COMBINING INTERACTIVE VISUALIZATION AND STATISTICS TO DETECT PATTERNS IN ENVIRONMENTAL DATA

Jean-Yves Hervé, Qing Liu
University of Rhode Island, Kingston, United States
Matthew Nicholson
U. S. Environmental Protection Agency, United States
Liliana González, Thomas Mather
University of Rhode Island, Kingston, United States
jyh@cs.uri.edu

One of the biggest challenges statisticians face when working with non-statisticians on applied problems is to be able to effectively communicate the statistical results. In this paper we discuss the use of interactive visualization as a tool to present the relationship between a binary response and a set of explanatory variables. The visualization system we present allows users to "manipulate" directly, dynamically, and interactively their data set. At a first level, this allows to integrate visualization with a classical statistical analysis by providing interactive 3D views of the data set. Beyond its potential use as a straightforward visualization tool, this new system opens up interesting possibilities for exploring data visually, by its better exploitation of the human visual system. The paper presents an example of exploring visual relationships between environmental variables and the presence/absence of Lyme disease in Rhode Island.

INTRODUCTION

One of the most difficult tasks faced by statisticians engaged in multidisciplinary collaboration is how best to communicate statistical concepts effectively and efficiently. In this paper we propose that a successful way to convey statistical information to a non-statistical audience is by means of visualization tools. Like the old saying “a picture is worth more than a thousand words,” in a statistical context we suggest that “a visualization tool is worth more than a thousand lines of statistical output.”

The use of graphical representation of data has a long history, beginning in the second half of the eighteen century with simple plots such as the scatter plots and time-series plots. But it wasn’t until Tukey’s pioneering book *Exploratory Data Analysis* (Tukey, 1977) that visualization was first used in a more structured way. Today visualization is employed extensively in data presentation as well as in data analysis. An important first step in data analysis is developing “a feel” for the data at hand and visualization methods offer the best vehicle to achieve this goal. Graphics are not only useful in exploratory data analysis but also in presenting results. “Graphics are instruments for reasoning about quantitative information” (Tufte, 2001).

Information visualization is the technology that tries to understand very large data sets by using enormous visual bandwidth and the remarkable human perceptual system. Information visualization focuses on data sets lacking inherent 2D or 3D semantics and therefore also lacking a standard mapping of abstract data onto the physical space of the paper or screen. A number of well-known techniques visualize (partially) such data sets, including x-y plots, line plots, and histograms. These techniques are useful for data exploration but are limited to relatively small low-dimensional data sets. A large number of information visualization techniques have been developed over the past decade, allowing visualizations of ever larger and more complex, or multidimensional, data sets (Keim, 2001; Soukup and Davidson, 2002). Among these, geometrically-transformed display techniques aim at finding useful information by using geometric transformations and projections. There is, however an infinite number of possibilities to project high-dimensional data onto the two dimensions of a standard display. “Projection pursuit” (Huber, 1985) attempts to locate projections that satisfy some computable quality of interest. A particular projection pursuit technique known as the GrandTour (Asimov, 1985) aims at automatically finding interesting projections or at least helping the user to find conclusion.

The GrandTour represents an interesting attempt at better exploiting the remarkable capabilities of the human visual system. Users can scan, recognize and recall images rapidly, and can detect subtle changes in size, color, shape, movement, or texture. They can extract efficiently “group” information: alignment, symmetry, “sameness,” and “togetherness.” At the same time, humans are easily deceived by a number of optical illusions based on relative dimensions,

intensity, or accidental alignments. These characteristics of the human visual system constitute one of the bases for rules of good graphic design as stated by Tufte (2001). Most geometrically-transformed display techniques fail to exploit human perception because they present static representations. We may live in a 3D world that we perceive through 2D sensors, but we do so by interpreting *dynamic* 2D projections of 3D scenes. 3D-to-2D projections typically exhibit ambiguities that can only be resolved by a change of point of view, which is what the GrandTour proposes. The effectiveness of this technique is limited by the fact that viewpoint changes are *passive*, that is, controlled by the system rather than by the user. However, it has been clearly established by Bajcsy (1988) and others that *active vision* is an essential aspect of scene understanding. It is by actively changing their viewpoint, eye vergence, focus, etc. that viewers acquire a better understanding of what they are looking at (even when what they are looking at is a flat picture or graph).

In this paper we discuss the use of interactive visualization as a tool to present the relationship between a discrete (binary) response and a set of explanatory variables. The visualization system we present allows users to "manipulate" directly, dynamically, and interactively their data set. This new system opens up interesting possibilities for exploring data visually, by its better exploitation of the human visual system. The ability of humans to detect subtle changes in size, color, shape, movement, or texture, and, beyond that, to extract "group" information (alignment, symmetry, "sameness") are considerably superior to that of any of today's computer-based system. This remarkable ability can be somewhat harnessed by letting the user's visual system guide the choice of transformations applied to the data set. The software tool we present here adds this new dimension to the visualization of large and complex data sets by allowing users to *manipulate directly and interactively* the data.

STATISTICAL PROBLEM

To contribute to our understanding of the spatial and temporal patterns of Lyme disease and the environmental factors associated with its presence/absence in the state of Rhode Island - an endemic area- a five-year study of the disease was conducted. To evaluate the spatial and temporal dynamics of the disease we gathered information on all cases of Lyme disease reported to the Rhode Island Department of Health (RIDOH) between 1993 and 1997. Each case included in the analysis met the Council of State and Territorial Epidemiologist/Centers for Disease Control and Prevention surveillance case definition for Lyme disease.

The identification of the risk factors associated with Lyme disease required comparison of Lyme disease patients to non-disease patients that were randomly selected from each of the 39 cities and towns of Rhode Island and with sizes proportional to their corresponding population sizes. The total sizes in both groups were the same for each of the years of the study. The home address of all reported Lyme disease cases and that of the non-disease cases was assigned geographic coordinates (as a surrogate for county) using a geographic information system (GIS). A circle of radius of 500m centered on the GIS-derived location became the estimate of the location of addresses and several environmental characteristics at the locations were measured and used in the analysis. Kriging was used to estimate tick abundance (Nymphs/hr) for each of the addresses and different surfaces were used for each of the years separately. To gather information for land uses at each of the locations of interest, the land use/land cover map developed by the Rhode Island Geographic Information System was used and the proportion of each habitat type within the 500 m radius circle was used as an estimate of the peri-domestic environment for that address. Other variables included were area classification of address (urban/rural), elevation (in meters), distance to closest fresh water (in meters), closest distance to forest edge (in meters), distance from location to primary road (in meters), distance to coastline (as a surrogate for climate) and total amount of forest edge within the 500 m radius circle around the home location (in meters). After including only the relevant land use classifications, there were a total of fourteen independent variables, where two were categorical and twelve were of the continuous type.

STATISTICAL RESULTS

From 1993 to 1997, 1371 Lyme disease cases were reported to the Rhode Island Department of Health. In the year 1993, 149 cases were reported, in 1994 they increased to 301 cases and decreased over the 1995 year to 246, with numbers increasing again in 1996 (366) and

decreasing once more in 1997 (309). In general, the Lyme disease cases increased in the even years and decreased during the odd years (with respect to the previous year) with a clear increasing trend over time. That is, Lyme disease in the years of the study showed a two-year cyclical pattern. The analysis of the presence/absence of Lyme disease indicates that the disease is more prevalent in rural than in urban areas. Moreover, there is a clear indication that Washington County is an endemic area, with 65.35% of the reported cases in the five-year period of the study. The number of cases reported in Kent and Providence Counties were basically the same over the study period (about 14% each), Newport reported 5.62% of the cases and Bristol County only 1.6% of the cases.

Logistic regression was used to assess the relationship between the presence/absence of Lyme disease in Rhode Island and all the other environmental variables considered in the study. Assessment of the associations between presence/absence of Lyme disease and each of the individual variables indicated that the strongest relationships, in order of importance, correspond to the *xy*-coordinates (used as a surrogate for county), total edge, percentage build-up and nymphs per hour. Model building indicated that the most important variables in predicting the presence/absence of Lyme disease in Rhode Island, after adjusting for the other effects present were, in order of significance, the *xy*-coordinates ($p<0.0001$), total edge ($p<0.0001$), year ($p<0.0001$), area classification (rural/urban, $p=0.0014$), percentage build-up ($p=0.0031$) distance to edge ($p=0.0066$) and nymphs per hour ($p=0.0482$). The likelihood ratio test for comparison of models indicated that there was no statistical difference between the full model and the final selected reduced model ($\chi^2=4.74$, df=6, $p=0.5776$) that included the previous eight listed variables.

VISUALIZATION TOOL

We have developed a first prototype of an interactive 3D visualization tool. The tool was developed in C++ with the Metrowerks CodeWarrior IDE, and using the OpenGL graphical library and the GLOW framework for user interface elements. It runs on the Windows 32 and Mac OS platforms. Data points are displayed as colored, tilted triangles. The user can at any time associate a specific column of the data to one of the 6 display dimensions (X, Y, Z, Hue, and two angles), as shown in Figure 1.

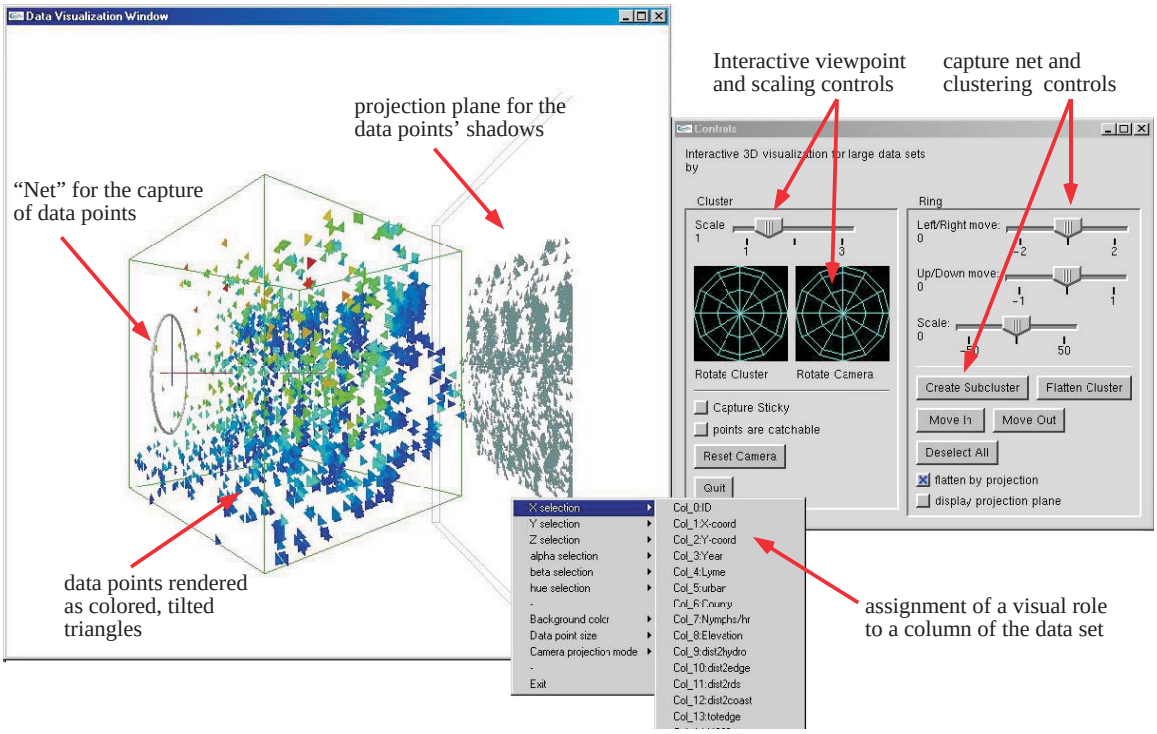


Figure 1: The interface of our first prototype. Left: Data visualization window. Right: Control panel allowing the user to change the viewpoint interactively, select data points, divide the data into clusters, etc.

The choice of the triangle as the basic shape to represent data points as opposed to, say, the sphere or the cube can be explained by the performance gains it offers: The triangle is the fundamental shape of computer graphics. A cube will be rendered as at least 12 triangles while a sphere will be rendered as a group of at least a hundred triangles, typically several hundreds. In addition, using the triangle as a basic shape allows the display of two additional data columns as the orientation of the triangle. The orientation of a uniform sphere is indeed imperceptible and the orientation of a cube among many other cubes is hardly perceptible at all.

Our tool allows us to perform "classical" visualization of the data. For example, we can plot under the form of maps and view side by side the presence/absence of Lyme disease and nymphs/hr, and notice that Lyme presence and high rate of nymphs/hr are mostly observed about Rhode Island's South County (Figure 2). However, the correlation between these data is much more striking in interactive 3D. Figure 3 attempts to convey some of that impression in a series of 3 flat screenshots. In these diagrams, the *x* and *y* data points are still plotted as the *X* and *Y* coordinates, but the nymphs/hr data is now displayed as the *Z* coordinate, while presence/absence provides the hue information.

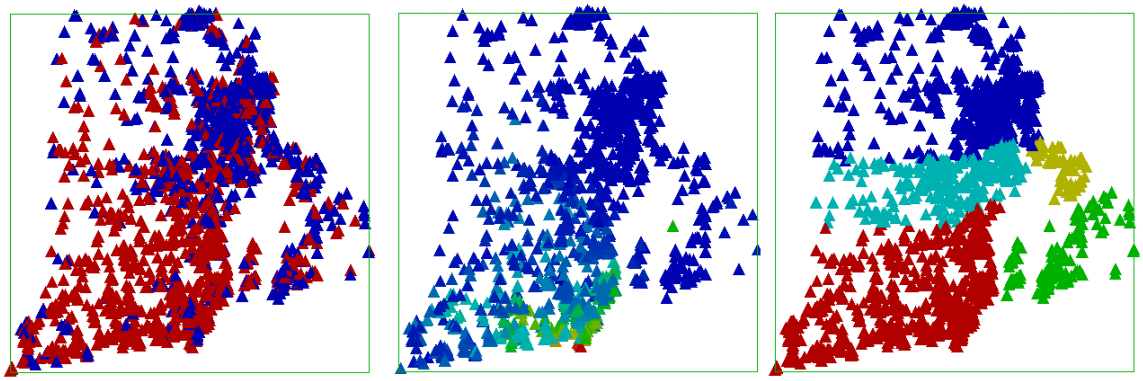


Figure 2: "Flat" representations (*x* and *y* drawn as *X* and *Y* and *Z* axis collapsed) showing presence/absence of Lyme disease (left), nymphs/hr (middle), RI County (right)

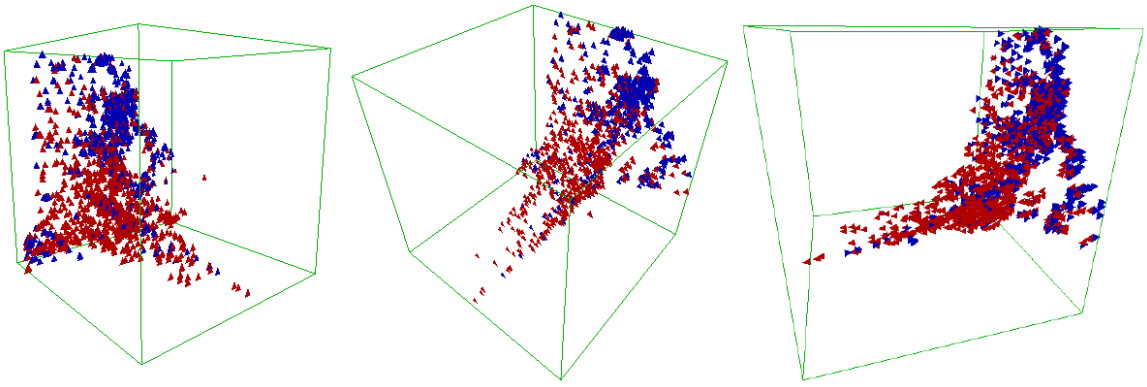


Figure 3: Screenshots from 3 different viewpoints of the mapping [*X*=*x*, *Y*=*y*, *Z*=nymphs/hr, Hue=presence/absence of Lyme disease]

It is interesting to note that in the chosen final model, the least significant variable was nymphs/hr and the most significant one the *xy*-location. This can be explained by the high correlation between the y-location and nymphs/hr, as visualized in the above screenshots. Once the y-coordinate was included in the model and being the most significant of all the variables, the significance of what would seem the most relevant variable in modeling Lyme disease became almost non-existent.

Our tool also allows us to represent the same data in a different visual role. For example, Figure 4 shows presence/absence of Lyme disease plotted against total edge-the second most important variable in our logistic regression model-, *x* and *y* being still plotted as *X* and *Y*. In the left and middle views, presence/absence is displayed as the Hue, while total edge provides *Z* information. In the right view, presence/absence of Lyme disease is used as *Z*, which splits the

graph into two planar maps on which we display total edge information as the Hue (in this view we changed to an orthographic projection).

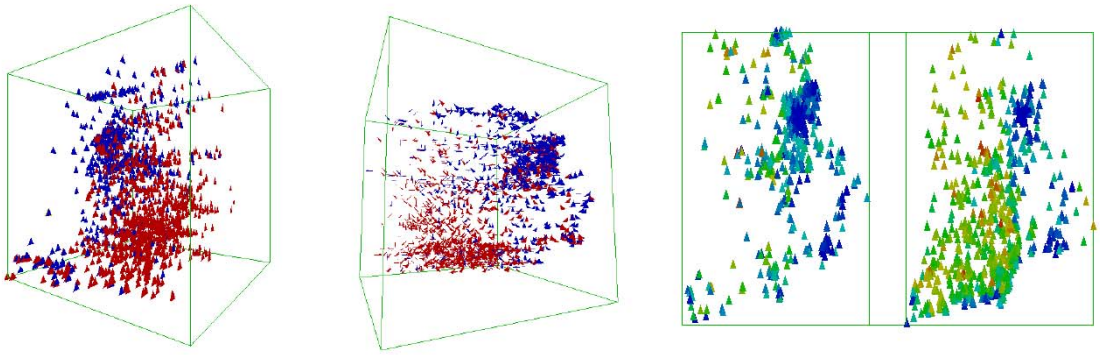


Figure 4: Left, Middle: *Z*=total edge, Hue=presence/absence of Lyme disease; Right: *Z*= presence/absence of Lyme disease, Hue=total edge

The next two most important variables associated with the presence/absence of Lyme disease in Rhode Island were year (1993-1997) and area classification (rural/urban). Our tool is now used to display these associations. Figure 5 below shows presence/absence of Lyme disease plotted against year-a categorical variable-, with *x* and *y* still being the *X* and *Y* coordinates. The left and middle pictures show the graph in perspective projection for 2 different viewpoints. The right graph is a flattened orthographic projection with the most recent year (1997) on the left and the most ancient (1993) on the right. Figure 6 shows the dependent variable plotted against area classification. This is a categorical variable as well, with two levels (urban/rural).

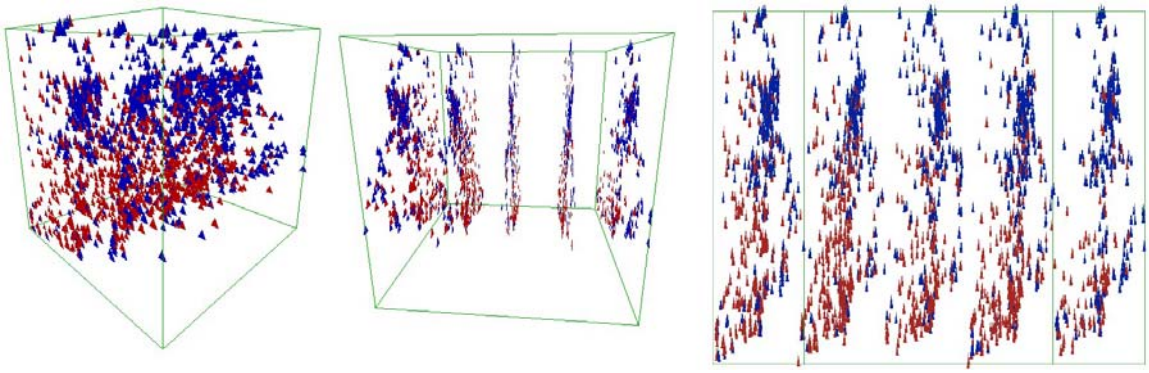


Figure 5: Left, Middle: *Z*=year, Hue=presence/absence of Lyme disease; Right: flattened orthographic projection with the most recent year on the left (1993) and the ancient on the right (1997), Hue=presence/absence of Lyme disease

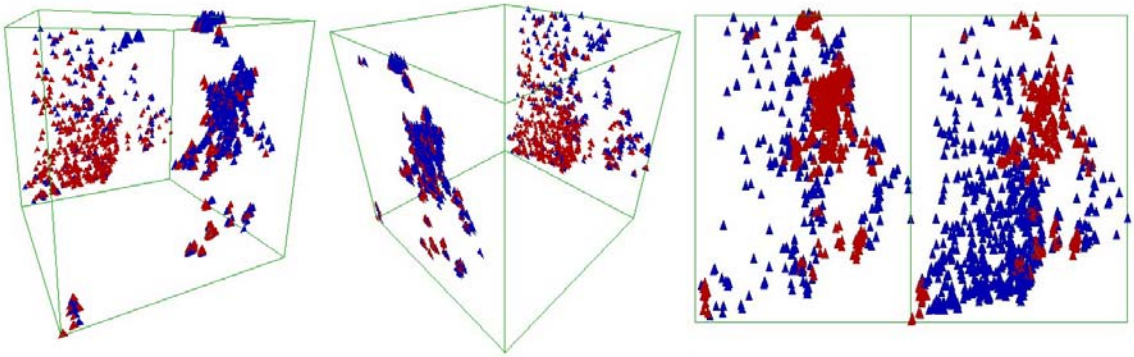
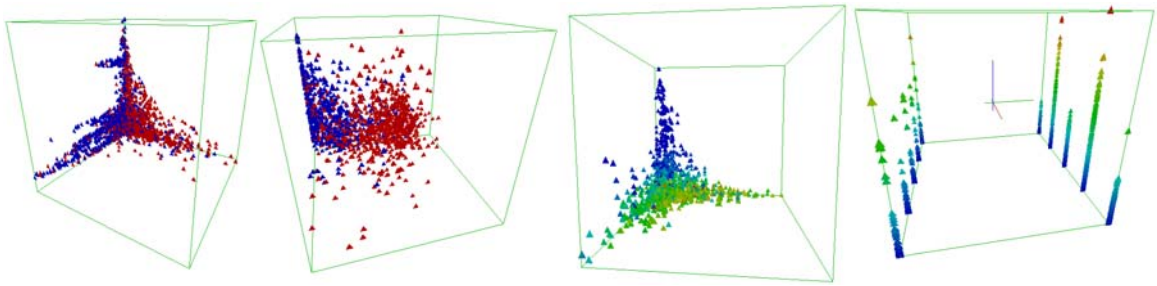


Figure 6: The left and middle graphs show Z=area classification (urban front, rural back) and Hue=presence/absence of Lyme. The right graph shows a collapsed orthographic projection of the mapping *Z*=Lyme (absence left, presence right) and Hue=red for urban, blue for rural.

EXPLORATORY VISUALIZATION

The tool also allows us to explore the data set directly, using solely our visual abilities as a guide to what constitutes "interesting" arrangements of the data as we assign different visual roles to the various columns of our data set. The power of this approach is also its main limitation. Because it is to some extent an uninformed study of the data, it cannot pretend to lead to an interpretation. Rather, its objective is to identify patterns in the data that would justify a guided statistical analysis. Figure 7 below gives examples of such interesting patterns that seem to indicate some form of correlation of the data that would deserve further investigation.



(a) (b) (c) (d)

Figure 7: (a) *X*=nymphs/hr, *Y*=closest fresh water, *Z*=distance to edge, Hue=Lyme
(b) *X*=total edge, *Y*=closest fresh water, *Z*=nymphs/hr, Hue=Lyme
(c) *X*=elevation, *Y*=distance to edge, *Z*=nymphs/hr, Hue=total edge
(d) *X*=Lyme, *Y*=Hue=Nymphs/hr, *Z*=Year

CONCLUSION

The use of an interactive visualization tool to aid in the exploratory phase of a statistical analysis, as well as in the presentation of statistical results, has been discussed in this paper. Sometimes visualization can fully replace the need for probabilistic inference, but in other cases, visualization is not enough and probabilistic inference is needed to help calibrate the uncertainty of a less certain issue (Cleveland, 1993). In the particular application presented in this paper a combination of inference with the visualization tool proved useful in developing a full understanding of the important environmental variables associated with the presence/absence of Lyme disease in Rhode Island.

The use of this tool is by no means limited to the problem presented here. The tool we developed for interactive visualization is flexible and could be applied to many other situations. As for future work, an extension of the present tool is being developed to interface with the programming language *R*.

ACKNOWLEDGEMENTS

The partial support of this research by the National Center for Research Resources, NIH, through grant number RR16457 is acknowledged. The authors also wish to thank Daniel Azuma whose freeware GLOW toolkit was used to build the GUI of the software presented.

REFERENCES

Asimov, D. (1985). The GrandTour: A tool for viewing multidimensional data. *SIAM Journal of Science and Stat. Comp.*, 6, 128–143.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, 76(8), 96-1005.

Cleveland, W. S. (1993). *Visualizing Data*. Summit. NJ: Hobart Press.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2), 435–474.

Keim, D. (2001). Visual exploration of large data sets. *Communications of the ACM*, 44(8), 176-188.

Soukup, T. and Davidson, I. (2002). *Visual Data Mining: Techniques and Tools for Data Visualization and Mining.* Indianapolis, IN: Wiley Publishing, Inc..

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (second edition). Cheshire, CT: Graphics Press.