APPENDIX A

LITERATURE REVIEW

LITERATURE REVIEW

Purpose and Structure

The term “scientific visualization” was coined by a panel of the Association for Computing Machinery (ACM) organized by the National Sciences Foundation’s Division of Advanced Scientific Computing (McCormick et al, 1987). Though the term became very popular, this tool is, by no means, completely new. Artists and scientists have been using visual aids for thousands of years. The history of discovery and invention, to some extent, can be viewed as a history of visualization (Robin, 1992). Nevertheless, the increasing power of modern computing makes visualization more accessible to many researchers. While researchers in the physical and engineering sciences have dealt with increasing data complexity by using scientific visualization, researchers in the behavioral sciences have been slower to adopt this tool (Butler, 1993).

To address this discrepancy, this review has the following goals:

1. To clarify misconceptions about visualization;

2. To define scientific visualization;

3. To introduce a taxonomy of multivariate visualization techniques;

4. To explain how the choices of visualization techniques are dictated by research goals and data types.

It is argued that successful visualization is a result of the proper alignment of graphical format, research goal/task type, and data type. This idea guides the discussion that follows and will be referred to as the alignment framework. A number of multivariate visualization techniques will be reviewed in light of the three taxonomies of graph, task and data. A diagram of the alignment framework is presented in

Figure 1.

Misconceptions of Visualization

A vague conceptualization of a discipline would mis-direct discussion and create unnecessary debate. Therefore, before defining what visualization is, I would like to clarify two common misconceptions about visualization.

Visualization and Computer Graphics

Many times the terms scientific visualization and computer graphics are used interchangeably though they are fundamentally different. Several scientists have emphasized a basic distinction between scientific visualization and computer graphics: Computer graphics aim to simulate a photo-realistic image, whereas scientific visualization often deals with abstract data. (cited in Wolfson, 1994). In other words, visualization does not necessarily produce a realistic image. Instead, the goal of visualization is to filter data noise. Sometimes a graph as simple as a histogram is sufficient to represent the data.

Further, computer graphic designers and animators devote effort to make electronic imitations of a known object or to actualize a planned draft. On the other hand, data visualizers usually do not know how the graphical representation of the data would turn out. Because of this data-driven nature, scientific visualization is also named “data visualization.”

Visualization and Multimedia

According to Ross (1993), visualization and multimedia are sometimes confused as well. Visualization is a process in which data is converted to an image for easier interpretation and analysis. Multimedia, on the other hand, should properly be used as an adjective such as a multimedia document. Used as a noun it describes a particular technique for the presentation of information involving a combination of video, audio, text, and so on. Although multimedia can enhance visualization, a visualization process does not necessarily involve multimedia.

What is Visualization?

In this paper I define scientific visualization as the process of exploring and displaying data in a manner that builds a visual analogy in the service of researcher insight and learning (Yu & Behrens, 1995). This entails finding a balance between the detail of the raw data and the parsimony of statistical summary. Each component of this definition is now addressed.

Visualization as Data Exploration

Visualization emphasizes data exploration, the methodology in which its origin can be traced back to the schools of Exploratory Data Analysis (EDA) and Total Quality Management (TQM), especially the former. Both schools of thought, to a large extent, are balanced against the school of Confirmatory Data Analysis (CDA), which is the intellectual inheritance of Sir R. A. Fisher and others.

Most statistical training in psychology focuses on CDA (cf. Aiken, West, Sechrest, & Reno, 1990). CDA is focused on hypothesis testing and probabilistic inference to populations. The emphasis of test statistics over graphical tools has become the paradigm of data analysis since the establishment of the Fisherian tradition. While commenting on the use of diagrams, Fisher (1932) said:

> Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data. (cited in Scott, 1992, p.4)

Pioneered by the work of John Tukey (1977, 1980, 1986a, 1986b, 1986c, 1988) and promoted by Cleveland (1985, 1993) and others, the tradition of EDA emphasizes seeking unexpected structure and developing rich descriptions through graphic summary, robust statistics, and model fit indicators. This approach contrasts the common methods of hypothesis testing that assumes an underlying structure to the data and proceeds with inference on the basis of such assumptions as normality, homogeneity, and independence. Since EDA emphasizes pattern-seeking and skepticism, EDA allows the researcher to use whatever tool is helpful. Graphical aid is an indispensable tool in EDA because data patterns can be unveiled by graphical manipulation (cf. Behrens, 1995; Behrens & Smith, in press). As Cleveland (1985) said, “The data...look innocent enough until they are graphed” (p.97).

TQM also uses graphical aids heavily but for a very focused purpose--statistical quality control. Although Edwards Deming, the founder of TQM, built his work on the Fisherian tradition, many of the Deming’s tools are indeed exploratory in essence (Schwabe, 1991). Deming’s inclination to exploratory statistics is manifested by his opposition to hypothesis testing and his heavy use of graphical exploratory tools. In Deming’s view, the goal of research on quality is to identify the pattern of changes and reduce this variation, which is on the contrary to the emphasis of centrality in many statistical textbooks. Therefore, Deming considered hypothesis testing is one of the evils taught in statistics courses (Boardman, 1994).

However, after four decades of criticism, the ritual of hypothesis testing still persists despite the availability of alternative tools such as confidence intervals, good enough principle, and graphical EDA (Cohen, 1994). In an article entitled *A Picture is Worth a Thousand p Values*, Loftus (1993) observed that even today journal editors do not accept the results reported in mere graphical form. Test statistics must be provided for the consideration of publication. Loftus asserted that “hypothesis testing, as it is customarily implemented, ignores two issues that are generally much more interesting, important and relevant: What is the pattern of population means over conditions, and what are the magnitudes of various variability measures?” (p.250)

Writing in a tone consonant with Tukey, Deming, and Loftus’ ideas, Cleveland (1993) argued that "visualization is an approach to data analysis that stresses a penetrating look at the structure of data" (p.5). Such work is deemed especially important when examining multi-dimensional data for which algebraic summaries are often difficult to interpret.

Visualization as Analogy Making

Visualization of phenomenon in the physical sciences is often striking because of the similarity of the computer-generated images to our expectation of how the process being model should look. At the same time, there are many physical systems that offer no direct visual analogy. For example, scientists at the Netherlands Research Foundations visualized multi-dimensional phenomena of 3-D flows of fluid dynamics (Hesselink, Post, & Wijk, 1994) for which there was no obvious physical analogy. Even though a single vector can be represented by an arrow, no compelling physical metaphor exists for a field of vectors and a tensor, the product of vectors. No wonder Keller and Keller (1993) argued that in scientific visualization "choosing techniques to represent the phenomenon may require some creative or artistic talent, especially if the phenomenon is abstract or has never been seen, such as the inside of a proton, or a black hole" (p.12).

The difficulty of lacking direct analogy also occurs in the visualization of the psychological sciences, because psychologists deal with abstract concepts such as cognitive models and personality traits. However, several researchers still conceptualize graphical information as some kind of “one-to-one-mapping” between the symbols and the world. For example, they prefer plots of raw data to plots of transformed data because the latter is not as “real” and “direct” as the former. This view overlooks the fact that raw data produced by various measurement methods such as Likert-scale, IQ test, MPPI, and many others are also man-made representations of an object of study. These analogies are not in any sense more direct than the analogy based upon transformed data. In short, all visualization is based on analogy and all analogy is incomplete. Visualizers should put aside the mental block of direct analogy and be creative in visualization as Keller and Keller suggest.

Visualization as Balancing Noise and Smooth

Data analysis is a process of reducing large amounts of information to parsimonious summaries while remaining accurate in the description of the total data. This often requires a balancing act between presenting masses of data that may be incomprehensible to the viewer, or presenting summaries that average over too many details of the original data. Visualization seeks to meet this challenge by portraying complex data in interpretable means so that aspects of both the messiness and smoothness of data can be discerned.

In this definition of visualization I have stressed the notion of exploration, analogy and balancing summary with raw data. This definition is broad and includes numerous statistical and other graphics. In order to make sense of the plethora of graphics available to psychologists for multivariate visualization, I present a framework for understanding graphics based on the idea of balancing summary with raw data. Following mathematical and statistical terms we discuss this as the balance of noise and smooth.

Bandwidth. The concepts of noise and smooth are perhaps best understood by using the well known histogram. The appearance of a histogram is largely controlled by the number of bars used to depict the data. When many bars are used the pattern of the data may look jittery as shown in the last histogram of Figure 20. Here the details are great and the reader may wonder if a simpler underlying form exists. On the other hand, the use of too few bars may obscure patterns in the data that are important to the viewer as illustrated in first histogram of Figure 20. In this case the summarization is great but the reader may wonder if some important detail is missed. The central panel of this figure presents an intermediate number of bars. In this view balancing smooth and noise is essentially balancing summary and raw data.

Structure Imposition. Another factor determining the noise level of a graph is the degree of data structure imposed by the data analyst. For example, a regression line summarizes the relationship between the variables and seeks to minimize residuals, but it assumes homogeneity of variance and linearity of the bivariate relationship. This assumption of structure may be very inappropriate in the early phases of data exploration. On the other hand, some procedures may be too flexible in that they overfit the data and inappropriately suggest structure that is unique to a sample. Just as in the case of balancing noise and smooth, balancing the imposition of structure versus the use of flexibility involves the subjective judgment and expectation of the data analyst.

Building on these ideas, graphical techniques can be conceived as occurring in the two-dimensional space of smoothness/noise and dimensionality of the data being depicted. Table 12 orders a number of statistical graphics along these dimensions. The horizontal dimension of smoothness/noise is conceived as a continuum, while the vertical dimension of variable dimensionality is conceived as discrete. Ranging from one- to multi-dimensional graphing techniques, there exist the tensions between much data and less data, and between little imposed structure and more imposed structure.

Visualization as an Interaction of Graph, Task, and Data Types

Conceptualizing graphical formats in terms of noise and smooth can help us to align research goal and data type with the proper visualization technique. In a lecture concerning multivariate data analysis, Holmes (1994) correctly pointed out that the choice of multivariate procedures depends on the aim of the analysis and the data available. The application of her notion is suitable to visualization, too. For instance, in some conditions such as spotting outliers with large amounts of data, a nosier plot showing all data points is helpful. In some situations such as examining relationships with four variables, a smoother graph depicting functions is desirable.

In order to understand how these three aspects of visualization interact with each other, taxonomies of the three aspects are constructed. In the following section, a number of statistical graphics meant to display one-, two-, three- and higher-dimension data are reviewed in light of their noise versus smooth characteristics. Their specific applications are also mentioned according to proper research goals and data types. Afterwards, suitable graphical tools for six major categories of research goals will be demonstrated. Last, the relationship between data type and graph type will be discussed with reference to graphs introduced in the previous two sections.

Variations in Graphics for Visualization

One-dimensional Graphs

The histogram is perhaps the most common graphic for displaying the distribution of a single variable. While constructing a histogram seems to be straightforward, the appearance of the histogram is arbitrarily tied to the interval width used as shown in Figure 20. As an alternative to a histogram, statisticians have developed several smoothing algorithms to estimate the underlying shape of the data. (Nadaraya, 1965; Hardle, 1991). The process can be thought of as constructing numerous histograms of differing interval widths and averaging the heights of the different bars--a sort of average all possible histograms. Figure 21 presents two density smoothers applied to the data depicted in Figure 20. Here the density shapes differ based on the smoothing algorithm used to average across data points.

A histogram with a large interval width may be smoother than a kernel density curve with a small interval width. Given that both kinds of graph use the same interval width, histograms and density curves are positioned on the continuum as shown in Table 12. Following current practice in the statistical literature we will use the term "bandwidth" rather than the more specific "interval width" since this term is more appropriate for discussions of continuous data and functions.

Two-dimensional Graphs

Bivariate data are usually presented in a scatterplot, which is also subject to the bandwidth problem. If there are thousands of data points, the scatterplot will appear to be a messy cluster of ink. This problem is called overplotting. There are five common ways to attack overplotting. These five types of graph, which are usually used with large data sets, are suitable to examining relationships among variables and discriminating clusters. However, because not every single datum point is depicted, these techniques are inadequate for identifying outliers.

Binning. One way to simplify noise is the binning approach suggested by Carr (1991). In this approach data points are grouped in bi-variate intervals, then plotted in a scatterplot with larger symbols indicating more data points in an interval. The size of the symbol is governed by the bandwidth (see Figure 22).

Sunflower plot. Sunflower plot, which is a function available in S-Plus, is also a common solution to overplotting (Cleveland & McGill, 1984). In a sunflower plot, the density of the bivariate coordinates is represented by glyphs. The more observations the spot contains, the more rays it emanates from the central point, and the more complete the "sunflower" is (see Figure 23). However, Schilling and Watkins (1994) pointed out that when there is only one observation in a spot, the representation is just a point instead of a sunflower. It is difficult for the viewer to reconcile both dots and sunflowers into a unified visual image. Moreover, when there are two observations, a single line extends from a point. It may give the incorrect perception that only one observation occurs there, especially if the central point is not large.

Median smoothing. Another way to simplify a overplotted scattergram is smoothing. Again, bandwidth choice inevitably becomes an issue. When encountering a noisy scatterplot, one can search for a pattern by dividing the data into several portions along the x-dimension, computing the median of y in each portion, and then look at the trend by connecting the medians (Tukey, 1977).

Mean rendering. Mihalisin, Timlin and Schwegler (1991) extended the preceding idea by using the mean rather than the median and introducing bandwidth as a variable. In Figure 24 the relationship between X and Y is depicted in this fashion called mean rendering. The data pattern is clear in the upper right graph where the bandwidth is wide. The bandwidth of the lower graph is three times smaller and thus gives a noisier appearance. A mean rendering imposes more structure on the data than a median smoothing, because interpretation of the mean generally depends on the normality of distributions.

Regression. The fifth way to reduce noise in a scatterplot is fitting a regression line. In terms of structure imposition, regression that assumes a linear function is more forceful than median smoothing and mean rendering, which allows local fluctuations departing from linearity. The positions of these graphics on the noise-smooth continuum are shown in Table 12.

Multi-dimensional graphs

Multivariate research is important in all areas of scientific inquiry. Take the most basic measurement of an element as an example. To describe the condition of an element in a phase space requires at least the coordinate of its 3D physical position, its temperature, pressure and density at a given time. In this simple case there are already seven variables together simultaneously. If the subject matter to be studied is more complicated, it will involve more dimensions. Also, in social sciences many variables interact with each other simultaneously. Multivariate visualization comes to the fore when researchers have difficulties in comprehending many dimensions at one time.

There are numerous multivariate visualization techniques. In this section we discuss techniques mainly related to examining relationships and interactions. Methods for other purposes such as spotting outliers, checking assumptions and discriminating clusters will be covered in later sections. Readers are recommended to consult Keller and Keller (1993) for additional techniques. The visualization techniques introduced here are star graphs, radar plots, stereo-ray glyphs, needle plots, volume models, surface plots, contour plots, image plots, cell means plots, Aiken and West's line plots, Johnson-Neyman plots, coplots, scatterplot-matrix brushing, and animated mesh surfaces. Their relative position in the noise-smooth continuum are presented in Table 12.

Multi-dimensional Graphs with Low Structure Imposition

Star plots, radar plots, stereo-ray glyphs, needle plots, and volume models are considered the nosiest techniques because they show all observations. Star plots and radar plots are only suitable for small data sets, while stereo-ray glyphs can be used in both small and medium data sets. For depicting all data of a large data set, a volume model is the best candidate.

Star plot. A star plot is a simple means of multivariate visualization, in which the multiple measurements of a case are depicted on equally spaced radii extending from the center of a circle and linked to form a star (Feinberg, 1979). For instance, in Figure 25 a star graph implemented in S-Plus shows 31 variables measured in nine states. This method provides an overall impression of change of variable value across subjects. This method is likely to work well with less complex data set. When there are too many variables and observations, a star plot will no longer be appropriate. This visual approach shows all data and thereby it is considered a noisy technique.

A star plot is not good for examining multivariate relationships in a still mode, because it is difficult for one to picture so many changes across subjects, especially when there are many observations. However, if individual stars can put together as a movie, the animated stars perhaps can provide us a clear picture of how the values of multiple variables vary across subjects or over time.

Radar plot. The idea of a radar plot is similar to that of the star plot. In a radar plot, the value of the measurement is also represented by radii stretching out from the center of a circle. However, here each radius stands for a subject instead of a variable. The subjects’ response on each variable is displayed by points of different shapes, colors, or both. In the hypothetical example shown in Figure 26, which is generated using Excel (Microsoft Corporation, 1994), achievement scores in three classes (Physics, Chemistry, and Biology) of eleven subjects are presented. The graph shows the frequencies of data series relative to one another. Apparently, the biology scores are the lowest and are not correlated to either physics or chemistry scores. However, physics and chemistry scores are fairly good predictors of each other. This approach suffers the same shortcoming as the star graph. When there are too many variables and subjects, the data pattern will be concealed. Moreover, the comparison among groups by a radar plot may not be intuitive when a color display is not available. A similar approach named spider plot (SU5 Group, 1991) brings enhancement to the radar plot. A spider plot is usually used for depicting cell means, and therefore it does not fit in the noisy end of the continuum. The detail of a spider plot will be introduced in the section concerning research goals.

Stereo-ray glyphs. A 3D plot with X, Y, and Z variables on three axes--called a spin-plot--is a common way to illustrate multivariate data. The user can rotate the plot to get a sense of depth. In practice, it is difficult to find out how the data spread in a 3D plot, because our binocular depth perception is of limited accuracy. There are several alternative ways to represent the depth (the third dimension) such as stereo-ray glyphs and the needle plot. They both share a common feature--compressing the 3D space into a 2D plane by varying symbols or colors.

The idea of glyph was proposed by Anderson (1960). Carr and Nicholson (1988) expanded this idea and added one more dimension to a 2D plot by using the analogy of a meter, or a ray. In a meter the value increases as the needle moves from the left to the right. In Figure 27 the change of the third variable is illustrated by attaching a glyph, which resembles a meter, to each data point. The angle of the "tail" of each data point indicates the size of change in the moderating variable. In order to view the 3D plot with an illusion of depth, Carr and Nicholson placed the same two graphs side by side and recommended the user to look at the graphs with a stereopticon.

Some individuals reported that viewing stereo-ray glyphs can get a more precise sense of depth than looking at 3D rotation alone (Scott, 1992). However, stereo-ray glyphs have at least three draw-backs. First, it is inconvenient for researchers to examine the graphs with a stereopticon. Second, some people who suffer from depth perception limitations may find this method useless. Last, although stereo-ray glyphs can accommodate more data points than a star plot and radar plot, stereo-ray glyphs, as a nosier technique, still suffer the problem of overplotting when the data set is too huge. In the situation of overplotting, the data pattern is buried by the noisy graph. This overplotting problem may be overcome by using a volume model, which is discussed below.

This addition of a meter is an example of the general strategy of changing the appearance of the symbol that represents an observation. Other packages such as DataDesk, XLISP-STAT, and S-Plus also allow the user to change the shape or color of observations based on values of a third variable.

<u>Needle plot.</u> The needle plot adopts a similar strategy to the Stereo-Ray Glyphs by compressing three dimensions into a two-dimensional plane. In Figure 28 two scatterplots are overlaid--a X*Y scattergram denoted by red dots and a X*Z plot depicted by green dots. The differences between X,Y and X,Z are attenuated by connecting them with needles. Imagine that one plot is in front and one is at back. In this

view the needles provide us a pseudo-sense of depth. Again, a needle plot is not effective in the situation of overplotting. Nonetheless, like stereo-ray glyphs, a needle plot can handle more observations than either a star plot or a radar plot.

It is important to note that in the overlaid plot, the Y and Z dimensions share the same scale and use the same tick marks. A needle plot can work better if the scales of Y and Z are the same. Otherwise, Y and Z must be converted into standardized scores.

Volume model. The volume model overcomes several problems occurring in other techniques such as overplotting that was mentioned previously, as well as perspective limitation, in which data mis-interpretation is possible when the viewer looks at a 3D graph from one particular angle. A volume model can be viewed as an enhancement of a 3D plot (see Figure 29). In a conventional 3D plot, the data points are symbolized by opaque dots and thus overplotting occurs in a large data set. In the volumetric visualization, each data value is denoted a translucent dot varying in intensity. The higher the data value is and the more data that lie along the region, the more opaque the line of sight is. In this way, the researcher can construct a transparent "data cloud" to counter overplotting.

In addition, Kaufman, Hohne, Kruger, Rosenblum and Schroder (1994) argued that a translucent volume model is perspective independent. However, clouds of data may also obscure views if the intensity of color is very high. In a volume model, the user can slice a vertical or a horizontal cross-section to look at the relationships at certain points as shown in Figure 29. Because of the versatility, Nielsen et al. (1994) asserted that the future emphasis in visualization will be on volumetric visualization. However, the strength of showing all data is also a weakness, since it takes some training to be able to present and interpret such displays properly (Fortner, 1992).

At the early stage of exploratory data analysis, a volumetric visualization is likely to be beneficial. A volume model shows all data, and thus the researcher can detect whether there exist relationships among variables and locate clusters of data.

Surface plot. When the completeness of a volume model is not necessary, a surface, a contour, or an image plot may be more desirable. Because all these graphs involve smoothing by the choice of bandwidth, they are considered smoother techniques. A surface plot is more vulnerable to perspective limitations while contour and image plots are more robust against this problem.

A surface plot is easily confused with a smoothed mesh surface plot. In the former, the surface of the raw data are depicted while in the latter a smoothed summary surface is presented. In a surface plot as shown in Figure 30, the data values of X and Z are plotted along the two horizontal axes while the data values of Y determine the height of the vertical axis. The appearance of a surface plot is tied to the grid size like the shape of a histogram is affected by the bandwidth. A small bandwidth will lead to a surface plot that appears with many spikes while a larger bandwidth leads to an appearance of smoother mountains. Because a viewer's perception of the surface plots depends on the viewpoint, they are sometimes called perspective plots. It is desirable for the researcher to vary the grid size and the perspective of the surface plot while exploring data.

Contour plot. In order to overcome the viewpoint limitation of surface plots, a contour plot takes a bird's eye view. In a contour plot, the Y-axis is hidden and the data values in Y are represented by connected lines at discrete levels. Although a contour plot is less viewpoint-dependent than a surface plot, it is still not as perspective-free as a volume model.

Another shortcoming of a contour plot is that it does not easily show "holes" in the data. For example, when a data set has a concentrated region of low-value data at the center, this depression in magnitude is represented by concentric rings--the same symbols that are used to show a concentration of high data values. Moreover, Grotch (1983) compared the efficacy of several multivariate graphing techniques and found that contour and surface plots failed to provide a good quantitative sense for the data. Last but not least, the bandheight of the isolines determines how a contour plot appears. Therefore, researchers should consider construct contour plots with different bandheights in order not to be mentally "stuck" in one depiction.

Image plot. An image plot is a bird's eye view of a surface plot, too. In an image plot the data values are often represented by different color hues. The advantage of this approach is that the maximum and minimum values are easily highlighted. However, choices of color hue, brightness, and saturation need to be made carefully. Bertin (1983) found that if the conventional color spectrum is used, red and blue, which are located at the two ends, are perceived as similar rather than different. And yellow, the lightest color at the center of the spectrum, looks more outstanding. Figure 31a illustrates this problem. In a similar vein, Encarnacao et al. (1994) argued that a color scale based upon perceived brightness is usually more effective. However, in some software packages, it is difficult, if not impossible, for the user to change the default setup of the color scale. Accordingly, an image plot with a monochrome scale varying in brightness, as in a gray scale shown in Figure 31b, is easier for the viewer to properly interpret the data (cf. Lewandowsky, Herrmann, Behrens, Li, Pickle, & Jobe, 1993).

Multi-dimensional Graphs with High Structure Imposition

The above techniques are well-suited for exploratory data analysis. At the later stage of analysis, smoother techniques are needed and structure imposition is inevitable. Cell mean plot, Aiken and West's Plot, and Johnson-Neyman plot with box plots hide raw data points and heavily impose structures on data by introducing centrality, which assumes normality and homogeneity of variances. The last two are more forceful than cell mean plot in terms of structure imposition because besides the assumptions of centrality, they also bring in regression assumptions.

Cell mean plot. If the multivariate data set contains grouping variables and the aim of the researcher is to look for possible interaction effects, a cell mean plot would be a convenient way for visualizing a two- or three-way interaction. When detecting a two-way interaction one simply looks at whether the joint lines of cell means are consistent, ordinal or disordinal. However, in the case of three-way interaction this is no longer intuitive. In three-way interactions, we compare the difference of a difference of a difference. For example, when there are A, B, and C factors and all factors have two levels, the analyst is interested in whether the A*B effect in C_1 and C_2 are consistent. If they are consistent, s/he could conclude that the three-way interaction is not present. For this type of visualization purpose, the researcher can adopt either the one-plot or the two-plot approach.

In the two-plot approach as shown in Figure 32a and 32b, A*B cell means are plotted on the condition of C_1 and C_2 separately. In this case it shows a two-way interaction of A*B in both conditions. However, the fact that the shape of two interactions are similar implies the absence of three-way interaction.

The one-plot approach is quicker than the two-plot method for investigating the three-way interaction. In Figure 32c the C_1 and C_2

means across all the A*B cells are drawn. The cell mean plot blatantly reveals a consistent effect across all A*B cells and thus a three-way interaction effect is absent.

Aiken and West's plot. It is well known that a multiple regression model can be conceptualized in terms of ANOVA and vice versa. Besides ANOVA, the cell mean plot is also applicable to the multiple regression model. In the ANOVA framework, interactions are often viewed as mean differences on one factor varying across levels of another factor. The regression counterpart concerns how slopes relating two variables change over the levels of a third variable. To detect the two-way interaction, Aiken and West (1991) suggested to plot the regression lines of X against Y on the conditions that Z is at the mean, one standard deviation above the mean, and one standard deviation below the mean (see Figure 33). This procedure can be extended to third-order interactions, in which the above three regression lines are drawn separately for data at the mean value of the fourth variable, and the value of one σ above and below the mean.

This smooth approach facilitates computation by using discrete points of the fourth variable. Also, it provides straightforward interpretation of the interaction term. Nonetheless, its strengths are also its disadvantages. First, the raw data are hidden and thereby the visualizer has no information about the residuals. Second, this discrete approach fails to depict the continuous nature of the function. Last, the depiction of three-way interactions is sometimes difficult to interpret since too many lines are used to describe four variables (Behrens & Yu, 1994).

Johnson-Neyman plot with Box plots. When a multivariate procedure includes a grouping variable, a covariate, and a continuous outcome variable, as in an Analysis of Covariance (ANCOVA), these elements and their inter-relationships can be expressed by a Johnson-Neyman plot (1936). Schafer (1992) asserted that identifying heterogeneous regression slopes in ANCOVA is analogous to identifying the interaction effect in factorial design. They recommended to combine a Johnson-Neyman plot with box plots for this particular goal. A Johnson-Neyman graph as shown in Figure 34 is similar to a regular regression line plot, with X as the covariate and Y as the outcome variable. But instead of fitting one regression line, the X scores of two groups separately regress against Y. An interaction effect is easily seen because the two regression lines are disordinal. Schafer argued that the difficulty of this presentation is that a bivariate display of the data does not unfold the univariate distributional characteristic of the two groups. Therefore, they suggested to put the box plots of the two groups for conveying the univariate information.

It is obvious that this technique is limited to situations where a grouping variable and a covariate are involved. When all variables are continuous in nature, a coplot, scatterplot brushing, and animated mesh surface, which will be introduced in the subsequent sections, are considered more helpful.

In addition to varying by the measurement scale of variables, coplots, scatterplot-matrix brushing, and animated mesh surfaces are different from the three previously discussed techniques in the control of noise level. The previous three techniques do not involve the choice of bandwidth, but the subsequent three methods do. The last three methods are positioned at the far end of the noise-smooth continuum because the researcher has the freedom to further smooth these graphs by altering the bandwidth. Also, it is important to note that both coplots and scatterplot-matrix brushing are suitable to three-dimensional data set while animated mesh surface is useful in a four-dimensional space.

Coplot. A coplot is an abbreviation for conditioning plot. This technique is helpful in detecting the presence of interaction effect of multiple variables. When viewing an interaction, different slopes are apparent between X and Y at different levels of Z. If Z is broken into a series of intervals, the regression of Y on X in each Z interval can be assessed with an eye open for differences in slope across the series of plots.

A coplot as implemented in the S-plus software (Statistical Sciences, 1993) is presented in Figure 35. The top panel is called the given panel, which shows a series of intervals across a third variable. The panels below are called dependence panels, which show a series of scatterplots of two other variables. In this example, the two variables on the dependence panels are number of points obtained in a mathematics class and a scaled value of perceived ability in mathematics. In the given panel, there is a scale for learning goal orientation and a series of overlapping lines. Each line represents the range of learning goal which is included in a corresponding scatterplot. The first line reflects the range of the learning goal scores included in the first (upper left) scatterplot, with the second line indicating the range of learning goal scores included on the next scatterplot and so on. The example presented in Figure 35 shows how the slope relating points and perceived ability fluctuates toward zero in the middle of the learning goal dimension while exhibiting positive slope elsewhere. Such a pattern would not be self-evident in examination of simple marginal distributions or unconditionalized scatterplots.

The reader may note that the intervals overlap--an aspect necessary to maintain the continuous influence of points on the conditional regression slope. Because the degree of overlap reflects the degree of local conditionalization for each scatterplot, this aspect of the plot is modifiable in the S-Plus implementation. The length of the conditioning intervals also varies because they reflect the density of points in different regions of the multi-dimensional space.

A coplot is a smoother technique than those discussed above, because the regression lines impose certain structures on the data. On the other hand, the number of the levels of dependence panels can be viewed as a kind of bandwidth.

Scatterplot-matrix brushing. A scatterplot matrix is a collection of scatterplots organized in a matrix analogous to a correlation matrix. Scatterplot matrices designed for exploratory data analysis are usually linked as found in the DataDesk computer package (Velleman, 1992). In this example we show only two scatterplots. Linked plots are indexed so that alteration of an observation on one plot leads to the same alteration of the case on all other linked plots. By brushing (coloring) cases along one scale (Figure 36a), the values along this scale can be perceived in other bivariate plots, which do not include this variable on the axes (Figure 36a). DataDesk has a built-in function for producing colored regression lines as shown in Figure 36b. The plot shows the regression lines of perceived ability against points conditioning to six levels of learning goal. Here it can be seen that the regression line for orange cases (those that are at the center of the learning goal scale) is relatively flat while other regression lines are otherwise positive.

Like a coplot, the number of sliced segments affects the conditional regression lines. Clearly the exploratory mode of such analyses leads to sets of possibilities and should not be confused with the results of confirmatory data analysis. Unlike a coplot in which the given panel has overlapping intervals, the colors in a brushing scatterplot cannot overlap since each case must be of one color or another. The choice of the number of color categories can also be viewed as a bandwidth decision. Several coloring schemes should be attempted to insure the slopes are not simply artifacts of an unusual grouping scheme. I recommend using both coplots and linked scatterplots to obtain a comprehensive view of the data. One drawback of brushing scatterplots is that the shift of conditional regression lines cannot be animated and the viewer has to mentally shift those lines to construct a pattern. This problem can be solved by the Animated mesh Surface model.

Animated mesh surface. A mesh surface plot is a simplification of a surface/perspective plot. Figure 37 illustrates how a mesh surface is formed by joining the regression lines of the predictor variable (perceived ability) against the criterion variable (self-regulation) across all levels of the second regressor (learning goal). In this example, three conditional regression lines are drawn. The first one is plotted given that the learning goal value is one standard deviation above the mean. The second one is plotted on the condition that the learning goal value is at the mean, and the last condition is one standard deviation below the mean of the learning goal value. In this example the procedure is implemented in Mathematica (Wolfram, 1991). The remaining plots illustrate the lines extended across the continuum to produce a surface and the surface being rotated to improve perspective. One merit of this approach is that in the first step it shows the regression lines in the three-dimensional context of the data. The final plot shows the surface with the data omitted.

By animation, this technique can be easily extended to the visualization of four-dimensional data. In Figure 38 there are three regressors--perceived ability, extrinsic motivation, and performance goal, and one outcome variable--deep thought processing. In the first box we connect the conditional regression lines of performance goal against deep thought processing across all levels of extrinsic motivation when the perceived ability is low. The same procedure repeats as the value of perceived ability increases. As a result, we produce a series of mesh surfaces as a movie. The user can either play the entire movie to get an overall impression or look at the graphs frame by frame. Interesting results may be discovered by this procedure. For instance, in the fourth box of Figure 38, it indicates that at a certain value of perceived ability, the mesh surface is flat and all main effects are absent.

However, the user should be cautious not to over-interpolate the function to areas where no data are found empirically. For example, in Figure 39 the lowest scores of the perceived ability did not reach the minimum of the scale, but the surface stretches all the way to the corner of the box as shown in Panel a. The graph of Panel b is a more accurate representation of the data.

The understanding of animated mesh surfaces is highly perspective-dependent. Although this shortcoming is possibly remedied by rotating the graph, a rotation will lead to the re-drawing of the entire movie. In order to compensate the perspective limitation, I recommend to present an animated mesh surface, contour plots and image plots side by side. It is important to note that here the contour plots and the image plots are no longer the bird's eye view of the rough surface plots. Rather smoothed mesh surfaces should be used.

Wickens, Merwin and Lin (1994) asserted that a mesh surface does not benefit the user in understanding the data. However, in their study only a few data points were presented and the task focused on individual values rather than the overall pattern. In addition, the surface used in that study was static in shape and animated only to change viewing position rather than changing the function itself as shown here. In short, the graphical format, task type, and data type were not properly aligned in their study.

Summary

In the first component of the alignment framework, visualization techniques are categorized with reference to the adjustment of noise and smooth. However, there is no optimal bandwidth and structure that can be applied to most situations. Researchers are encouraged to look at the data in different ways. The general rule is that in the exploratory stage of data analysis noisier techniques seem to be more appropriate. The relationship between research goal and graph type will be discussed in the next section.

Besides the bandwidth decision, the choice of visualization techniques also depends on perspective limitation. The sample size is a crucial factor in determining whether a perspective-free or perspective-dependent graph should be used. Further, as mentioned before, some graphs such as star and radar plots are suitable to small data sets while some graphs such as stereo-ray glyphs and volume model are useful in larger data sets. The appropriateness of graph is inevitably affected by the number of observations and the number of variables. The relationship between data type and graph will be discussed in a later section.

Research Goals and Visualization

In the previous section, I introduced noise and smooth as a means of understanding visualization techniques. The choice of using noise or smooth technique is closely tied to one’s specific research objective. Keller and Keller strongly endorsed (1993) the practice of beginning data visualization by first identifying the goal, because knowing the goal the researcher may recognize new sources of techniques. Even considering existing techniques, it is beneficial for a researcher to make the graph and data types fit each other. Because the task type is driven by the research goal, in this paper I will use the two terms interchangeably.

Several existing frameworks attempt to enhance the effective use of graphs by matching task and graph types. Chernoff (1978) outlined a list of purposes (e.g. illustration, analysis, or computation) to correspond to the graph type classified by accuracy (e.g. precision or distortion) and dimensionality. However, the taxonomy of task developed by Chernoff is not specific enough to be a practical guide. Tan and Benbasat (1990) decomposed graphical elements in attempt to match data extraction tasks and graphical representations. Nonetheless, their taxonomy includes only two-dimensional, but not multi-dimensional graphical formats.

To compensate for the limitations of existing taxonomies, a more practical framework is provided. In my taxonomy, there are six major categories of research goals as the following:

1. spotting outliers;
2. discriminating clusters.
3. checking assumption;
4. examining relationships and interactions;
5. comparing group differences;
6. controlling quality.

All these goals can be accommodated in the noise-smooth continuum. For the first three goals, nosier techniques are desirable. For the last three objectives, smoother methods are recommended. Each of them is addressed now.

Spotting Outliers

It is often helpful for a researcher to begin visualization with the goal of detecting outliers. Because if subsequent procedures are parametric statistics, the tests will be sensitive to extreme cases. Moreover, this procedure should be implemented before discriminating clusters, checking assumptions, and doing remedial data transformation due to the following reasons. First, some clustering procedures such as Ward's method perform poorly with the presence of outliers (Everitt, 1993). Second, some transformation methods are ineffective for normalizing data that is highly skewed due to outliers (Rasmussen, 1989). Also, the need for data transformation can be unduly influenced by one or two outliers (Atkinson, 1985). After removing outliers the assumptions may no longer be broken. At this exploratory stage it is preferred to include all observations and not to impose any structure on the data.

Guttman (1973) regarded an outlier as an observation that does not belong to the target population. Chatterjee and Hadi (1988) defined an outlier as an observation with a large residual. There are many other definitions, but all of them share a common characteristic: outliers lead to a lack of fit to the model that would potentially bias the result of the analysis. No wonder outliers are sometimes named "contaminants" (Davies & Gather, 1993).

There are numerous tests to spot outliers, such as Leverage points, Studentized residuals, DFITS, Cook's Distances, DFBETAs, Mahalanobis Distances, and the Andrews-Pregibon statistic (McGinnis, 1991; Jarrell, 1992). However, many test statistics for detecting multivariate outliers are susceptible to either the problems of masking, swamping, or both (Jarrell, 1991; Hadi & Simonoff, 1993). And no single method is adequate for identifying all outliers (Hecht, 1992). Even in a univariate case, a heavy-tailed distribution may introduce a misleading result. Thus, visual approaches should be used as a supplement to test statistics for avoiding these problems.

Heavy-tailed distribution. On some occasions the so-called "outliers" are just data points from a heavy-tailed distribution. In other words, the "deviated" observations are due to the variability inherent in the data (Grubbs, 1969; Hawkins, 1980). Therefore, it is important to visualize the shape of the population and predict the tendency of the two tails. For example, although Figure 40 shows several remote observations (shaded) appearing to be "outliers," the distribution is indeed flatter than normal. If a curve is overlaid on top of the histogram, those so-called "outliers" are likely to be included.

Swamping. Besides a long tailed distribution, swamping may also cause the mis-classification of good observations as outliers. Swamping occurs when some good observations are close to another remote subset of observations (Hadi & Simonoff, 1993). In this case test statistics may mislead us. In Figure 41 the shaded observations near the authentic extreme scores may be mis-identified as outliers. By utilizing such a simple graphing technique as a histogram, the analyst may be able to separate the good observations from the cluster of true outliers when he sees the location of those "outliers" relative to the entire distribution.

The previous two examples are demonstrated with univariate cases. In multivariate cases the situation would not be so straight-forward and a 3D spin plot is always recommended.

Masking. Masking is the opposite of swamping. Masking refers to the problem that some outliers are undetected when two or more outliers are bundled together on the same side of the sample (Hadi & Simonoff, 1993). The larger the sample is, the more masking may occur because of the larger number of outliers (Beckman & Cook, 1983). A spin plot, which enables the visualizer to examine the data from multiple perspectives, is helpful to unmask bundled outliers. For example, in Figure 42a there is a possible outlier at the bottom right of the graph. After a rotation, three outliers at the lower left are exposed in Figure 42b.

Overplotting. A regular spin plot may not be effective in spotting outliers when there are too many data points, because an overplotted graph fails to give you a clear sense of distance among observations even if you rotate it. Both DataDesk and SAS/JMP offer a rescaling feature to attack the problem of overplotting. By pressing a button the data points will spread out according to a new scale. For example, The top panel of Figure 43 shows a cluster of dark ink in which the presence of outliers is unknown. In the bottom panel of Figure 43 two possible outliers emerge after the data points have zoomed out.

Ellipse boundary. If two variables are bivariate normal, the data points should form an ellipsoid pattern. SAS/JMP provides a tool for spotting outliers by showing bivariate normal ellipses in a scattergram with different probabilities. In the example shown in Figure 44, the outside ellipse (the red line) includes 99 percent of the data and no observation is considered an outlier. The middle ellipse (the green line) tightens the coverage to 95 percent and three outliers are detected. The inside ellipse (the blue line) yields the same result as the green one.

A similar solution to the previous approach is to construct an ellipsoid of data concentration and the boundary of rejection according to the standardized, Studentized, leverage, and Cook’s Distance method. This procedure can be implemented in a computer program DRAWREG written by Hecht (1992).

Leaving-one-out. As mentioned before, outliers can be defined as contaminants that would seriously affect the analysis result. Leaving-one-out (LOO) is a powerful technique to see what the result would turn out if the suspected outlier is deleted (D'esposito, 1992). This is a special case of the Jackknife procedure described by Mosteller and Tukey (1977). In both R-code (Cook & Weisberg, 1994) and SAS/JMP (SAS Institute, 1989) this function is performed in an interactive manner i.e. the user can exclude any data point from the graph and the regression line will be re-fitted according to the remaining observations. The example in Figure 45 is implemented in SAS/JMP. The red regression line is fitted with the suspected outlier at the bottom whereas the green line is fitted without that observation.

Discriminating Clusters

The goal of discriminating clusters may be implemented at the early stage of data analysis. It is because the latent groups that are un-noticed by the researcher may bias the result if s/he treats all subjects as one group. Moreover, grouping of subjects can provide an informal means for assessing dimensionality, identifying outliers, and suggesting hypotheses regarding to relationships (Johnson & Wichern, 1992).

There is no single best clustering algorithm. Sometimes noise level and the presence of outliers affect the effectiveness of various clustering methods. Milligan (1980) found that when the data were noisy, single linkage, centroid and the median method might not be capable of recovering the cluster structure while Ward's method and group average had better performance. On the other hand, in the presence of outliers Ward's method and group average were ineffective but single link, centroid, and median methods provided the best results.

Jain and Dubes (1988) noted that “cluster analysis is a tool for exploring data and must be supplemented by techniques for visualizing data.” (p.7) In a similar vein, Hair, Anderson, Tatham and Black (1991) contended that cluster analysis is much more of an art than a science. In other words, the process involves creativity and subjective judgment. Hair et al. (1991) also warned that cluster analysis can be dramatically affected by the inclusion of only one or two inappropriate or undifferentiated variables. The researcher is encouraged to examine the results and eliminate variables that are not distinctive across the derived clusters. No doubt visualization such as Chernoff faces, parallel coordinates plot, color enhanced generalized draftman’s plot and many others would be useful for this type of exploration.

Scatterplot-matrix brushing and 3D spin plot. One simple visual approach for clustering is applying both scatterplot-matrix brushing and 3D spin plot together. Figure 46 is an example implemented in DataDesk. In this example, the researcher detects two distinct groups in X*Y plot and attribute the two groups with two different colors--red and green. The distinctness of the two groups holds in Y*Z plot but not in X*Z plot. The researcher can change the group assignment of individual subjects until the clusters are well discriminated in all dimensions. A 3D spin plot can help to verify whether such discrimination is successful or not.

Chernoff faces. Chernoff faces, as the name implies, was invented by Chernoff (1973) for representing multivariate data. The procedure is simple but effective for discriminating clusters. In this procedure, each facial feature denotes a particular variable. For example, X_1 can be associated with the size of the mouth, X_2 with the size of the nose, X_3 with the size of the eyes, and so on (see Figure 47).

The power of a Chernoff face is its interpretability and its interesting way of presentation. Differences of facial features are easily spotted for clustering groups and integrating summary of multiple dimensions. Repetitious viewing of large tables of data is tedious, but Chernoff faces can significantly improve data digestion in terms of both accuracy and speed (Moriarity, 1979; Scott, 1992). Also, Brown (1985) found that Chernoff faces provided better cluster discovery than some other techniques such as Andrew plots (c.f. Andrews, 1972) and box plots, irrespective of data dimensions and distance.

A major drawback of Chernoff faces is that the subjective assignment of facial expressions to variables affects the shape of the face. Chernoff and Rizvi (1975) found that the permutations of the assignment of features caused an error rate of as high as 25 for the task of classifying faces into groups. It means that classifying two faces as “fairly similar” is greatly influenced by the assignment of variables to specific features. Further, Brown (1983) found that the mouth feature had a more dominant influence on visual perception. Last, some researchers such as Flury and Riwdwyl (1981), and Turner and Tidmore (1980) criticized that the symmetrical feature of Chernoff faces is redundant.

Parallel coordinates plot. A parallel coordinates plot (Inselberg, 1985) is a technique to cluster subjects by response in multiple variables. In a parallel coordinates plot as shown in Figure 48, variables are displayed on horizontal axes in a parallel fashion. Same subjects’ responses on each scale are connected by straight lines so that the response pattern can be revealed. In this hypothetical example, the researcher can classify two groups according to their achievement scores in four tests. The red-line group did poorly in mathematics, a little better in sciences, poorly in verbal skills, and a little better in social studies. The green-line group has weak performance in mathematics, better in sciences, extremely good in verbal skills, but very poor in social studies. A black and white parallel coordinates plot may be affected by overplotting. Coping with this problem, some software packages such as BMDP/Diamond present parallel coordinates plots in color so that the response pattern can be more outstanding. However, the parallel coordinate plot takes only the subject space into account, but not the variable space. For techniques that use the distance within subject space and between variable space, readers are recommended to read Dawkins (1995).

Color enhanced generalized draftman's plot. A color enhanced generalized draftsman's plot (Ware & Beatty, 1988) can be used for spotting outliers and discriminating clusters at the same time. A generalized draftman's plot is another name for a scatterplot matrix. In a regular scatterplot matrix, values of multiple variables are plotted in different panels in a pairwise manner. Besides the physical space of two axes, Ware and Beatty added the color space for higher dimensions. In their example, red, blue, and green are used to represent other dimensions.

For the ease of discussion, the example given in Figure 49 uses only two colors. In each panel of the scatterplot matrix, the values of the first three variables are shown by X-Y-Z coordinates, the fourth dimension can be represented by the luminance of red while the fifth, blue. In this way, the group clusters can be identified by both the location in the XYZ space and the mixture of colors. For instance, if the subjects are low in the fourth dimension and high in the fifth, the data points will be shown in pink. If the observations are high in the fourth but low in the fifth, their color will be purple. If the group has a middle value in both dimensions, magenta will turn out. When both dimensions have low values, the hue is gray (see Figure 50).

Ware and Beatty (1988) found that there was a clear advantage in using physical space and color together for identifying multiple clusters. However, this advantage was not shown in the case of only one cluster. Moreover, users must be cautious about the following limitations of this graphing approach. First, this method is only helpful in detecting the existence of clusters, but not exactly where they are located. Second, the purpose of this graphical format is mainly for detecting groups rather than looking up data values. Using color for depicting values is prone to perceptual error as hue variations do not imply magnitude (Hubbold, 1992). For instance, Cleveland and McGill (1983) found that users tend to perceive an area represented by red is larger than that shown in green. Besides the illusion resulted from hue, saturation and intensity also may bias our judgment. The RGB system used by electronic displays and the YMCB system used in print media produce different visual effects. If data values are converted into the YMCB system, low values will yield a light color. As you see in the last panel of Figure 50, by mixing RGB low values in two dimensions generate a gray appearance. It is somehow counter-intuitive because we always associate low value with light color and high with dark. Further, when the value is too low, structures presented in the variables mapped to hue and saturation are not observable. For example, a viewer cannot detect the difference of saturation in black.

<u>Other clustering procedures.</u> There are many other cluster analytic procedures such as clustered histogram, profile diagram, cluster diagram, dendrogram of hierarchical clustering (tree plot), and castle plot. Freni-Titulaer and Louv (1984) found that among many

clustering techniques trees performed best in many situations while castle plots were considered the least helpful. For the detail of castles and trees, consult Kleiner and Hartigan (1981). For a comprehensive review of clustering methods, consult Jain and Dubes (1988) and Everitt (1993).

Checking Assumptions

With the goal of ensuring the legitimacy of parametric procedures, the researcher should apply test statistics, visualization, or both, to find out whether the data structure meets several assumptions. For example, a Quantile-Quantile plot (Q-Q plot) is usually used for checking homogeneity of variances. At this stage certain structures may be imposed on the data such as transformation to normality and linearity, stabilization of the variance, or all of the above.

Checking normality. The assumption of normality can be tested by the Shapiro-Wilks test, Kolmogorov-Smirnov test, histograms, and normal probability plots. These procedures are usually effective in testing univariate normality, but not multivariate normality. If a variable is multivariate normal, it is also univariate normal. But the reverse is not necessarily true. Multivariate normality is difficult to test. The validity of measures of multivariate skewness and kurtosis are subject to certain limitations (Horswell, 1992).

Surface plot. There are a few graphical ways of testing multivariate normality such as the Goodness of Fit test (Ozturk & Romeu, 1992) and χ^2 probability plot (MULTINOR) (Thompson, 1990; du Toit, Steyn & Stumpf, 1986). A less effective but much easier visualization technique for this purpose is to test bivariate normality by examining all cases of bivariate data. This procedure can be implemented in Mathematica (Wolfram, 1991) as shown in Figure 51. If the joint distribution of two variables appears as a Texan hat, and most points cluster around the centroid (bivariate means), with fewer and fewer points as one moves away from the joint means, these variables will be considered bivariate normal. It is important to note that this example is a hypothetical data set. In reality the researcher seldom gets such a clean surface plot. While dealing with a messy data set, one should alter the bandwidth or apply Kernel estimation to unmask the underlying data structure.

Contour plot. A surface plot as shown above is perspective-dependent. The visualizer may consider to use a contour plot to overcome this limitation. However, the contour plot lacks the exact detail and visual impact available in a surface plot. Scott (1992) recommended a way of expressing the difference by hanging a corresponding contour plot over a perspective plot.

With a contour plot, the analyst can check not only the bivariate normality, but also the bivariate relationship, and the existence of distinct groups. This will be discussed further in the section of examining relationships and interactions.

Spotting multicollinearity. In regression when several predictors are highly correlated, this problem is called multi-collinearity. The assumption of the absence of multi-collinearity is essential to multiple regression model for avoiding high standard error and parameter estimates. Variation inflation factor (VIF), tolerance value, and condition index are common methods for checking this assumption. However, there is no generally agreed cut-off for VIF (SAS Institute, 1990). Neither is tolerance value nor condition index.

Visual aids such as a 3D spin plot can be used as an supplement to the previously mentioned methods. In a 3D plot, without rotation data points may seem to spread out even if multi-collinearity exists (see

Figure 52a). However, during spinning the problem of multi-collinearity may emerge. Figure 52b shows that from a certain perspective the data points appear to form a vertical and lean cluster along the dimension of the outcome variable. This lean cluster suggests a high degree of multi-collinearity (Cook & Weisberg, 1994).

Using this approach one is capable of spotting very high multi-collinearity. Nonetheless, data points do not form a narrow line if the multi-collinearity is moderate. In this case, a mesh surface can provide more visual cues for the detection. Figure 53 is a 3D plot of three hypothetical variables-X, Y, and Z, with a regression mesh surface, implemented in XLISP-STAT (Tierney, 1991). The mesh surface is a function defined by the data points. Geometrically, we can think of the plane as an object that is supported by the points. There exists the threat of collinearity between Y and Z if the dots do not spread out in the X space enough to provide a stable support for the plane. Readers may consult Stine (1995) for other graphical techniques of spotting multi-collinearity.

Examining Relationships and Interactions

Examining relationships and interaction effects plays a major role in bivariate and multivariate statistics. However, Keller and Keller (1993) argued that the choice of techniques should be determined by what the researcher is trying to learn from the image, but not by the relationship among the dimensions. Dimensions always associate with the dependency of variables. Conventionally, the y-axis is regarded as the dependent dimension. Keller and Keller stated that the relationship between dependent and independent variables can be expressed as a function such as $y = f(x)$. In using visualization techniques it often does not matter whether you have two variables, or one dependent and one independent variable. In other words, a functional equation can be rewritten in different ways as the following:

density = mass / volume

mass = density * volume

volume = mass / density

mass/volume = density

density * volume = mass

mass / density = volume

This argument is by no means new. As a logical positivist, Russell (1913) cited functional relationships as evidence to attack the notion of causation. Readers who are interested in the philosophical argument of functional relationships and causation may find Cook and Campbell's work (1979) helpful. In practice, it is advantageous for natural scientists and engineers to visualize dimensions in terms of functions. The same strategy may also be helpful in psychology and social sciences at the exploratory stage of visualization like using scattergrams. However, many advanced dimension-reduction procedures in social sciences such as CHAID and XAID do specify dependent and independent variables (du Toit, Steyn & Stumpf, 1986).

Many visualization techniques are useful for performing relationship examination including contour plots, cell mean plots, Johnson-Neyman plots, coplots, brushing, and animated mesh surfaces. However, these techniques hide raw data, assume structures, or both. Usually they should be performed at the later stage of data analysis.

Contour plots are noteworthy because such a simple tool can help the researcher to look at joint distributions, relationships, and clusters simultaneously. A series of contour plots in Figure 54 show the combination of different distributions, relationships, and grouping (Wand & Jones, 1993). Figure 54a is an uncorrelated normal bivariate distribution while Figure 54b is an correlated normal case. Figure 54c is a moderately correlated skewed distribution and Figure 54d is a slightly correlated but kurtotic distribution. The example of Figure 54e is a bivariate correlation with two distinct groups. In this case a scattergram can detect such blatant groups, too. However, the two groups in Figure 54f may not be easily picked up by a scattergram if too many subjects cause the problem of overplotting. In this case, the contour plot can unveil the two clusters out of the ellipse.

Comparing Group differences

Comparing group differences for the evaluation of treatment effectiveness is a common practice in psychology. Parametric procedures such as t-tests and F-tests are widely used for this purpose. However, those procedures based upon centrality may mislead the researcher, especially in the case of heterogeneity of variance. Some researchers (e.g. Cleveland & McGill, 1984; Cleveland, 1993; Feingold, 1995) advised that other than the centrality the researcher should also look at other aspects of distributions.

Diamond plot. SAS/JMP (SAS Institute, 1989) provides diamond plots to visualize variability, as demonstrated in Figure 55. Figure 55 condenses a lot of essential information--the grand sample mean, group means, raw data, CIs and quantile information. The grand sample mean is represented by a horizontal dot line. The diamond is the CI for each group and the horizontal lines inside the diamonds are the group means. In addition to CI, JMP also provides the option of overlaying a boxplot showing quantile information. Visualization of CI can also be applied to F-test by using a leverage plot (SAS Institute, 1989).

Spider chart. A spider chart (SU5 Group, 1991) can be used for condensing the summary information of a factorial design. A spider graph (see Figure 56) and a radar plot (see Figure 26) are basically the same, except that in the spider plot the joined values form a colored area to make group differences more outstanding. Figure 56 is an example of a spider plot generated by DeltaGraph Professional based upon a fictitious 3*5 factorial design. One factor is the computer language being taught whereas the other is the teaching medium. Each radius is the scale for each cell. The cell means are plotted and connected by different colors. In this example, online tutor is more effective in teaching Common Lisp and Borland C++ while Video is better for HyperCard, Visual Basic and Turbo Pascal. Lecture is the least effective medium for all languages.

Mosaic plot. If the data are categorical, the group differences can be visualized by a mosaic plot (SAS Institute, 1989) as shown in Figure 57. A mosaic plot is a pictorial version of a break down frequency table. The number of observations in each cell is represented by the area. In Figure 57 it shows the usage of computer platforms by gender. Overall there are more males than females in this sample. There are the same percentages of male and female UNIX users, but men tends to favor IBM while Mac is preferred by women.

Controlling Quality

In some situations such as quality control, spotting outlier is not just a preparation for further analysis. Instead, this activity plays a central role of the data analysis. The movement of statistical process control (SPC), widely known as Total Quality Management (TQM), was launched by Edward Deming. The primary target audience of TQM is in industry. Nonetheless, in recent years there are more and more applications of TQM to educational research (Macchia, 1993).

According to Deming (1993), variation is inherent in all processes. Some variations are due to special causes while some are owing to system causes. It is the responsibility of researchers to spot and resolve special causes of variation. For the convenience of subsequent discussion, I consider his theory of variation in conventional terms: there exists random fluctuations. Some sampling fluctuations are within the population but some are not belonging to the population. Here the goal of the researcher is to examine whether there are some outliers and to study what causes those extreme cases.

In Deming's school there are many graphing techniques to achieve the preceding goal. For example, Pareto diagrams, fish bone diagrams, least median square (LMS) regression plots, control charts, relational diagrams, affinity diagrams, tree diagrams, matrix diagrams, matrix data-analysis diagrams, Taguchi's plot, and many others. In addition, because quality control is concerned with the consistency of performance over time, many SPC methods involve time-series study. Schwabe (1993) provided a detailed description of computer support for the above visualization needs. Among these techniques, LMS regression plots and control charts are more exploratory and data-driven.

LMS regression plot. Figure 58 is an example of LMS regression plot. In this example the change of Y values is plotted against the change of time. Two regression lines are superimposed on the raw data points. The lower regression line is based upon traditional least sum of squares (LSS), which is sensitive to outliers, or special causes of variations. In contrast, the upper line uses the least median of squares (LMS) introduced by Roussessuw and Leroy (1987), which is robust against extreme cases. By comparing the robust and non-robust regression lines, the visualizer can determine whether there exist special causes of variations and where the sources of variations are. In this example the plot indicates that two lines diverge at the later years and therefore the problem of variation emerges recently. While compensating for the shortcomings of LSS, LMS may pull us to another extreme--the LMS method tends to detect too many outliers (Fung, 1993). And thus it should be supplemented by some other techniques such as control chart, which will be introduced below.

Control chart. A control chart, which is also known as a Schewart chart, is based upon some distributions, and usually the normal distribution. Essentially, a control chart is a standardized residual plot with a center line (CL) and two control lines. One control line is located at 3 standard deviations above the CL, which is called the upper control limit (UCL). The other one situates 3 standard deviations below the CL, which is labeled as the lower control limit (LCL). Figure 58 is an example of control chart showing the standardized residuals against time. Tiernan (1991) recommended the following criteria for determining when a process is out of control:

1. One point falls outside the 3 σ control limit;
2. seven successive points are on the same side of the CL;
3. seven successive points show an increase or decrease;
4. two out of three successive points are both on the same side of the CL and outside the 2 σ zone;
5. four of five successive points are on the same side of the CL and outside the 1 σ zone;
6. other non-random patterns occur such as trends, cycles, or an unusual spread of points within the control limits.

Regression models, including LSS and LMS methods, require several assumptions such as normality, linearity, homogeneity of variances, and the absence of multicollinearity. Using standard deviations as the control limits in control charts assumes a normal distribution of the data. Other kinds of control chart, as noted above, also rely on some other theoretical distributions and the independence of observations. Jacobs (1990) contended that these assumptions are often unrealistic.

He suggested visualizing the data and transforming them if necessary. Therefore, SPC visualization generally take places at the later stage of data analysis.

Summary

This discussion illustrates that research goals strongly dictate the choice of graphical techniques. However, although in this section specific goals are discussed in the sequence of the noise-smooth continuum, it does not imply that the researcher must follow this rigid order. As a matter of fact, several procedures such as contour plots and coneplots can serve multiple purposes. Further, in many situations, the combination of test statistics and graphing techniques can enable the researcher to have a more thorough understanding of the data.

Data Type and Visualization

Data type is the third component of the alignment framework. Referring to Figure 1, readers can see that there are four aspects of data type: origin, format, complexity, and distribution. In this section I will explain why all these aspects are crucial to the appropriate choice of graphing techniques. Data origin is concerned with the meaning of the data while data format is about the measurement scale. There are two factors contributing to data complexity, namely, the number of observations and the number of variables. Data distribution is classified in terms of normality or non-normality.

Data Origin

Convinced by their experience, Keller and Keller (1993) argued that recognizing data type would lead to the cumulating of mental blocks in data visualization. They observed that many scientists and engineers have been conditioned to regard data as some entity with inviolate properties. Consequently, this rigid thinking narrows the choice of visualization techniques. To eradicate mental blocks, they suggested one treat data only as numbers that a computer knows about devoid of reference to data origin (discipline or application).

On several occasions, the origin of data does determine the use of techniques. For example, if the data are from medical scanning devices, in which surgeons analyze them for surgical aids, precision is the prime criterion for a successful visualization. Under this circumstance all or most data must be included and a translucent volumetric model is probability most useful. On the other hand, data for economic forecasts may need a smoother visualization technique, rather than a volumetric model, to indicate a trend. For applications of visualization to various disciplines, readers may consult Quarendon (1992).

Data Format

Data format is defined as the level of measurement, or measurement scale: Nominal, ordinal, interval, and ratio. Some researchers argued that this classification is outmoded (e.g. Velleman & Wilkinson, 1993). This line of thought can be traced back to the school of "measurement-independent," which stresses the position that "the numbers do not know where they came from." (Lord, 1953; Gaito, 1980). Keller and Keller (1993) hold a similar position that visualizers should not worry about how the numbers are measured and represented. Data of any format van be easily converted by computer. Ignoring data conversion errors, especially in the exploratory phase, encourages rapid evaluation of visualization techniques.

In computer and engineering sciences, it is a common practice to cast variables of integer, floating point, double floating point, unsigned...etc. back and forth. Also, visualization in these fields, which primarily utilize matrices, can afford data conversion errors. However, in behavioral sciences, variable casting may bring detrimental effects. For instance, Pedhazur and Schmelkin (1991) warned that a great threat to the validity of research occurs when a non-experimental design with continuous independent variables is cast into the form of factorial ANOVA. First, the research result may suffer serious information loss from categorizing continuous variables. In addition, it creates the false impression that a non-experimental study has been sanctified into an experimental study. Following this rationale, Behrens and Yu (1994) asserted that Aiken and West's plot, which casts variables of continuous nature into a cell-mean-type plot, suffer tremendous loss of information and is not an optimal representation of regression models. In brief, data format in terms of measurement scale should be taken into consideration for the selection of visualization techniques in psychology and social sciences. In the following I will introduce frameworks introduced by other researchers pertaining to how the measurement scale is related to multivariate procedures and graphical elements.

Holmes (1994) developed a scheme of how data type links with methodology. Table 10 and 11 present her summary of the relationship between methods for studying and representing data, and variable types. Basically, Holmes classified data type into continuous and categorical variables. Although the taxonomy is concerned with multivariate statistical procedures rather than multivariate visualization techniques, the tables are still a good reference for indirectly relating data format to visualization.

Mackinlay (1986) analyzed graphic elements such as position, length, angle, color ... etc. and ranked their appropriateness according to measurement scales. For example, with quantitative (interval/ratio) data, length, angle, and slope in a line graph is better for showing comparisons than a change in shape or color; with ordinal data changing levels of gray scale are superior to other graphical elements for showing ranks; with nominal data color hue is better for showing clustering of categories than others. The detail of Mackinlay's framework is presented in Table 12.

Ware and Beatty (1985) focused on the relationship between measurement scale and a specific graphical element--color. They asserted that color is helpful in distinguishing one nominal category from others because color has the advantage of being analyzed "pre-attentively" by the human visual system. However, color is not a good tool for presenting ordinal data. A gray scale from black to white is perceptually ordered, but a scale based on the physical spectrum is not--people do not intuitively know that green lies between red and blue. For interval and ordinal data color is definitely a poor choice because the color spectrum is incapable of conveying the change of continuous value.

As a guideline for choosing visualization models according to data format, the above frameworks are either too broad or too narrow. Holmes left the reader to find the appropriate graph for the procedure while Mackinlay, Ware and Beatty discussed specific graphical elements instead of graphs. To provide a better guideline, in the following I will illustrate how research goals and measurement scales interact with each other in determining specific graphs.

Discriminating clusters. Cluster analysis is based upon data format of either pattern matrix or proximity matrix. In the former case, the variables can be nominal, discrete or continuous. In the latter case the variables are mostly binary--similarity or dissimilarity (Jain & Dubes, 1988). Consequently, the choice of clustering visualization is subject to the data format. For instance, when the variables are continuous, a clustering of neighboring members is allowed. However, in image processing if we apply this clustering algorithm to line-art data, which is on or off in nature, the image will be blurred instead of being enhanced!

Examining relationships. Suppose a researcher aims at examining relationships and interactions. If the data contain grouping variables as factors, a cell mean plot can be used to display the presence or absence of main effects and interaction effects. When the data include a grouping variable, a covariate and a continuos outcome variable, a Johnson-Neyman plot with box plots is the qualified candidate. When all variables are either interval or ratio, scatterplot brushing, coplot, animated mesh surface would be more appropriate. Provided that the data set has both categorical dependent and independent variables, a graphical exploratory technique named CHAID can be helpful. In the case that the dependent variable is continuos and the predictors are categorical in nature, then another graphical method called XAID is preferably used.

Comparing groups. When one’s goal is group comparison as in a one-way ANOVA where the dependent variable is interval and the independent variable is ordinal or nominal, diamond plots can be helpful. When there are a few grouping factors, leverage plots with confidence intervals are suitable. On the other hand, ordinal or nominal variables necessitates the use of mosaic plots.

Controlling quality. Assume that the researcher is interested in quality control. Under the circumstance that the data are continuous, s/he can apply X-control charts which utilitize means, W-control charts that are based upon range, or control charts using standardized residuals. On the other hand, given that the data are resulted from dichotomous measurements, P-charts which rely on the binomial distribution and the Poisson approximation should be the choice (Cherland, R. M., 1992; du Toit, Steyn & Stumpf, 1986).

By SAS/JMP convention, control charts in which the quality characteristic is in a continuous scale are called control charts with variables. When the number of non-conformities (defects), which are discrete in nature, are used as the indicator of quality, this type of control chart is named the control chart with attributes. Figure 60 shows an example of a control chart with attributes. Unlike the control chart with variables as shown in Figure 60, control chart with attributes allows a localization of mean and control limits by sub-sampling i.e. the data are sliced into several segments, and the mean and control limits of each segment are plotted.

Data Complexity

Although data complexity is a crucial aspect in visualization, there are quite a few misconceptions in this regard. For example, the assertion made by Wickens et al. (1994) that interactive visualization techniques do not help data analysis is made without the consideration of data set size. Moreover, some researchers tried to resolve the problem of data complexity in visualization by applying the findings in cognitive psychology. While discussing the bandwidth dilemma in data representation, Aaronson (personal communication, November 10, 1994) suggested to think about the rule of “Seven plus and minus two,” which was introduced by Miller (1956/1994). Aaronson is not alone. Keller and Keller (1993) also endorsed this rule for scientific visualization:

> A good guide for determining complexity is “The Magic Number 7 plus and minus 2.” Oriental artists typically restrict compositions to five elements. Studies indicate the human brain can best handle seven unrelated elements, more or less, simultaneously. The spread from five to nine can be a function of viewer’ familiarity with the topic, viewers’ fatigue, or

> distraction in the viewing environment. We strongly advise creating a second image whenever the number of elements exceeds seven. (p.24)

According to my past experience as a fine art major, many oriental paintings indeed carry more than seven elements, but they do not affect the appreciation by the viewer. The simplicity of some arts is due to the aesthetic reason rather than human perceptive capability.

Further, the term “element” is not clearly defined. In a computer environment a menu bar usually contains several sub-menus. Is a single sub-menu or the entire menu bar considered an element? In a 3D plot is every axis treated as an element? Or should we perceive all three axes as a single element? The rule of “seven plus and minus two” is concerned with chunking--short term memory is limited by the number of chunks of meaning. When the short term memory is overloaded, humans can blend information or perform selective perception. For instance, in the computer environment where I am writing this paper now, there are far more than seven elements on the screen, but I do not lose my orientation. I can group all choices in the menu bar as a single object, then select a particular sub-menu I need and temporarily ignore all the rest. By the same token, I can blend all axes of a 3D plot as a multi-dimensional representation, then choose the axes of interest for manipulation. The application of “seven plus and minus two” to data visualization needs more theoretical refinement and empirical verification.

Nevertheless, it is the right direction to conceive that the application of visualization is a function of data complexity. When the data set is univariate and contain a small sample, Scott suggested (1992) that histogram is superior to many other graphs. But this is no longer true when the researcher faces “the curse of dimensionality” (p.5). Some visualization methods are inherently deficient in dealing with large data sets such as star plots, Chernoff faces, and radar plots. For example, in Figure 61 a radar plot shows 150 subjects and 6 variables. It is obvious that the graph is too entangled for pattern-seeking.

Number of observations. In order to manifest a latent pattern out of a data set with many observations, different smoothing techniques should be employed. There are three major kinds of bandwidth choices, namely, fixed kernel estimators, the nearest neighbor approach and variable kernel estimators. In the fixed kernel estimator the bandwidth remains the same throughout all subjects. The nearest neighbor approach assigns weights to bandwidth according to sparsity of the data near the point of estimation. The last method is a combination of both (Bowman & Foster, 1993). It is important to notice that different smoothers impose different assumptions on the data. Even simple graphs such as contour plots may appear differently owing to different smoothing algorithms. For example, S-Plus and XLISP-STAT produce completely different shapes of density smoothing, even if the same bandwidth and kernel estimator are used in both software (see Figure 62a and 62b). Moreover, we usually expect a contour plot will have round isolines. But the triangle smoothing method will yield an angular contour plot as shown in Figure 63. Silverman (1986), Hardle (1991) and Scott (1992) have complete demonstrations referring to Kernel density estimation.

Number of variables. While encountering multi-dimensional data sets, the visualizer can apply dimension-integration or dimension-reduction techniques to make the data more interpretable. Dimension integration retains all variables (dimensions) but portrays them as a single object (Carswell & Wickens, 1987; Barnett and Wickens, 1988). On the other hand, dimension reduction cuts down the number of variables and presents summary components. Gabriel biplots and Ideal Summary Plots are good examples of such reduction.

A Gabriel biplot (Gabriel, 1981) is a visualization technique for principle component analysis, which is available in SAS/JMP. To use a Gabriel biplot, the data should be considered as a matrix, in which the column space represents the subject space while the row space represents the variable space. “Biplot” simply means to plot the subject and variable spaces at the same time. In the example shown by Figure 64, the vectors labeled as P1 and P2 are eigenvectors, which depict the eigenvalues. The longer their length is, the more variance it can explain. Also, the user is allowed to specify the number of components and rotate them visually. In Figure 64, R1 and R2 are rotated factors. In this way the researcher can determine whether the unrotated or rotated vectors pass through as many points as possible to retain the simple structure.

Dimension reduction is also applicable to regression. Cook and Weisberg (1994) suggested that a three-dimensional regression model can be compressed into a two-dimensional one. Both DataDesk (Velleman, 1992) and R-code (Cook & Weisberg, 1994) are capable of performing this reduction. The following example is illustrated with the use of R-code. Before rotation, data points seem to scatter all over in the three-dimensional space, as shown in Figure 65a. When the researcher rotates a 3D spin plot, s/he may discover a strong linear trend as shown in Figure 65b. This perspective of the plot can be treated as an Ideal Summary Plot. As long as the linear trend holds, this Ideal Summary Plot can replace the full 3D plot without loss of any statistical information. The R-code can record the screen coordinates, which are the linear combinations of variables. To check the adequacy of the summary plot, the analyst can select a vertical slice of the plot and then look behind the slice by turning the points side way. If the selected points within the slice look like a horizontal band, the summary plot is considered adequate. The R-code can produce a slider bar to examine sliced segments bit by bit (see Figure 65c).

Before applying these techniques, the researcher should ask, “Why so many variables?” Some researchers started with many variables and then applied techniques such as stepwise regression and factor analysis for data reduction. Stevens (1992) criticized that these methods are very sample-specific and the result is subject to instability. Rather than beginning the analysis and visualization of multi-dimensions, we should evaluate the meaning of each variable. Perhaps the most effective dimension-reduction technique is simply to drop the irrelevant items!

Data Distribution

Many researchers prematurely adopt parametric tests without looking at the distributional structure of the data. The common rationale is that in the long run the central limit theorem (CLT) would restore the equilibrium--different types of members of the population would eventually have a proportional representation in the sample (Tversky and Kahneman, 1982). Thus, it was mistakenly believed that the CLT can magically make the parametric test legitimate even if the data distribution is non-normal. This misconception was clarified by Yu, Anthony, and Behrens (1995). It is essential for a researcher to obtain the distributional information before proceeding to further parametric procedures. Afterwards, the analyst may select subsequent visualization models corresponding to the data structure of normality or non-normality.

For a normal distribution, viewpoint dependent graphs such as a surface plot is acceptable because the obscured side is symmetrical to the front. Viewpoint independent techniques such as stereo-ray glyphs, volumetric model, image plot, contour plot, or a combination of those, are favored while analyzing a non-normal data set.

Skewed distributions. If non-normal data are skewed, a visualizer may apply interactive software to conduct data transformation and check the result immediately. Rasmussen (1989) asserted that data transformation is beneficial because it can yield more accurate Type I error rates and substantially increase power. On the other hand, some researchers are strongly opposed to data transformation because they believe that inferences based upon the transformed figures cannot directly apply to reality. This argument overlooks the fact that data visualization is a process of analogy-making. Some other researchers such as Motulsky (1993) prefer nonlinear regression to transformation. He argued that if the data has no experimental error, the linear regression line should fit the data well. Transformation is said to distort experimental error and so violates the assumption of linear regression. Tiny scatter around the theoretical curve can be transformed into huge scatter around the line. As a result, the parameters that results from linear regression are not optimal. Nevertheless, if the analyst chooses to use a non-linear regression, visualization tools such as the “NonLinearFit” package in Mathematica (Wolfram, 1991), nonlinear regression in R-code (Cook & Weisberg, 1994) and in SAS/JMP (SAS Institute, 1989) are beneficial in checking how well the non-linear model fits the data. Figure 66 shows an example implemented in SAS/JMP.

Bimodal and multimodal distributions. If the non-normal data are bimodal or multimodal, visualization techniques for discriminating groups and examining interactions can be helpful. Some data are inherently non-normal such as the racial composition of schools. This can be illustrated by a profile of grade schools in Illinois from 1991-1992. The data set contains several variables such as the percentage of white teachers and black teachers. In America, Caucasians are the majority and it is expected that the percentage of different races would not have normal distributions. However, this non-normal data can guide further investigation--the researcher may visualize the racial composition with other variables such as social economic status (SES) and achievement scores. Take percentage of black teachers as an example. By using a histogram the researcher can clearly discriminate two groups: few-black-teacher schools and half-black teacher schools (see Figure 67). Also, two outstanding groups, almost-all-white teacher schools and half-white teacher schools, emerge out of the variable “white teacher percentage” (see Figure 68).

By using scatterplot brushing, students from half-black teacher schools can be clustered in one color while students from few-black teacher schools are grouped into another segment. Then the interaction between performance of students and SES can be examined conditional on racial composition. The same strategy can also be applied to assess performance of students and SES between almost-all-white teacher schools and half-white teacher schools.

Summary

On some rare occasions relaxing the conception of data type may facilitate creative use of data visualization. For example, the inverse scree plot for cluster analysis is distribution-free (non-parametric) and relatively N-independent (Lathrop & Williams, 1989). We may put aside data distribution and complexity while using inverse scree plots. However, in most situations, data format, complexity, and distribution govern the appropriateness of the technique.

Data complexity is noteworthy because this aspect is swamped with misconceptions. Evaluation of the effectiveness of high-dimensional visualization techniques would not be fruitful if we do not take data complexity into account. Moreover, we should be careful not to over-apply cognitive psychology such as "seven plus and minus two."

Conclusion

This paper discussed the alignment framework relating to graphical format, task type, and data type. While previous research has examined isolated aspects of graphics (mostly perceptual issues), examination of the interaction of these aspects have been overlooked. Although the user interface issue in graphics is receiving increasing attention from psychologists, many psychologists tend to pinpoint specific graphical elements such as color, brightness, and motion, rather than evaluating the appropriateness of certain graphical techniques for specific task and data combinations. Although some researchers (e.g. Wickens et al., 1994) evaluated the efficacy of specific visualization techniques, mis-aligned graph, task, and data types make their generalization questionable. To rectify these problems, this alignment framework begins the move toward a set of comprehensive and practical guidelines for social scientists to apply scientific visualization. Moreover, this logical analysis calls for subsequent research to empirically verify the claim that the visualization effectiveness is determined by the proper alignment of graph type, data type and research goal.