Dead Graphs -- RIP

(How Interactive Graphical EDA Techniques Can Markedly Enhance the Analysis/Display of Our Data)

Dave DesJardins, Statistical Research Division, U.S. Bureau of the Census

Abstract:

The purpose of this paper is to highlight the important difference between "dead" graphs and "live" graphs. With live graphs, users can dynamically change the display of their information based on a quick, interactive cross tabbing of the underlying/input data. These live graphs will be a key component of a Census Historical Atlas CD-ROM that is to be published by the Census Bureau next year. With "live" graphs, we hope to bring an incredible capability to the general public -- to gain a quantum leap in our ability to see and understand our data.

The concept of this new data display technology is based on the existing, new, and very powerful interactive graphics software packages that have already revolutionized the statistical analysis of our data. These packages create "live" graphs -- which are the fundamental building blocks of a new *interactive* Exploratory Data Analysis (EDA) methodology course currently being taught on an ongoing basis by the author. This methodology is based on easy-to-learn 3rd generation point-and-click software tools (SAS/Insight® and JMP®). Using this software, the author has devised an "EDA Plan of Attack" that has become a key component of the interactive techniques of his courses. The plan of attack is likened to a battlefield strategy used by the military -- wherein the unique strengths and weaknesses of planes, tanks, and artillery (tools) are used in combination with a practiced strategy (techniques) to gain an objective. Using powerful, newly developed live graphs, this EDA strategy has been devised to reveal the hidden features of our data. Thus, it uses a combination of the best features of a number of types of graphs (the tools: box plots, scatterplots, data profile plots, etc.) -- in combination with the interactive features of SAS/Insight and JMP software (the techniques: animation, brushing, dynamic color assignment, 3D rotation, etc.)

These interactive techniques are powerful in another way as well. Often, graphs can make even complex statistical concepts easily understood. Accordingly, the author has sometimes taught his EDA class to Subject Matter Specialists who have little statistical background training. A key goal of these classes is to give our Analysts a fundamental *understanding* of the nuances of their data -- as opposed to some of our conventional statistical techniques that simply blindly edit "outlier" data. And, since these Subject Matter Specialists are the individuals closest to the data, acquiring these powerful tools has led to any number of breakthroughs in the analysis of our data.

Shown in figure 1 in this printed form is a "dead" Data Profile graph devised by the author. Although it is a quite powerful graph even in this "dead/printed" format, this static display is but a small fragment of the information it could reveal in a "live" interactive format -- where you could, for instance, brush across outlier points. This paper will attempt to explain the extent of the information that could be revealed in an interactive session designed for the analysis as well as the display of our data.

A Dead Graph -- RIP

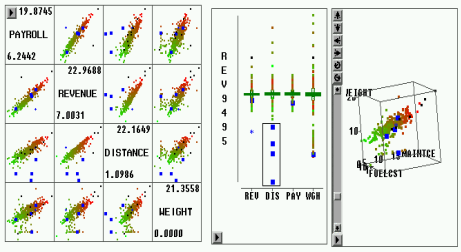

Figure 1

Background --Conceptual Framework:

Data Display:

The author is currently the project manager of an interactive Historical Atlas CD-ROM project. This CD is envisioned as a showcase product of the Bureau -- highlighting a number of major Census data themes in an *interactive graphic format* -- *"live" graphs.* The framework of this atlas would be digitized images from the outstanding 1870 and 1890 Census Atlases. The 1870 and 1890 Census Atlases were the first and third atlases of our nation -- landmark publications, long noted for their outstanding maps and graphs. However, these digitized atlases would be but a mere skeleton of the information that will be incorporated in the project.

Conventional chart books and atlases produced for publication in this country rarely seek to provide more than a compilation of simple charts, graphs, and general reference maps. Where an attempt is made to include the display of information illustrative of some specific aspect of American life, it is usually presented as "dead" graphs in a rigidly two-dimensional, aerial format and, often, relegated to a peripheral section of the volume. Our CD project, however, would be a very different kind of atlas in that it would also feature "live" graphs. Using hyperlinks, it would draw together the immense quantities of statistical information available from our current censuses (from our web page) and show these data in an historical perspective. Again, these data would be portrayed in an interactive graphics format -- to produce a vision of the United States that has literally never been achieved before.

Data Analysis:

We live in a very dynamic society -- with significant changes, such as those represented by the Internet, all around us. From a *data analysis methodology* perspective, the advantages of EDA methodology and "live" graphs (over blind computer algorithms that rely on fixed relationships, "typical" companies/industries, normal distributions, and preset formats) are only too evident when it comes to change. For instance, comparing variables and using ratios for data editing and analysis requires a very good understanding of the often unique and unpredictable relationships between each of the variables. These relationships can vary markedly for different point cloud clusters -- for instance, for industries with different SIC codes, or simply for smaller or larger companies. Additionally, time series variances such as business cycles and periodic anomalies could also systematically affect these ratios. Basic, historical relationships for data sets can also vary substantially over time-- making fixed, traditional methodology useless. By using powerful interactive graphics tools in combination with an individual trained in EDA methodology, these problems (which are often □invisible□ to fixed computer algorithms) can be promptly found and dealt with. In addition to quickly identifying rather obvious outliers, these techniques often find hidden "inliers".

Introduction:

"In the land of the blind, a one eyed man is king." (Anon.)

This paper focuses on the fruits of a true “revolution” in the analysis and display of our data. Like the introduction of gunpowder to warfare, "live" graphs are the key ingredient of this revolution. This paper will show a few examples of how these live graphs can significantly enhance the analysis and display of our data. (Previous papers by the author have highlighted examples of where EDA has shown itself to be significantly better at data editing, data analysis, identifying outliers and inliers -- and discovering flaws in our traditional data analysis methodology at the Census Bureau.)

Four key factors contribute to this "revolution" in data analysis/display -- and make the introduction of "live" graphs in conjunction with EDA methods at this time a momentous opportunity:

* Today's news media is awash with reports of how increasingly powerful computer technology is forever changing our world. These same very powerful computers also give our data analysts (and the general public) the ability to generate hundreds of graphs in a matter of mere minutes -- a feat that would have taken months to do just a few years ago. These "live" graphs can provide us with a "window" into these computers -- for the first time truly unlocking their real potential.

* Likewise, new, very powerful, general purpose, graphical data analysis/display software packages also give our Subject Matter Specialists the *easy to use (point-and-click)* tools required to quickly generate these graphs. Once formulated for our key applications, the graphs from these packages give us unique new insights into our data -- and the ability to interact in real time with the data that is displayed therein.

* These software packages can often be acquired at very low costs. For instance, the student version of SAS Institute's JMP software (JMP-IN®) only costs about $60. This "best buy" software -- with a very powerful statistical data analysis and data manipulation capability -- requires no annual license renewal fee and comes with a 500+ page statistical data analysis manual!

* By using the above hardware and software tools, we have developed new graphical forms and special techniques that greatly enhance the speed and efficiency of our data editing/analysis tasks. Accordingly, our Analysts no longer need to waste their time (and valuable subject matter expertise) trying to edit their data with "blind" CPU- formatted algorithms and cumbersome, boring, tabular printouts. In addition, since graphs have the extraordinary ability to communicate across a wide area of expertise, they can often make even very sophisticated statistical concepts clear to laymen. Thus, our Statisticians can now more quickly and effectively explain to our Analysts the fundamental concepts behind these new graphical data analysis techniques -- and focus the majority of their efforts instead on improving our methodology. The next step in this process is to use these "live" graphs to more effectively present these data to the general public.

DATA DISPLAY -- DESIGN CONSIDERATIONS:

Unfortunately, Census data are all too often published in a tabular format. Worse, even when the data are portrayed in a graphic format, those graphs are then published in the conventional "dead-graph" format. Even worse still, this (conventional) format for atlases and chart books is usually a simple repeat of the same basic set of charts and maps, page after page. No attempt is made to show the interrelationships between the individual bits of information chosen for display -- and, often, much of the data could have been presented with greater clarity and precision if left in numerical tables. The result is neither good art nor a particularly useful reference.

The essential concept behind our "live graph" Historical Atlas CD, in contrast, is an approach to graphic design that is concerned primarily with a *multivariate interactive display/analysis* of our data. It would show "drill down" and "comparative/ juxtaposed" multi-formatted displays -- with the goal of showing the graphic display of interactive cross tabs of data. This format would thus display how these databases and subsets interact. This methodology will allow the user to explore the relationships between the bodies of data, and the broader context of the factual information it presents.
Another aspect of the conceptual framework of these displays will be to demonstrate the ways in which great quantities of precise detail can be combined in order to illustrate or reveal central principles and relationships. As such, the major displays can be viewed at several levels of detail depending upon the expertise and interest of the individual reader.
Obviously, good design and intuitive/interactive software are a necessary prerequisite for the success of this project. However, as complex as many of those displays/designs would be, the alternative would be individual pages and pages of tables, graphs, and charts -- at the end of which no reader would be able to perceive the interaction between the parts as readily as he or she could perceive through a single, multivariate, "live" graphic display.

For a simple example of the limitations of conventional displays, let's imagine a "dead" map displaying the number of elderly in the USA. This single-dimensioned graphic display provides a very poor overall picture of the elderly. This map would clearly show California with the highest concentration of elderly. However, California is a "young" state -- and would rank way below the top if we were to show the elderly as a percent of the total population. Thus, we could, and should, at least, display two "dead" maps in each of these formats to help the user gain a better understanding of the nuances of these data. As will be seen below, in a "live" format, we would offer "percent elderly"

and "total elderly" as just two of a number of default bars for these data.

DATA DISPLAY -- AN EXAMPLE OF LIVE GRAPHS:

The proposed format for the Historical Census Atlas CD is not only live graphs, but live graphs in a multi-database (complex) format. In this format, the interactive software of the CD would allow users to gain a unique understanding of these data. Imagine, for instance, a drill down map of the USA. Further, imagine that users could simply click on any state and then county boundary to access a display right down to the Census tract level of the data. For this level of data, a key, preformatted, Census variable (with a set of associated bar charts) would automatically be displayed. By default, it would show, for instance, the geographic distribution of poverty for that area -- with bar charts showing, say, the race/ethnicity, the age/education, and the income distribution for the individuals in this geographic area. (An example of this is shown below.) Even in a static format, this would be a very informative display of Census data. However, the key feature of this graphic display would be that these graphs would be "alive" and would respond to the EDA technique of brushing. Using brushing, a user could simply click on any bar in the bar chart to highlight a data subset (say, the Hispanics) -- and the map would then automatically display where this group of individuals are located within the map display of this county. Further, this brushing would darken portions of the bars to highlight how this group of individuals makes up the components of the information in all the other graphs. By then clicking these subsets, we can sub-select from within this group in the other bar charts. Thus, in seconds, we can cross tab these data to see where, say, only the young, highly educated, low income Hispanics are located in this county.

In figure 2, we see a map displaying the distribution of the poor in the New York City area. (NOTE: For privacy purposes, these are fictitious data.) To illustrate the concept of live graphs, I will focus on the interactive nature of a set of related data in the form of bar charts that would automatically accompany this map. In figure 3, for this hypothetical area, we see bar charts showing the race/ethnic distribution, household income, number of children, age distribution, and gender of these individuals. Here we see (by the darker shading) that the "white" race category for this tract-level display has been selected. We can see (by the shaded portions of the other bar charts) that this white sub-category is well represented across all levels of household incomes-- and also seems to be proportionally represented by gender as well. (Here we only see a change in the bar charts -- in a "live" display, the map would likewise change to show the distribution of these "white" individuals.)

By simply clicking on any of the segments of the bar chart display, we can quickly cross tab these data to, for instance, take out the youngest and oldest whites. In figure 4, we have selected only the white individuals between the ages of 30 and 50. As can be seen from the portions of the bars that remain shaded, most of these individuals are part of the households with incomes between the values of $45,000 and $50,000.

Another example is shown in figure 5. Here we have simply clicked to select the tracts with the highest incomes for this geography. As can be seen from the shading of the other variables, these "rich individuals" are rather well represented across all of the races, but a larger proportion of this "rich" tract is composed of males.

The most important aspect of the above example is that these displays can be created in this interactive format in mere seconds. As such, this seductive methodology invites the user to further explore these data -- so a true, broader understanding of its implications can be gained. For instance, at this point, I would now ask the reader if she/he would be interested in clicking on the highest levels of the income distribution shown here? Imagine being able to quickly see where these high-income families are distributed on the map by age, gender and the other variables. Would the reader be likewise interested in seeing the distribution of these data for their own home town -- and then compare this display with another US area?

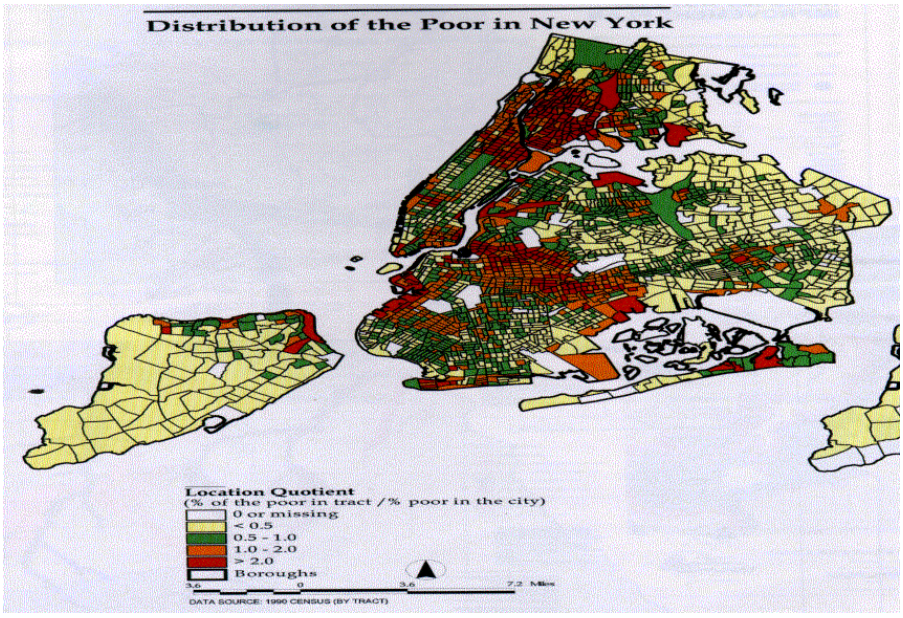

Figure 2

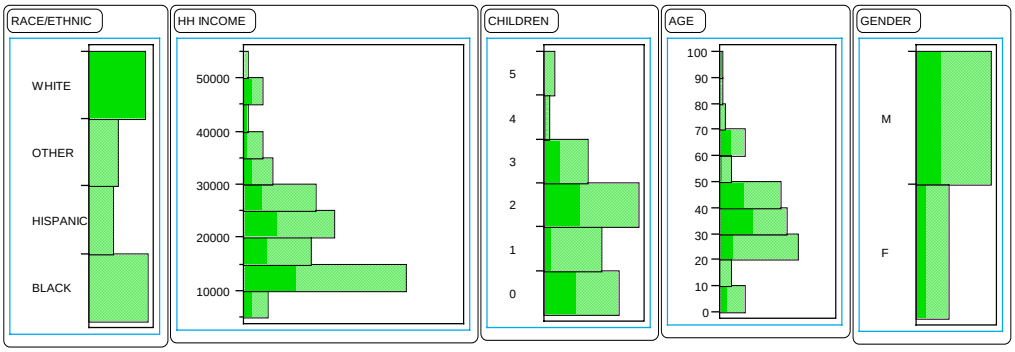

Figure 3

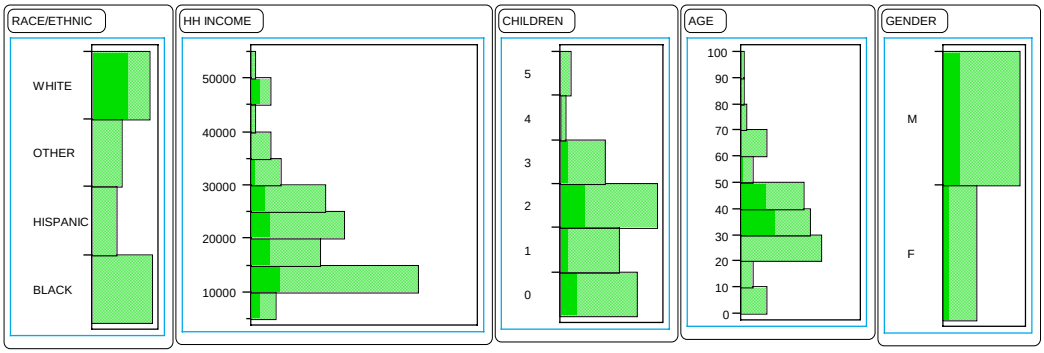

Figure 4

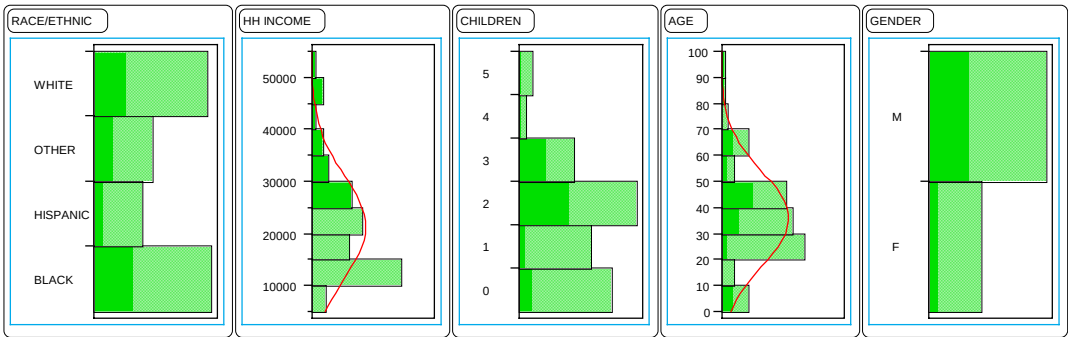

Figure 5

Figure 6

This example is, by no means, meant to be inclusive as to the data. Our display would not be solely restricted to these few demographic variables. The Census Bureau, the "Factfinder of the Nation", is

incredibly rich in the variety of the demographic, agricultural, and economic information we gather. For example, figure 6 is a copy of Census Bureau's web page. Imagine how these "live" displays could be incredibly enriched by allowing the reader to select from a dozen interrelated data sets. Imagine selecting from a menu that would include things like age/education/income, housing cost, type of business, type/productivity of agriculture, population density, welfare payments, migration trends, etc. -- and then to be able to cross tab and interact with these displays!

Data Analysis Methodology Example -- Leverage Plots:

Live graphs also play a significant role in the statistical analysis of our data. Leverage plots are another example of an EDA methodology that has proven itself to be very helpful in identifying outliers/inliers. For instance, in figure 7, we see another two scatterplots of values from our trucking company data (values are in log dollars). Here we see the reported log revenues versus reported fuel costs for these companies -- and their revenues versus reported leasing costs. Aside from a few obvious outliers and the fact that the lease cost data is a lot less correlated with revenues (than fuel costs); nothing jumps out at us.

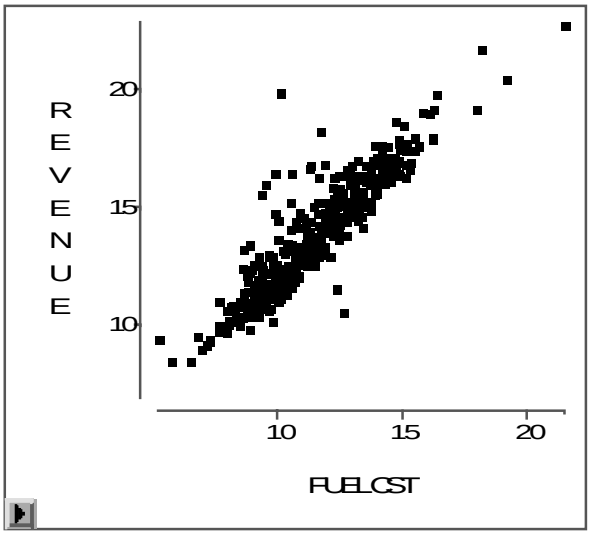

Figure 7 (a)

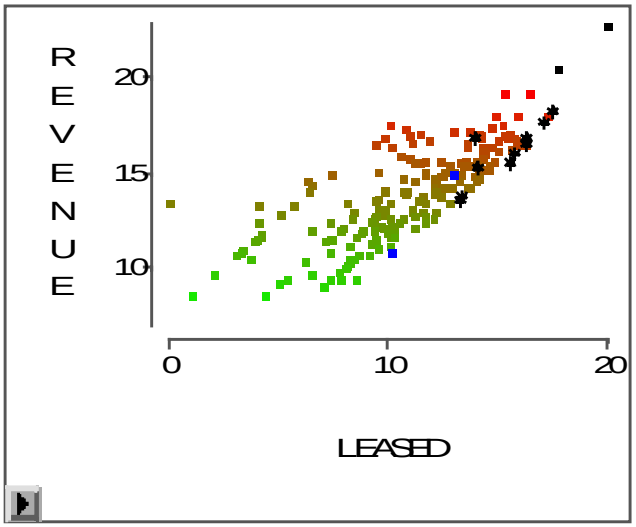

Figure 7 (b)

However, when we plot a fit (and the leverage plots) of these reported revenues versus the related variables (figure 8), we begin to suspect something. Even if we know nothing about leverage plots, we can see that there is a subset of very suspicious points in the lease cost plot (highlighted within the square with a *). These points also now show themselves to be very influential points in the leverage plots of payroll and fuel costs.

Returning now to a repeat of the original scatterplots of these data (figure 9), we can see how these points show up in the revenue versus fuel costs data -- as points that stand out with higher revenues and lower costs than the majority of these other "normal" companies. In the scatterplot of revenues and leasing costs, most of these points are clearly inliers -- until now, not the least bit suspicious points. They only show up along the lowest edge of these data -- having the lowest leasing costs and highest revenues. Again, my EDA course strives to go beyond "blind" algorithms that simply edit outliers to these teaching these techniques that give our Subject Matter Specialists an understanding of their data. Here, clearly, we can see a need to gain a better understanding of the trucking industry. It suggests that companies that lease vehicles need to be treated differently -- certainly, their different revenue profile would affect the imputation methodology that we use.

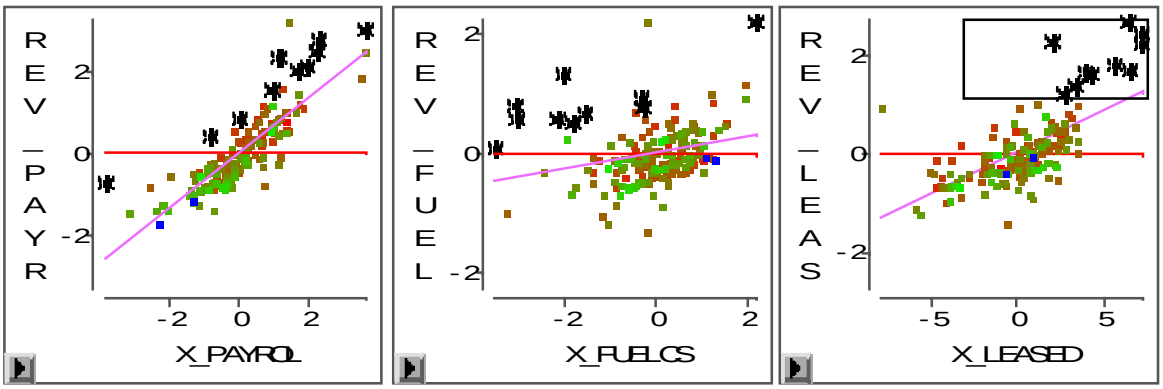

Figure 8

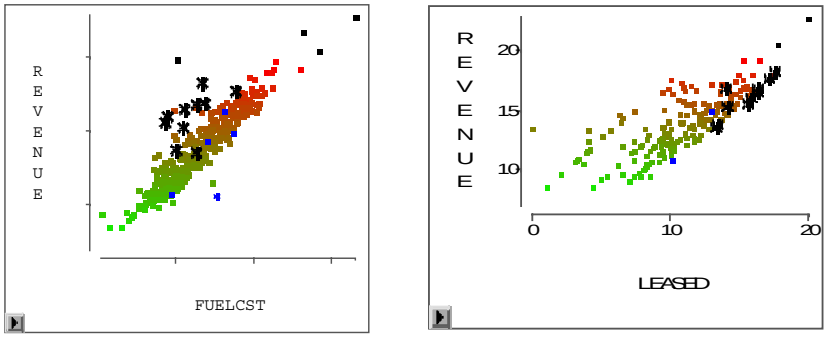

Figure 9

Summary:

This paper has highlighted the important difference between "dead" graphs and "live" graphs. These live graphs are the fundamental building block of a new interactive Exploratory Data Analysis (EDA) methodology currently being taught at the Census Bureau on an ongoing basis by the author. (The author also teaches this EDA methodology in a graduate level statistics course and as a 2-day short course in various non-government forums).

Using live graphs we have also shown how this data analysis methodology could also be a revolutionary new way to display our data to the general public. As was shown, live graphs can dynamically change the display of information by allowing a quick/interactive cross tabbing of input data by the user. The software for the Historical Atlas CD-ROM -- which will need to very user friendly for use by novice/casual users -- is currently under development. Hopefully this new CD software will reflect the power and function of the interactive EDA software (SAS/Insight and JMP) currently being used by the author in his classes.

Author Information:

David DesJardins
Statistical Research Division
Washington, DC 20233
Tel: 301-457-4863
david.l.desjardins@ccmail.census.gov

References:

Cleveland, William, (1993), “Visualizing Data”, *Hobart Press*.

DesJardins, David (1998), “New Graphical Techniques for the Analysis of Census Data”, *Statistics Canada Conference Proceedings*.

Fellegi, I.P. and Holt, D. (1976), “A systematic approach to automatic edit and imputation”, *Journal of the American Statistical Association* 71, 17-35.

Granquist, L. (1997), “Macro-Editing - The Aggregate Method Statistical Data Editing”, *UN Conference of European Statisticians Statistical Standards and Studies, Geneva (Switzerland).*

Hidirogou, M.A. and Berthelot, J-M. (1986), “Statistical editing and imputation for periodic business survey”, *Survey Methodology* 12, 73-83.

Hogan, Howard (1995), “How Exploratory Data Analysis is improving the way we collect business statistics”, *Proceedings of the American Statistical Association,* August 1995.

Sall, John (1990), "Leverage Plots for General Linear Hypotheses", *The American Statistician*, Volume 44, No. 4.