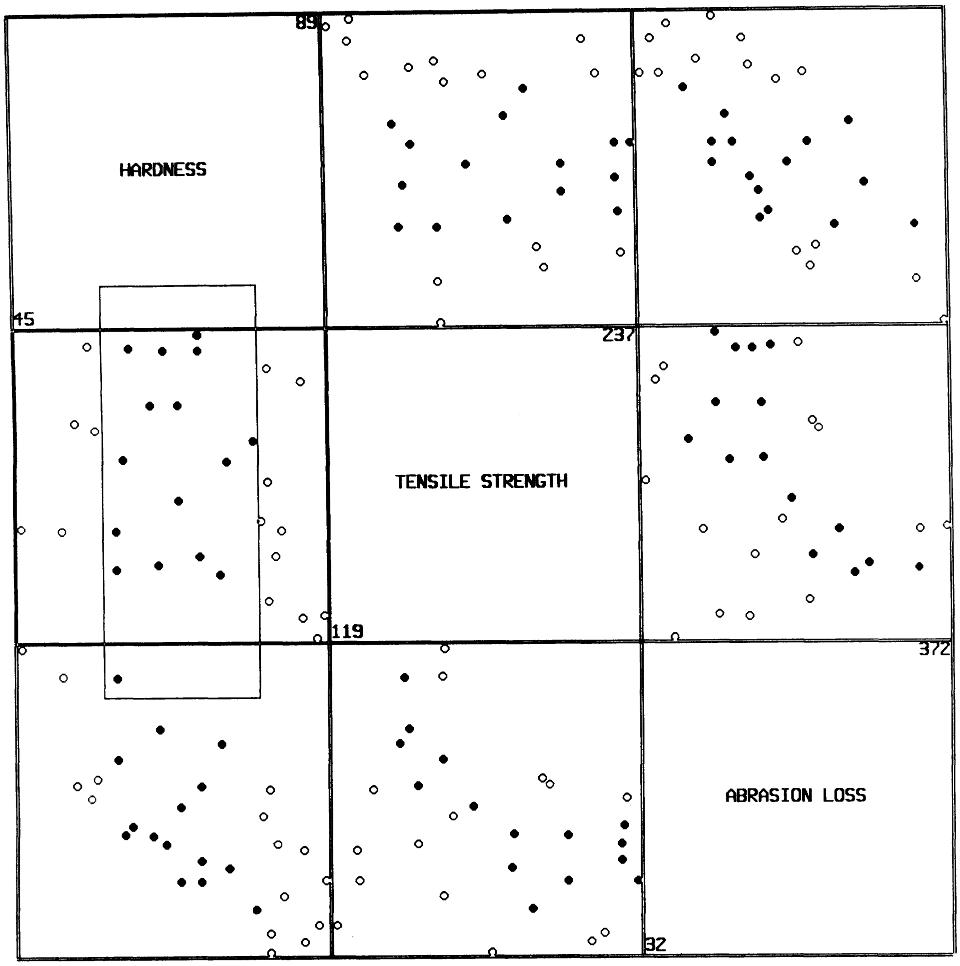

Figure 6. Abrasion Loss Experiment. The highlighted points of the (3, 2) panel show the dependence of abrasion loss on tensile strength for middle values of hardness.

lated, but photochemical theory suggests that temperature itself is not responsible for greater ozone generation, since the rates of the reactions change negligibly over the range of temperatures in the data set. Temperature, however, is correlated with the meteorological stagnation that can lead to high ozone; when stagnation is high, temperature tends to be high. For example, in the (3, 4) panel of Figure 1 we can see that temperature tends to be high when wind speed is low.

Is temperature a necessary surrogate for stagnation in this data set, or is there little explanatory power beyond solar radiation and wind speed? Figure 8 helps us to see. We have conditioned on both wind speed and solar radiation in the (2, 4) panel. The operation is now shadow highlight. This behaves like highlight except that only the points that are highlighted on the active panel appear on the other panels. The principal use of shadow highlight is in situations in which, because there are a large number of points, highlighted points cannot be easily visually detected because of overlap. This is the case for the data of Figure 8.

Figure 8 suggests that temperature is needed. Panel (1, 3) shows a relationship between temperature and ozone. The (1, 2) and (1, 4) panels show no relationship; this assures us that the temperature variable provides explanatory power beyond that which is provided by wind speed and solar radiation. When we move the brush around the (2, 4) panel, a similar temperature-ozone relationship always appears, but when the brush is in a region of low solar radiation the temperature-ozone slope tends to decrease, which suggests a temperature-radiation interaction.

2.5 Subsetting Categorical Variables

Often, multidimensional data can be broken into groups, which can be described by a categorical variable. We can incorporate such a categorical variable

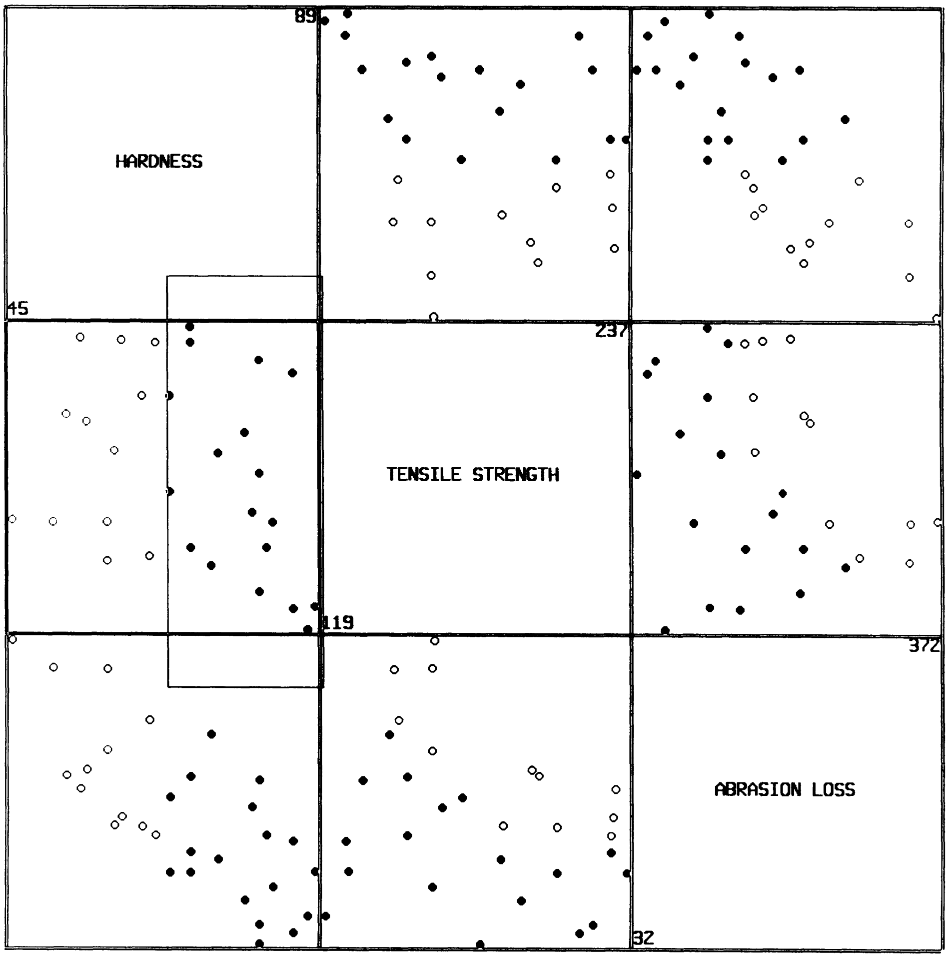

Figure 7. Abrasion Loss Experiment. The highlighted points of the (3, 2) panel show the dependence of abrasion loss on tensile strength for high values of hardness.

by defining a numerical variable that takes on values from 1 to k, where k is the number of groups, and adding it to the list of variables shown by the scatterplot matrix; this enables us to study the structure of each group.

An added categorical variable is illustrated in Figure 9. The data are from a survey of graphs in scientific publications (Cleveland 1984). One quantitative variable is the base 10 logarithm of the ratio of legend area to graph area for 50 scientific journals; another quantitative variable is the base 10 logarithm of the fraction of space devoted to graphs (not including legends) in 50 scientific journals. The second variable is a measure of the amount of use of graphs in a journal, and the first variable is a rough measure of the average amount of explanation a graph receives in the legends. The four journal categories are as follows:

1. Biological—biology, medicine
2. Physical—physics, chemistry, engineering, geography
3. Mathematical—mathematics, statistics, computer science
4. Social—psychology, economics, sociology, education.

A categorical variable was defined by assigning one of the numbers 1–4 to each journal, according to its category, and then adding a random number from the interval from −.25 to .25. This *jittering* (Chambers et al. 1983) is done to alleviate overlap, but in such a way that each journal's category can be correctly identified. The jittered categorical variable is included in Figure 9. The brush on the (1, 3) panel is highlighting the first group, which is biological journals. We can see that these journals have a greater amount of explanation than the other categories and that the amount of explanation per unit area of graphs for this category does not depend on the fraction of space devoted to graphs. Using the brush, we can quickly shift to the other journal groups and also study their behavior.

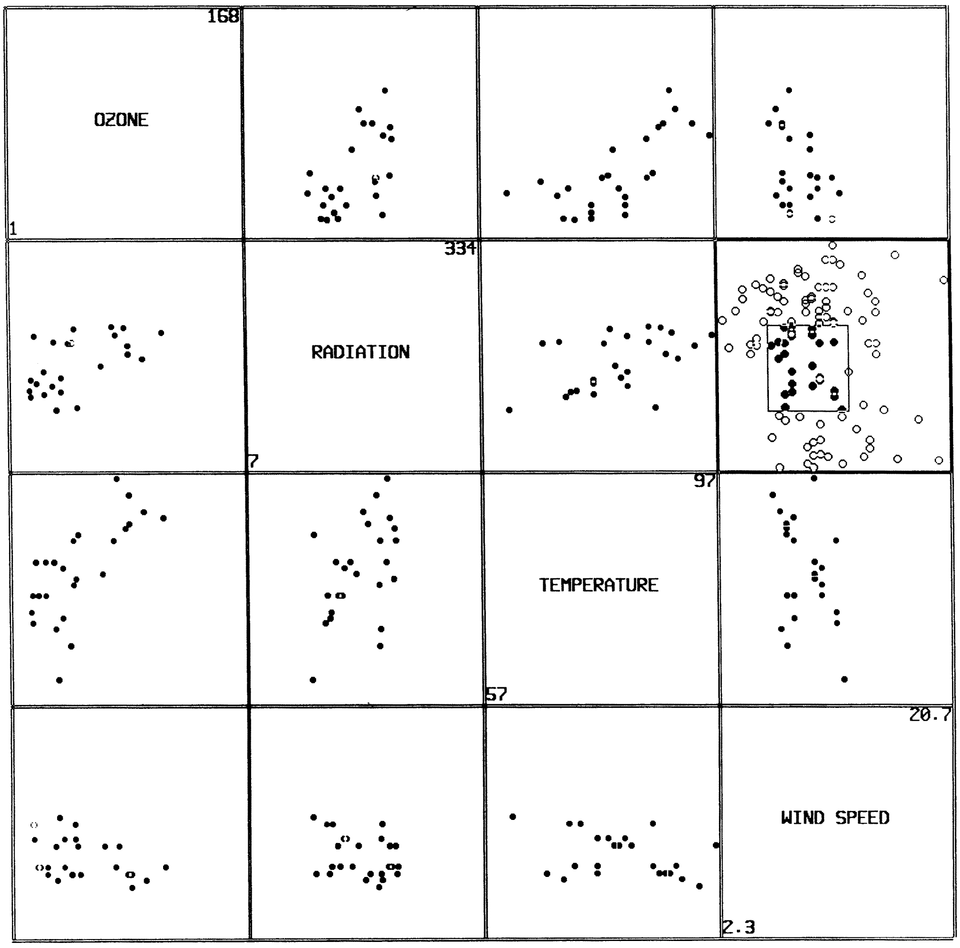

Figure 8. Ozone and Meteorological Data. The figure illustrates conditioning on two variables and the shadow highlight operation. The points of the (1, 3) panel show the dependence of abrasion loss on temperature for radiation and wind speed held fixed to the ranges determined by the brush in the (2, 4) panel.

2.6 The Lasting and Undo Paint Modes and Irregular Cluster Linking

When the paint mode is transient, we can quickly move from highlighting one rectangular region to highlighting another; this is important for the three brushing techniques discussed in the previous three sections. But how can we perform cluster linking when the cluster region on a panel is irregular and cannot be enclosed in a rectangle without including other points? In other words, how can we highlight an arbitrary subset of points on the panel?

We solve this problem of irregular cluster linking by the *lasting* paint mode; points inside the brush remain highlighted even after the brush no longer covers them. (In our implementation, the analyst invokes the lasting paint mode by pressing button 1 of a 3-button mouse.) With this capability we can highlight an irregular region much as we would paint it; this is illustrated in Figure 10 for jittered values of the famous four-dimensional iris data (Anderson 1935). A subset of points has been highlighted by painting the cluster in the upper left of the (1, 2) panel, which is a plot of sepal width and length. Now we can readily see from panel (3, 4) that this corresponds to the cluster of flowers with large petal lengths and widths.

The *undo* paint mode can be used to remove highlighting. When the brush is in this mode, any lasting highlighting inside the brush is undone, so if we moved the brush over the highlighted cluster in Figure 10 in this mode, we would remove the highlighting. (In our implementation the analyst invokes the undo paint mode by pressing button 2 of the mouse.) This undo operation allows us to separately study a variety of clusters.

2.7 Other Variables in the Scatterplot Matrix

It is important to emphasize that variables other than the original measurements can be displayed by

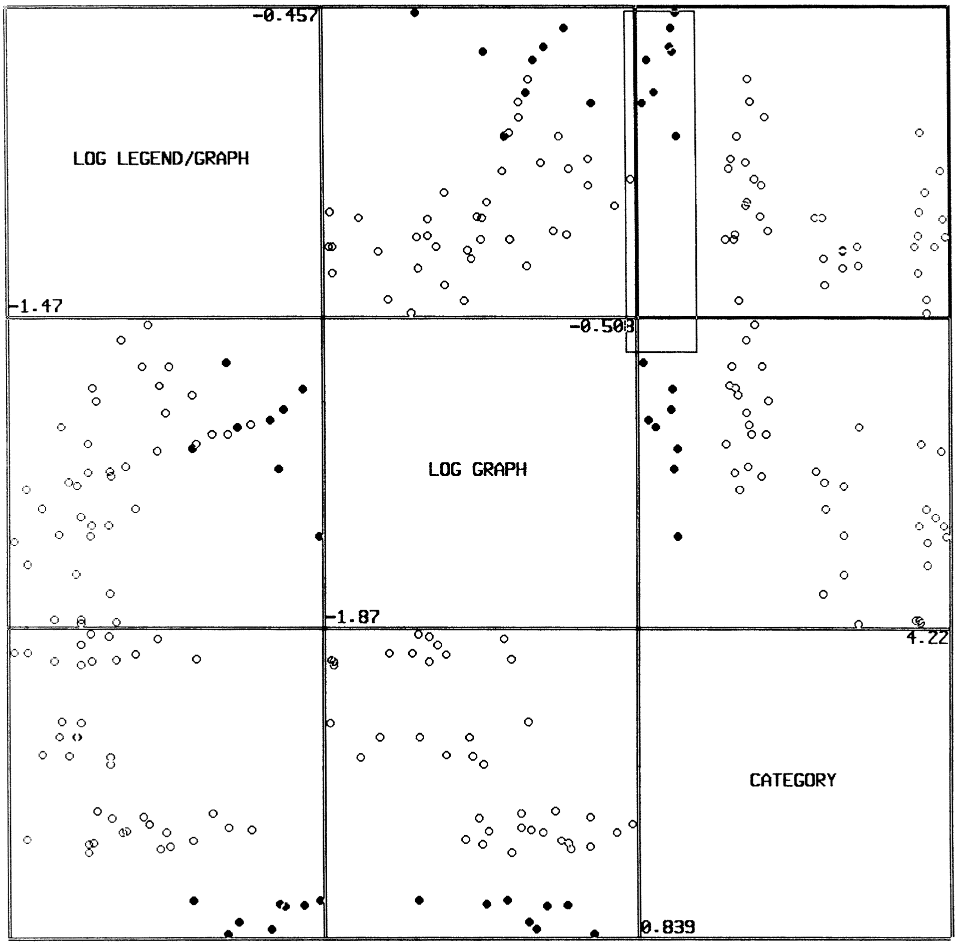

Figure 9. Journal Data. Brushing can be used to show a breakup of the data into subsets by a categorical variable. In this example the brush is highlighting the observations in the first of four categories.

the scatterplot matrix and brushed. For example, if we are involved in a regression, putting residuals in the scatterplot matrix can be helpful. One possibility is to replace the dependent variable by the residuals from regressing the dependent variable on all variables; in this case the variables of the matrix would be the residuals, together with the explanatory variables. If some explanatory variables have been decided upon as important and necessary in the regression and r additional variables are under consideration, then the variables of the scatterplot matrix can be the set of residuals from regressing the dependent variable on the entered explanatory variables and r sets of residuals from regressing each of the r not-entered variables on the entered ones.

3. THE DELETE AND LABEL OPERATIONS AND MORE BRUSH TECHNIQUES

3.1 The Delete Operation

Another brushing operation is *delete*. It works in the same way as highlight except that points are deleted instead of highlighted. The transient, lasting, and undo paint modes also operate in the same way; button 1 makes the deletion lasting rather than transient, and the deletion can be undone by holding down button 2 and painting the region of deleted points.

3.2 Stationarity Probing of a Time Series

In the usual approach to analyzing a time series, $x(t)$, the autocovariance is estimated as if stationary; that is, the autocovariance

$$c(k) = \operatorname{cov}(x(t), x(t + k))$$

is assumed independent of t. But in real life $c(k)$ can vary appreciably through time. It can slowly drift or, if $x(t)$ has a strong seasonal component, $c(k)$ can be cyclic.

Suppose we want to investigate the autocovariance at lag l. Brushing can be used to probe nonstationarity by defining three variables: $x(t)$, $x(t - l)$, and $d(t)$, where $d(t)$ might be t if we want to check for

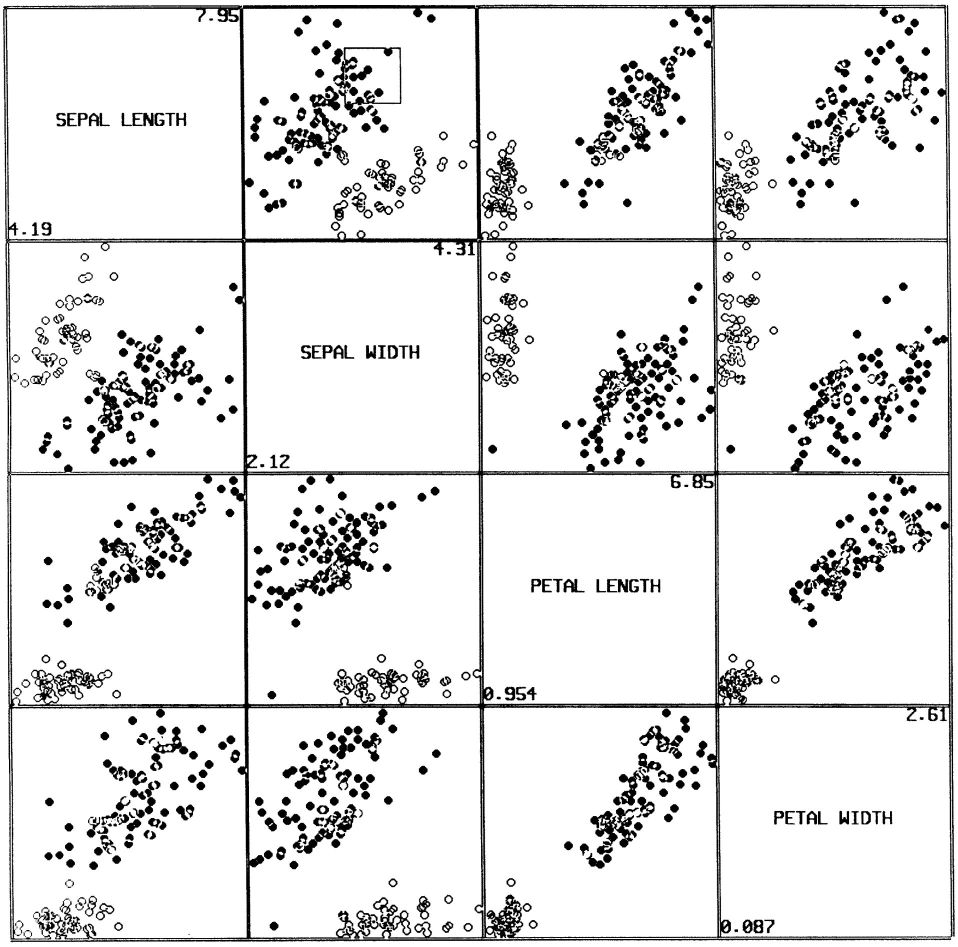

Figure 10. Iris Data. Brushing with the highlight operation and the lasting paint mode (points once highlighted remain highlighted) has been used to highlight the points of an irregular cluster in the (1, 2) panel. The lasting highlighting can be removed by the undo operation.

drift, or $d(t)$ might be t mod p if we want to check for a cycle of period p in $c(k)$.

Figure 11 is a scatterplot matrix in which the three variables are $x(t)$, $x(t - 2)$, and t. The (1, 2) panel is a graph of $x(t)$ against $x(t - 2)$, and the (1, 3) panel is a graph of $x(t)$ against time. The time series is the residuals of monthly atmospheric CO_2 concentrations (Keeling, Bacastow, and Whort 1982) after a seasonal component and trend component, computed using the SABL procedures (Cleveland, Devlin, and Terpenning 1982), were subtracted. The (1, 2) panel suggests that there is correlation at lag 2; indeed, the autocorrelation coefficient at lag 2 is .44.

Brushing, however, shows that the autocorrelation at lag 2 is not a stationary phenomenon. This is illustrated in Figure 12. Using the delete operation, we have removed all data from a narrow time slice; with these points removed, the autocorrelation disappears. Figure 13 shows a shadow highlight; it is clear that the points in the narrow time slice have a high lag 2 autocorrelation and that they are the sole contributor to the autocorrelation of all of the data.

3.3 The Label Operation

Very frequently in the analysis of measurements of two or more variables, each observation of the variables has a name, or label, associated with it. An example is shown in Figure 14. The data are values of four variables for the 48 contiguous U.S. states (*The World Almanac* 1983, pp. 116, 126)—per capita income in 1975 and 1979, per capita state government debt in 1980, and per capita state taxes in 1980. In this case each observation (the four values of the variables) has a label, the name of the state.

In most situations in which the observations are labeled, we find ourselves wanting to know some or all of the labels of the points on graphs portraying the data. Usually, however, we cannot display all

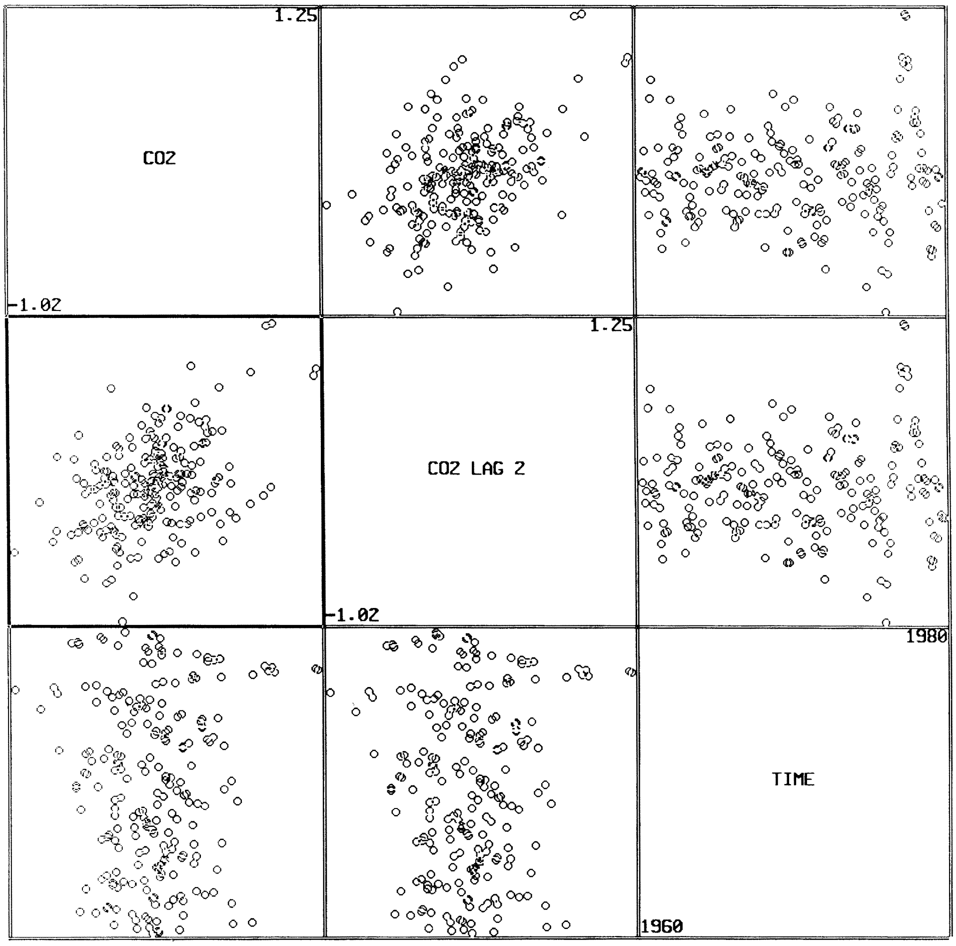

Figure 11. CO_2 Data. The three variables in this scatterplot are $x(t)$, $x(t-2)$, and t for a time series, x. The (1, 2) panel suggests that $x(t)$ and $x(t-2)$ are positively correlated.

labels simultaneously, since the overlap of the labels produces an uninterpretable mess; this would be the case for the graph in Figure 14.

Brushing, however, provides a convenient mechanism for studying the labels, because we can use the brush to turn labels on and off. If the *label operation* is used, the effect of brushing is to cause the names to appear, as in Figure 14. As with the highlight, shadow highlight, and delete operations, label can be used in the transient, lasting, or undo paint modes; brushing in the lasting mode can be used to turn on a collection of labels and brushing in the undo mode can be used to turn them off.

4. DISCUSSION

4.1 Linking

There is a long history of attempts to design static graphical displays that allow the data analyst to carry out linking of corresponding points on different scatterplots (Chambers et al. 1983). In fact, the scatterplot matrix is a static method that arose, in part, because it provides partial linking.

The first high-interaction method for linking of which we are aware was developed by McDonald (1982). In his method a cursor is moved around one plot. Points close to the cursor on the plot are red, points intermediate in distance are purple, and far points are blue. All points corresponding to the same observation on any other scatterplots have the same color. McDonald did not apply this method to the scatterplot matrix, although one could, of course, do so. This method of linking gives quite striking results in certain cases, however, the method does not allow the highlighting of many regions that are of interest in data analysis—for example, irregular clusters, such as the one in Figure 10, and the long and narrow rectangular regions of conditioning on a single variable. The aspects of brushing that provide the needed generality are the highlight operation, the three paint modes, and a rectangular brush whose shape can change.

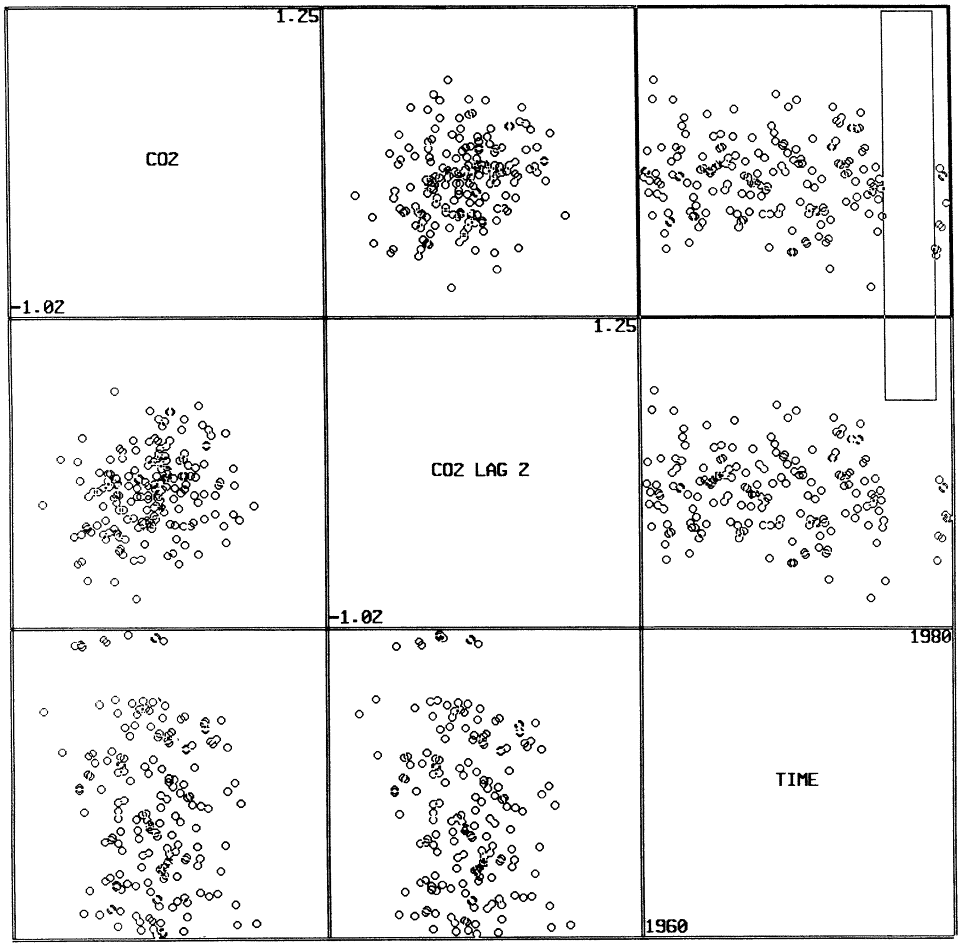

Figure 12. CO_2 Data. The points in a narrow time interval have been removed by the brush in the (1, 3) panel and the delete operation. Now the points in the (1, 2) panel appear to have no correlation.

4.2 Other Methods and Extensions

Brushing is just one of many dynamic methods that are available to data analysts. Others are reviewed by Becker, Cleveland, and Wilks (in press). Moreover, brushing can be extended to more general configurations of graphical displays than the scatterplot matrix. For example, one can brush a scatterplot and have the result appear also on a plot of three variables shown by rotation (Becker, Cleveland, and Weil, in press; Stuetzle, in press) or on histograms of the variables (Stuetzle, in press).

Brushing by itself is not, of course, a sufficiently rich environment for data analysis. For the methodology to be helpful in analyzing data, it must be embedded in a software environment that includes data management and standard statistical methodology, both graphical and numerical. For example, our implementation is within the S system for data analysis and graphics (Becker and Chambers 1984) and the Wang and Gugel (in press) implementation is callable from SAS (SAS Institute 1985). Several other dynamic capabilities can enhance brushing as a data-analytic tool. For example, after points have been deleted it is often helpful to rescale the scatterplots of the matrix so that the remaining points fill up the plotting regions of the panels. It is also helpful to have a locate capability, which is the opposite of label; a name is chosen from a pop-up menu (controlled by the mouse) and the points of the scatterplot matrix that have this name are highlighted (Becker and Cleveland 1984).

4.3 Brushing and Point-Cloud Rotation

An important consideration is how brushing a scatterplot matrix compares with rotation, since both are oriented toward analyzing multidimensional data. Our experience in applications is that rotation is better suited for certain tasks and brushing is

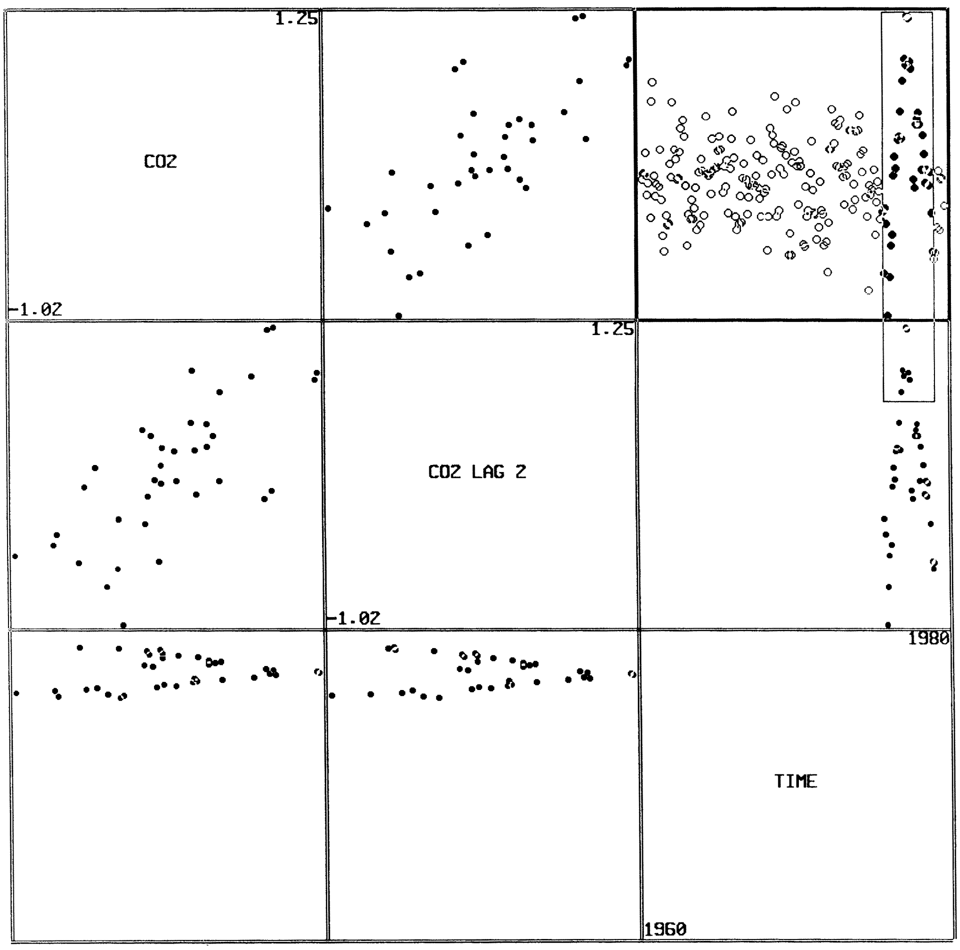

Figure 13. CO_2 Data. The same time interval has been selected by the brush as in Figure 12, but now the operation is shadow highlight. The points in the (1, 2) panel have a strong correlation. Thus the deletion and shadow highlight of Figures 12 and 13 show the correlation in $x(t)$ and $x(t - 2)$ results only from the data in a narrow time interval.

better suited for others. Thus both are worthwhile tools to have available for data analysis.

Suppose there are three variables; that is, the data are in three dimensions. Point-cloud rotation can be useful for locating groups of points that cluster. Many clusterings are relatively easy to discover with rotation but not with brushing. One example is the *randu* random numbers; detecting their planar clusterings (Marsaglia 1968) appears impossible with brushing. If, however, the goal is to determine how one variable depends on the other two, brushing is usually far more informative than rotation. One reason is perceptual. Perceiving dependence on a two-dimensional scatterplot can be difficult; it is even more difficult on a three-dimensional point cloud because our perception of depth is not nearly as effective as our perception of position along the horizontal or along the vertical. Furthermore, it is quite easy to become befuddled in viewing a rotating point cloud and lose a sense of the axes. Another reason for the better performance of brushing is that the brushing technique of conditioning on a single variable is such a powerful method for understanding dependence in three-dimensional data.

4.4 Computer Implementation

Brushing can be implemented in many different computing environments. Our original implementation was on an AT&T Teletype 5620 graphics display terminal together with a host machine running under the control of the UNIX™ operating system. Details of this implementation were given by Becker and Cleveland (1984). Since then, we have implemented brushing on an IRIS graphics workstation (Becker, Cleveland, and Weil, in press), Tierney (personal communication, 1986) has implemented it on a MacIntosh™, Stuetzle (in press) has implemented it on a Symbolics workstation, and Wang and Gugel (in press) have done a less interactive color

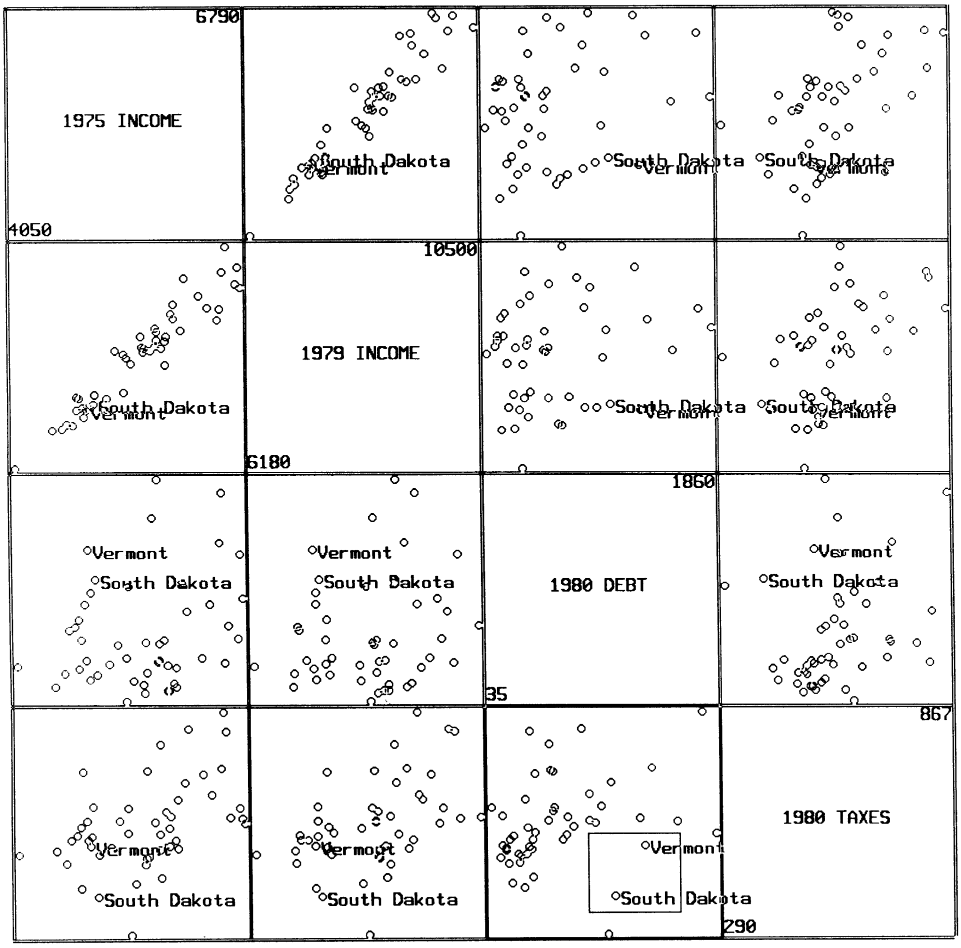

Figure 14. State Data. The label operation is being used. The two points inside the brush on the active panel have their labels displayed as well as corresponding points on other panels.

implementation using the DI-3000™ graphics system.

In developing dynamic graphics software it is important to maintain uniformity in the operation of the mouse. Without general concepts that cut across many operations, the data analyst can become confused by the mouse buttons. For example, in our implementation the operation of the four brushing operations is the same; with no button depressed brushing is transient, with button 1 depressed it is lasting, and with button 2 depressed it is undone.

Space is a problem for the scatterplot matrix on the computer screen; many plotting symbols need to be squeezed into a small area. To give the matrix as much room as possible on the screen, we put the variable names inside the main-diagonal panels, instead of around the sides; we omitted tick marks and put extreme scale values in the corners of the main-diagonal panels; and we squeezed the panels as close as possible, allowing just enough space to have good visual separation of the panels.

ACKNOWLEDGMENTS

Jon Bentley, John Chambers, Brian Kernighan, and Alan Wilks gave us many helpful comments on an earlier version of this manuscript.

[*Received November 1985. Revised October 1986.*]

REFERENCES

Anderson, E. (1935), "The Irises of the Gaspe Peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

Anscombe, F. S., and Tukey, J. W. (1963), "The Examination and Analysis of Residuals," *Technometrics*, 5, 141–160.

Becker, R. A., and Chambers, J. M. (1984), *S: An Interactive Environment for Data Analysis and Graphics*, Monterey, CA: Wadsworth/Brooks Cole.

Becker, R. A., and Cleveland, W. S. (1984), "Brushing a Scatterplot Matrix: High-Interaction Graphical Methods for Analyzing

Multidimensional Data," technical memorandum, Murray Hill, NJ: AT&T Bell Laboratories.

Becker, R. A., Cleveland, W. S., and Weil, G. (in press), "The Use of Brushing and Rotation in Data Analysis," in *Dynamic Graphics for Statistics*, eds. W. S. Cleveland and M. E. McGill, Monterey, CA: Wadsworth/Brooks Cole.

Becker, R. A., Cleveland, W. S., and Wilks, A. R. (in press), "Dynamic Graphics for Data Analysis," *Statistical Science*.

Bentley, J. L., Johnson, D. S., Leighton, T., and McGeogh, C. Cole (1983), "An Experimental Study of Bin Packing," in *Proceedings, Twenty-First Annual Allentown Conference on Communication, Control, and Computing*, Urbana-Champaign, IL: University of Illinois Press, pp. 51–60.

Bruntz, S. M., Cleveland, W. S., Kleiner, B., and Warner, J. L. (1974), "The Dependence of Ambient Ozone on Solar Radiation, Wind, Temperature, and Mixing Height," in *Symposium on Atmospheric Diffusion and Air Pollution*, Boston: American Meteorological Society, pp. 125–128.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Monterey, CA: Wadsworth/Brooks Cole.

Cleveland, W. S. (1984), "Graphs in Scientific Publications," *The American Statistician*, 38, 261–269.

——— (1985), *The Elements of Graphing Data*, Monterey, CA: Wadsworth/Brooks Cole.

Cleveland, W. S., Devlin, S. J., and Terpenning, I. J. (1982), "The SABL Seasonal and Calendar Adjustment Procedures," in *Time Series Analysis: Theory and Practice 1*, ed. O. D. Anderson, Amsterdam: North-Holland, pp. 539–564.

Davies, O. L. (ed.) (1958), *Statistical Methods in Research and Production*, London: Oliver & Boyd.

Fisherkeller, M. A., Friedman, J. H., and Tukey, J. W. (1974), "PRIM9: An Interactive Multidimensional Data Display and Analysis System," in *Data: Its Use, Organization, and Management*, New York: Association for Computing Machinery, pp. 140–145.

Fowlkes, E. B. (1969), "User's Manual for a System for Interactive Probability Plotting on Graphic-2," technical memorandum, Murray Hill, NJ: Bell Telephone Laboratories.

——— (1971), "User's Manual for an On-Line Interactive System for Probability Plotting on the DDP-224 Computer," technical memorandum, Murray Hill, NJ: Bell Telephone Laboratories.

Keeling, C. D., Bacastow, R. B., and Whort, T. P. (1982), "Measurement of the Concentration of Carbon Dioxide at Mauna Loa Observatory, Hawaii," in *Carbon Dioxide Review: 1982*, ed. W. C. Clark, New York: Oxford University Press, pp. 377–385.

Marsaglia, G. (1968), "Random Numbers Fall Mainly in the Planes," *Proceedings of the National Academy of Sciences*, 61, 25–28.

McDonald, J. A. (1982), "Interactive Graphics for Data Analysis," technical report, Stanford University, Dept. of Statistics.

Newman, W. M., and Sproull, R.F. (1979), *Principles of Interactive Computer Graphics*, New York: McGraw-Hill.

SAS Institute, Inc. (1985), *SAS User's Guide: Basics*, Cary, NC: Author.

Stuetzle, W. (in press), "Plot Windows," *Journal of the American Statistical Association*, 82.

Tukey, P. A., and Tukey, J. W. (1981), "Graphical Display of Data Sets in 3 or More Dimensions," in *Interpreting Multivariate Data*, ed. V. Barnett, Chichester, U.K.: John Wiley.

Wang, C. M., and Gugel, H. W. (in press), "High-Performance Graphics for Exploring Multivariate Data," in *Dynamic Graphics for Statistics*, eds. W. S. Cleveland and M. E. McGill, Monterey, CA: Wadsworth/Brooks Cole.

World Almanac and Book of Facts, The (1983), New York: Newspaper Enterprise Association.