The Benefits of Statistical Visualization in an Immersive Environment

Laura Arns[1], Dianne Cook[2], Carolina Cruz-Neira[1]
[1]Iowa Center for Emerging Manufacturing Technology
Iowa State University, Ames IA 50011-1210
[2]Department of Statistics
Iowa State University, Ames IA 50011-1210

Abstract

We have created an immersive application for statistical graphics and have investigated what benefits it offers over more traditional data analysis tools. This paper presents a description of both the traditional data analysis tools and our virtual environment, and results of an experiment designed to determine if an immersive environment based on the XGobi desktop system provides advantages over XGobi for analysis of high-dimensional statistical data. The experiment included two aspects of each environment: three structure detection (visualization) tasks and one ease of interaction task. The subjects were given these tasks in both the C2 virtual environment and a workstation running XGobi. The experiment results showed an improvement in participants' ability to perform structure detection tasks in the C2 to their performance in the desktop environment. However, participants were more comfortable with the interaction tools in the desktop system.

1. Introduction

Often, statistical data consists of sample sets measured on more than two variables. For example, a data set may consist of water samples, where each sample was tested for various contaminants such as sulfur, chlorine, ammonia, phosphorus, and other pollutants. Such data is called high dimensional or multivariate data, and the number of variables is denoted as p. Because p is generally greater than two, the two-dimensional plots created by traditional statistical packages do not adequately convey all the information contained in such data sets. Many approaches have been used to visualize high-dimensional statistical data. While many traditional static sorts of displays such as multiple histograms, scatterplots, Chernoff faces and parallel coordinate plots are included, sophisticated graphics software exploits interactive and dynamic querying of data to provide more powerful tools for understanding high-dimensional relationships. Currently, workstation-based tools such as XGobi [5] are used to view data that has been sampled on multiple variables, referred to as multivariate data. These tools display 1D and 2D graphics on a traditional computer monitor. However, virtual reality devices offer the capability to display graphics in stereoscopic 3D, allowing the user to better use another dimension of information. To date, there has been little virtual reality research in statistical graphics. We believe that alternate display methods enabled by virtual environments (VE) may be useful in analyzing this type of data. This experiment was designed to test the benefits and limitations of our statistical application.

When determining what type of VR device would be appropriate for our statistical graphics application, we considered the fact that statisticians often wish to collaborate when analyzing data. Additionally, they require the ability of having an expert or instructor to demonstrate a technique or a particular feature in the data to a group of colleagues or students. Thus it is extremely important that the VR device used allow for multiple viewers. Projection-based systems are ideal for implementing applications that require multiple observers. The C2, developed at the Iowa Center for Emerging Manufacturing Technology, is such a device. It consists of three projected walls, and a projected floor, creating a 12'x12'x9' virtual environment [3].

In order to determine if the C2 provides an improved exploratory data analysis tool over XGobi, we developed a statistical analysis application for the C2, which provides much the same functionality as XGobi. We describe these functions in section 2, and the reasons these functions are important in section 3. Section 4 explains the XGobi tool, and section 5 covers the C2 statistical application. In section 6 we introduce the tests conduced to compare the environments, and the remainder of the paper covers the results, their interpretation and significance, and future work that could be conducted as a result of this experiment.

2. Statistical Background

Statisticians often need to analyze large sets of data points. Because finding relationships among the data points (also called detecting structure in the data) is often easier to do visually than numerically, computer graphics is used as a visual aid for exploring the data. This may involve identifying anomalies in the data set, that is, sample points that have unusual attributes. For example, in a data set consisting of water samples, a sample containing a concentration of nitrates which is much higher than other samples in the region might be an anomaly. The point may be anomalous along only one variable, or along several. Conversely, instead of trying to locate anomalous points, a statistician may be interested in trying to find samples which have something in common. Perhaps a group of cities in the same state all have similar climate. Again, the samples may be similar in more than one way. Section 3 discusses several types of information a statistician may wish to identify.

There are several graphics methods a statistician may use to analyze a data set. In sections 2.1 and 2.2 we describe two such methods: brushing and the grand tour, and explain how they aid statisticians in discovering relationships among the data.

2.1 Brushing

Brushing is a method of marking data points to distinguish them from each other, to indicate interesting features, and to identify groups of data with similarities. Data points displayed on a computer screen can be brushed with different shapes (called glyphs, which may also vary in size) and colors. For example, when we find a group of cities with similar climate, we can brush them all as red circles to differentiate them from the other data points (which may be displayed with some default color and shape).

Typically, a two-dimensional computer screen is used to display the data sets. However, because multivariate data contains more than two variables, it is difficult to display all the information. We can create different views of the data by choosing to display information about only certain variables. Perhaps one view of data about cities would display the climate rating vs. the housing cost. Another view would then plot the cities with latitude vs. longitude (a map). We can simultaneously show both views on the computer screen, each view in its own window. When we discover points in one view which are interesting for some reason, we would like to know how these points are arranged in the other view. The method of brushing points in one view and having points in another view appear brushed similarly is called *linked brushing*.

Newton [4] first introduced brushing in the late seventies, and it has become a popular tool for statistical graphics applications. For a detailed analysis of brushing, see [6].

2.2 Grand Tour

As discussed above, high-dimensional data consists of more than two pieces of information for each sample, so the data can no longer be simply plotted on the XY Cartesian plane. We could create multiple 2D plots of the data, each one comparing two of the variables. However, for high-dimensional data, the number of plots created in this way may become large. Additionally. this way of viewing the data is limiting, because there may exist relationships in the data which involve multiple variables in different combinations.

A better approach is to create projections from the high-dimensional space into two or three dimensions for display.

Different combinations of the variables are used to create the different projections. The *grand tour* algorithm uses such an approach. The tour creates many possible projections of the data, and then steps through them, displaying each for a short amount of time. During the tour, the data plot appears to rotate. This effect happens because the projections are chosen so there are small differences in the data points locations from one projection to the next. The process can be compared to moving a light source around a three-dimensional object, creating various two-dimensional projections (shadows). Software packages that have grand tour tools include XGobi [5], ExplorN [1], and XploRe [2]. Details of the grand tour algorithm can be found in [7].

Because brushing and the grand tour are so often used in statistical analysis, we felt they would be good indicators for comparing our environments. These methods are implemented in both XGobi and the C2 statistical analysis application described in sections 4 and 5. In section 6 we discuss how the methods were utilized by the experiments.

3. Structure in High-Dimensions

The statistical methods discussed in section 2 are often used to identify structures in a high-dimensional data set. Such structure may reveal interesting characteristics in the data samples. Three typical structures are discussed in this section: clusters, intrinsic dimensionality, and radial sparseness.

3.1 Clusters

A *cluster* in a plot is a group of data points that are located in close proximity to each other in space. Clusters are said to be unconnected if the clusters do not overlap or touch each other. Data points that do not belong to a cluster are called outliers. Figure 1 shows a data set containing two clusters and some outliers. Because a cluster represents a set of sample points with several common features, it is useful to know if the data is clustered, and how many clusters are present in the set. For example, the cities data set may contain a cluster of sites which all have excellent climate and high cost of living. It is obviously important to discover and note this correspondence. Additionally, it is useful to note whether samples that appear as a cluster in one projection remain clustered in other views or projections. Consider a set of stream chemistry information: Points representing water samples which always remain clustered and have high concentrations of certain chemicals could indicate nearby pollution.

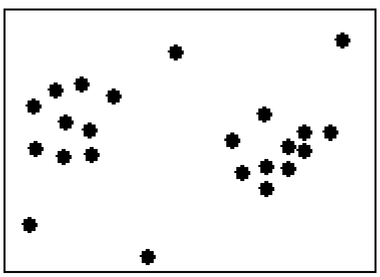

Figure 1: A data set with two clusters and some outliers

3.2 Intrinsic Dimensionality

Often high-dimensional data doesn't fill the entire high-dimensional space. For example, with two variables, there may be a linear relationship so that one variable essentially explains the variation in the other variable (for example, a chart plotting people's height vs. their weight usually has this type of structure). In statistical terms, this data is said to be intrinsically one-dimensional. Figure 2 shows a data set that is *intrinsically* one-dimensional.

A data set with three (or more) variables may all lie on a two-dimensional plane in three-space. This data is then intrinsically two-dimensional. If the intrinsic dimensionality is less than *p* (the total number of variables in the dataset), the number of variables needed to explain the data can be reduced. The ability to reduce the number of variables makes it easier for a statistician to find relationships among the remaining variables.

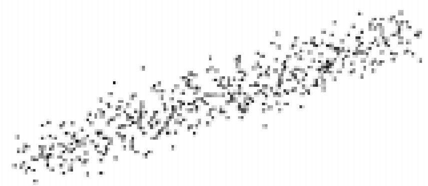

Figure 2: An intrinsically 1D data set

3.3 Radial Sparseness

a for what types??

For certain types of statistical data, it is often interesting to determine whether the data points lie on the surface of an object such as a sphere, or whether the points are spread throughout the object. The first case is referred to as *radial sparseness*. Figure 3-a shows a radially sparse data set on a circle. Figure 3-b shows a data set that is not radially sparse.

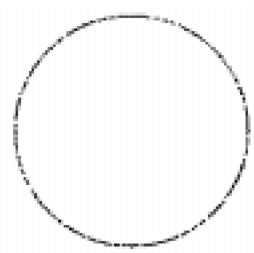

Figure 3-a: A radially sparse data set

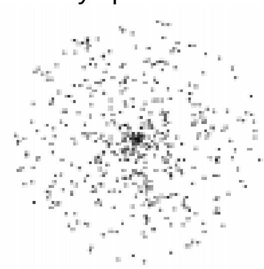

Figure 3-b: A non-radially sparse data set

4. XGobi

As mentioned in section 2.2, a variety of programs exist for analysis of statistical data. XGobi is a popular application well-known in the statistical community and readily available as a public-domain tool [5].

XGobi is a tool for visualizing and interacting with high-dimensional data, designed to run on UNIX®[1]

[1] *UNIX* is a registered trademark of UNIX Systems Laboratories.

workstations with the X Window System™[2]. It creates plots of multi-dimensional data on a two-dimensional display device (workstation monitor) which the user can interact with through conventional desktop devices such as keyboard and mouse. Although the plots are displayed in the two-dimensional monitor screen, they may be created by a projection from high-dimensional space as in the grand tour, and thus contain information about many variables. The application includes many features for statistical analysis, some of which are not described here as they do not directly relate to the virtual environment we developed or the experiment performed. For a thorough description of XGobi, see [5]. In XGobi, the primary features we are concerned with are the ability to link multiple plots, view data "rotated" through a grand tour, or stop the tour to brush points. The brushing tools provide several colors and glyphs.

Figure 4 shows an XGobi screen. The plot area is shown in the center. The data set displayed contains three clusters, one of which is being brushed. The circles on the right represent the variables. Each one indicates how the associated variable is being used by the grand tour to create the current projection. Also, in the lower left corner of the plot area this information is repeated by means of axes, one corresponding to each variable.

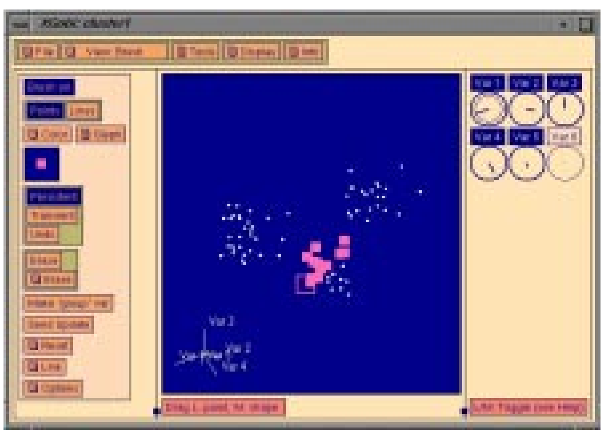

Figure 4: An XGobi screen

The major XGobi functions are accessible from the pull down menus accessed from the buttons appearing at the top of the screen. These menus are labeled "File", "View", "Display", and "Info". Like most "File" menus, this one provides input/output abilities. For example, the user could chose to save files containing information about the current colors or glyphs of the data points. The "View" menu determines the current method of displaying the data, such as a two-dimensional dot plot, grand tour, or brushing mode. More advanced statistical methods such as projection pursuit may be accessed through this menu. The "Tools" menu may be used to open additional windows for linked brushing, to obtain listing of variables, and other statistical tools. The "Display" menu allows users to chose whether to display data points and/or lines (some data sets may have pre-defined lines between pairs of data points). And the "Info" menu provides general information about XGobi and also provides access to the on-line help.

The following describes how a typical work session with XGobi might proceed. First the user views the data as each projection is displayed by the grand tour. The speed of the tour can be controlled using one of the tools on the left of the screen. When the user finds an interesting feature in the data and wishes to brush the data points, he switches to brushing mode through the "View" menu. This action causes the buttons to the left side of the screen to change from grand tour controls to controls with which the user chooses a brushing color, glyph shape, and glyph size. In Figure 4, the user has chosen to brush with medium sized squares. A paintbrush appears in the plot area. In Figure 4 the paintbrush is shown as a small square. To brush data points, the statistician uses the mouse to drag the paintbrush over the plot area. Any points touched by the paintbrush are displayed with the glyph the user has selected.

5. C2 Statistical Program

The statistical program developed for the C2 is loosely based on XGobi and includes many of the same features and tools such as brushing and the grand tour. For compatibility purposes, the C2 application can read and write files in the same format as XGobi so that data analysis results from one can be seen in the other. Figure 5 shows the C2 statistical application.

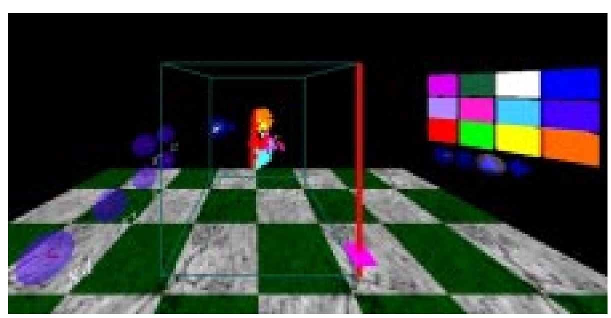

Figure 5: The C2 statistical program

Interaction with the tools is done through simple hand gestures, such as pointing with the index finger, using

[2] *X Window System* is a trademark of MIT.

a CyberGlove. We assumed that pointing hand gestures would work well as a selection operation in our environment, because humans are accustomed to pointing to indicate a selection, both in the real world and with window-type computer environments. A similar assumption was made when we chose a closed fist gesture to indicate operations involving moving an object, because it emulates the familiar action of grasping a real world object. For more abstract operations, a joystick with three buttons (referred to as a wand) is used. We will see in later sections that these assumptions produced unexpected results in the experiments.

When the C2 application runs, the user is presented with a three dimensional environment which includes the tools for operating the grand tour and brushing features. The large wireframe cube in the center is where data is plotted. In the center of this box in Figure 5 we see some plotted data points. On the front right edge of the cube is a slider pole used to control the speed of the grand tour. With one pointing finger, a user can drag the slider handle up or down the pole to increase or decrease the speed of cycling through the data projections. With a closed fist gesture, the entire plot can be moved or resized by grabbing one of the box's vertices. The data box and its contents can also be rotated using a special rotation tool.

To the left of the plot box are the variable spheres. These operate in a manner similar to XGobi's variable circles, but in three dimensions instead of two. By pointing at a sphere, the user may toggle the axis on or off in order to control which variables are included in the current plot and grand tour.

To the right of the plot box are the painting tools. Three rows of colored boxes are used for determining the current brush color. The user picks a color by pointing at the block containing the desired color. Four different glyphs (box, pyramid, sphere, and star) are located directly beneath the color boxes. A glyph may be selected by pointing at it. To the right of these tools is a short slider bar. By moving the slider on the pole, the user can change the size of the glyph.

The user enters the painting mode by pressing a wand button. The paintbrush appears around the user's hand when using the painting mode (it can be thought of as the "area of influence"). A user selects points to brush by moving the paintbrush over the point. When a point is in contact with the paintbrush, it becomes colored with the chosen color and glyph.

Many additional tools are available to the user through a menu that can be accessed by pressing a wand button. This menu is part of a larger virtual reality windowing system for the C2 currently under development [8]. Additional types of menus and windows are available, and can be easily placed inside other C2 programs. By using one pointing finger, the user may press any of the buttons on the menu. The menu includes the ability to turn on and off the display of predefined connecting lines between data points, or bring up a help file with information on how to use the program. Several save buttons are also provided. These allow the user to save information about the color, shape, or size of each glyph, or to save information about the current projection.

6. Experimental Design

To compare the C2 application with XGobi, we designed a set of exercises to be performed in each environment by a group of subjects. The testing group was recruited from Iowa State University. Subjects were asked to volunteer if they had an interest in either statistics or virtual environments. Most were graduate students in statistics, with some XGobi experience. Only a few subjects had previous experience with any type of virtual environment. The number of volunteers was 15. While this number may not be large enough to prove conclusively that one environment is better than the other, it is enough to provide strong indications of the potential benefits that each environment provides, and areas of the applications which need improvements.

Each subject completed four tests, each taking from one to five minutes, in each environment. Three of the tests were designed to compare the users' ability to visually identify important features of the data set, and concerned clustering, intrinsic dimensionality, and radial sparseness. The forth test was an interaction test and used the brushing technique. For every subject, the order in which tests were taken, and the data set used for each test were chosen randomly. Before beginning the tests, each subject was given a short tutorial that included five minutes of training in each of the C2 and XGobi.

6.1 Visualization Tests

The main goal of these tests was to quantify a user's ability to correctly identify structure in a high-dimensional data set in both the desktop and immersive environments. Because virtual environments require a large investment in terms of money, time, and personnel, it is important to determine whether there is any significant improvement in users' performance to justify the use of the environment. If users are more accurate when analyzing data in the virtual environment, then using the environment for analyzing future data sets could prevent errors and increase productivity by decreasing the time needed for analysis. The three tests were designed as follows;

Cluster Test: Six data sets were available for this test. Two of the sets were relatively small, containing only seventy four data points with six variables, while the other four sets contained five hundred data points and five variables each. The number of clusters present ranged from two to five, and all data points appear with the same shape and color. The subject was asked to determine the number of clusters in the data set by viewing the data in a grand tour for one minute. Brushing of data points was not allowed.

Intrinsic Dimensionality: Two data sets, each containing five hundred points and five variables, were used for this test. One of the data sets was intrinsically 1D and the other 2D. Subjects were allowed to view the data set in a grand tour for one minute in order to determine if the data set was intrinsically 1, 2, or 3D. Again, no brushing of data points was allowed.

Radial Sparseness: Two data sets were used for this test, each composed of randomly chosen points from a 4D sphere. In one data set the points were chosen to lie on the surface of the sphere (radially sparse) and in the other the points were distributed throughout the interior of the sphere. Again the subject was allowed to watch the data in the grand tour for one minute, without brushing data points, to determine which set he was viewing.

6.2 Interaction Test

Working under the assumption that directly interacting with the data in three dimensions using our hands is more "natural" and "intuitive" than interacting with it on a two dimensional monitor, we designed an experiment to measure users' performance in both environments.

For this test, the data set used for the cluster test was used again. The subject was instructed to stop the grand tour when he observed a cluster, and brush the data points in the cluster. If several clusters were present, the subject was asked to only brush one of them. Up to five minutes could be used to complete this task (though not one of them came close to this limit). The subject was allowed to select any color and glyph desired, including the default (C2 only). The time used in selecting a brush color and glyph was recorded, as was the time used to brush the cluster.

7. Results

7.1 Visualization Tests

Table 1 shows the results of the visualization tests. From this table, we can see that subjects perform nearly twice as well on the cluster test in the C2 than using XGobi, correctly identifying the number of clusters present 80% of the time using the C2, but only 46.7% of the time with XGobi. Subjects did marginally better in the C2 on the sphere test, answering correctly 66.7% of the time in the C2 compared with 53.3% in XGobi. The dimensions test was performed equally well in both environments, with an 80% correct response rate.

Table 1: Results of Visualization Tests

Subject Number	Test Order[b]	Sphere Test[c]		Dimension Test[c]		Cluster Test[c]	
		C2	XGobi	C2	XGobi	C2	XGobi
1	2	1	1	1	1	1	0
2	1	0	1	1	1	0	0
3	1	1	0	1	1	1	1
4	1	0	1	1	1	1	1
5	2	1	1	0	0	1	0
6	2	1	1	0	1	1	0
7	2	1	0	1	1	0	0
8	2	1	1	1	1	1	1
9	1	0	0	1	0	1	0
10	2	0	0	0	0	0	1
11	1	1	0	1	1	1	0
12	1	1	0	1	1	1	0
13	2	1	1	1	1	1	1
14	1	0	0	1	1	1	1
15	2	1	1	1	1	1	1
Correct Response Rate		0.66	0.53	0.80	0.80	0.80	0.46
Standard Deviation		0.47	0.49	0.40	0.40	0.40	0.49

[a] Cluster Size: 1=Small, 2=Large
[b] Test Order: 1=XGobi First, 2=C2 First
[c] Response Accuracy: 0=Incorrect, 1=Correct

Interestingly, subjects were more likely to respond correctly in the second testing environment on both the dimensions test and clusters test, even though they did not necessarily observe the same data set in both environments. This suggests there may be some learning curve to correctly perceiving structure in high-dimensional data. Also, participants who were tested first in the C2 environment were more accurate on the sphere test in both environments. Perhaps viewing the sphere data in the C2 first gave subjects a better understanding of this type of data, enabling them to use this knowledge in the XGobi environment.

7.2 Interaction Test

Table 2 shows the results of the brushing test. The time used in selecting a brushing tool is about the same in both environments (19 seconds), but actual brushing times are significantly lower using XGobi (50.9 seconds vs. 12.9 seconds). Table 3 shows that subjects with previous virtual reality experience have much lower brushing times in the C2 than subjects with no previous experience (61.2 seconds vs. 30.4 seconds).

Table 2: Results of Interaction Tests

Subject Number	Size[a]	Previous Virtual Reality[c]	Select Time[d]		Brushing Time[d]	
			C2	XGobi	C2	XGobi
1	1	2	5	22	58	12
2	1	2	5	15	12	5
3	1	2	0	12	25	6
4	1	2	0	8	32	5
5	1	1	30	7	12	3
6	1	1	10	12	50	8
7	2	1	38	24	48	68
8	2	1	0	10	63	17
9	2	1	0	46	120	8
10	2	1	0	39	35	15
11	2	2	0	15	25	8
12	2	1	0	8	90	6
13	1	1	38	20	88	8
14	2	1	10	28	75	22
15	1	1	0	20	31	3
Correct Response Rate			19.4	19.0	50.9	12.9
Standard Deviation			14.1	11.0	30.3	15.6

[a] Cluster Size: 1=Small, 2=Large

[b] Test Order: 1=XGobi First, 2=C2 First

[c] Previous Virtual Reality Experience: 1=None, 2=Limited

[d] Time in Seconds. Averages and standard deviations were calculated using only non-zero entries

Table 3: Brushing and selecting times in the C2

Previous Virtual Reality Experience	Average Brushing Time	Average Selection Time[a]
None	61.2	25.2
Some	30.4	5

[a] Time in Seconds: Averages were computed using only non-zero times

8. Conclusions

We anticipated that subjects would perform better on visualization tasks in the C2 than with XGobi. The experimental results support this hypothesis: Subjects are consistently able to identify clusters better in the C2, and they perform better on the sphere test when they view the data in the C2 first. One possible explanation is that the C2 virtual environment provides the ability to visualize data in "true" three dimensions by immersing the user in the data, instead of the two dimensions of visualization provided by a workstation. When viewing high dimensional statistical data, the added dimension allows a user to better make decisions regarding the structure of the data.

These visualization tests may not be conclusive, but they indicate that the C2 environment has the potential to significantly improve user's productivity over XGobi for structure and feature detection tasks involved in analysis of high dimensional data. They also suggest that more substantial studies should be performed to examine how people perceive high-dimensional structure generally, which may be useful for designing new visualization tools.

The higher interaction times shown in Table 2 for the C2 suggest that subjects had more difficulty interacting with the C2 than with a traditional workstation running XGobi. However, the large standard deviation indicates that the results need to be analyzed carefully. On a first look at the results, it is obvious to note that most people have significantly more experience interacting with a desktop environment than with virtual environments, which justify the overall performance results. However, a closer look at the results combined with the subject's backgrounds revealed that subjects with any level of virtual reality experience were able to interact more easily and quickly with the C2. This also suggests that the learning curve for a virtual environment is not particularly steep due to the intuitive nature of the environment. Another factor we believe impacted the results is that users may have felt intimidated by the C2. When encountering a new technology for the first time, many people are uncomfortable and unsure of their actions, therefore being more careful in their interaction with the equipment, which, again, impacts their performance in the system. Finally, the type of interaction we provided in the C2 statistical program may not be the ideal method. We expected that direct hand interaction would be a very natural and intuitive option in particular for first time users, but the results did not reflect this expectation. Many factors may have contributed to this, some related to the hardware we used, such as C2 system performance and accuracy of the hand tracking devices, while others are related to the method we chose for brushing of data points. Wills [6] discusses various interaction paradigms and their benefits.

9. Acknowledgments

The research reported in this article has been funded by the National Science Foundation grant DMS9632662.

The authors thank the testing subjects for volunteering their time to the experiments, and the staff and students of the Iowa Center for Emerging Manufacturing Technology for their assistance.

10. References

[1] Carr, D.B., Wegman, E.J. & Luo, Q. ExplorN: Design considerations past and present, Technical Report 129, Center for Computational Statistics, George Mason University. 1996.

[2] Härdle, W., Klinke, S. & Turlach, B.A., XploRe: An Interactive Statistical Computing Environment, Springer-Verlag, New York, 1995.

[3] Iowa Center for Emerging Manufacturing Technology website. www.icemt.iastate.edu

[4] Newton, C. Graphica: From Alpha to Omega in Data Analysis, in P.C.C. Wang, ed., 'Graphical Representation of Multivariate Data', Academic Press, New York, NT, pp 59-92, 1978.

[5] Swayne, D.F., Cook, D. & Buja, A. 'XGobi: Interactive Dynamic Graphics in the X Window System', *Journal of computational and Graphical Statistics*. 1998.

[6] Wills, G. 'Selection: 524,288 Ways to Say 'This is Interesting'', *Journal of Computational and Graphical* Statistics. Forthcoming.

[7] Asimov, D. "The Grand Tour: A Tool for Viewing Multidimensional Data," *SIAM Journal of Scientific and Statistical Computing*, Vol. 6, No. 1, 1985, pg. 128-143.

[8] Heath, D. J. "Daniel's HomePage," accessed Jan 12, 1998. http://www.cs.iastate.edu/~dheath