THE GRAPHICAL EDITING ANALYSIS QUERY SYSTEM

Paula Weir, Energy Information Administration, Robert Emery, SAIC, John Walker, SAIC
Paula Weir, 1000 Independence Ave., S.W., EI-431, Washington, D.C., 20585

Key Words: Exploratory, Visualization, Top down, PowerBuilder

BACKGROUND

In 1990 the Data Editing Subcommittee of the Federal Committee on Statistical Methodology released the Statistical Policy Working Paper No. 18, "Data Editing in Federal Statistical Agencies." The paper presented the subcommittee's finding that median editing cost as a percentage of total survey costs was 40% for economic surveys. The committee felt that the large proportional cost was the direct result of over identification of potential errors. Hit rates, the number of identified potential errors that later result in a data correction divided by the total number identified, were universally very low. As a result, a lot of time and resources were spent that had no real impact on the survey results. The report cites research by the Australian Bureau of Statistics concerning the use of graphical techniques to find outliers at both the micro and macro level. A similar graphical approach to editing used by the U.S. Bureau of Labor Statistics (BLS) for the Current Employment Survey is also described. The BLS Automated Review of Industry Employment Statistics (ARIES) system identifies true errors quicker and results in fewer man-hours to edit the data than the previous hard copy error listing report method.

Subsequent to the efforts of the Data Editing Subcommittee, a working group of analysts, research statisticians and programmers was formed within the Bureau of Census to examine the potential use of graphics for identifying potential problem data points in surveys. It was felt that the existing procedure of flagging cases failing programmed edits and reviewing each edit on a case-by-case basis, had three main disadvantages: 1) analysts see neither the bigger industry picture nor the impact of the individual data point on the aggregate estimate, 2) analysts, therefore, examined more cases than necessary, 3) edit parameters or tolerances were derived from previous surveys, implying constant relationships over time. The group felt that the tools of exploratory data analysis combined with subject matter specialists' expertise were well suited for identifying unusual cases. The working group developed a prototype that made use of box plots, scatter plots and some fitting methods, as well as transformations. Two other systems, the Graphical Macro-Editing Application at Statistics Sweden, and the Distributed Edited Deposits Data System Editing Project (DEEP) of the Federal Reserve Board, have further demonstrated the efficiency of graphical editing. The recommendations of the subcommittee included focusing on the need for survey managers to evaluate the cost efficiency and timeliness of their editing practices and the implications of technological developments such as microcomputers, local area networks, and various communications links, in conjunction with traditional subject matter specialists' expertise.

THE CONCEPT

The Graphical Editing Analysis Query System (GEAQS) is being developed by the Energy Information Administration (EIA) as a tool to reduce survey costs and reduce the amount of paper generated. It combines and builds on the features of the four other systems mentioned above, the ARIES system, the Census Working Group prototype, the Graphical Macro-Editing Application, and DEEP. The GEAQS borrows from the ARIES system the concept of an anomaly map which summarizes the relationship of various levels of aggregates and flags questionable aggregates through the use of color. This top down method of editing provides the user the ability to drill down through the aggregates to the respondent level.

From the Census Working Group prototype and recommendations, GEAQS makes use of the tools of Exploratory Data Analysis. Box and whisker plots summarize aggregate changes from the previous period to the current period through multiple boxes for the "children" of the select higher level aggregates. Further subaggregates are visible and identifiable within each box. Scatter plots are used to further drill down and display respondent and imputed data for the current period versus the last period for the select aggregate. Data points with high influence are indicated by color. High influence points that also visually deviate the most from the trend contribute the most to the overall change. Outliers of low influence, if not systematic, are not as cost effective to pursue, and they contribute to over editing.

GEAQS builds upon the need for a Windows' application as developed by Statistics Sweden. This allows the user to point-and-click on an aggregate in the anomaly map or the box and whisker plot, as well as a data point on the scatter graph. The user can take advantage of tool bars, dialogue boxes, and icons. Resizing is built in to enable the analyst to focus on particular parts of a graphic. Tiling, on the other hand, allows the analyst to maintain the previous graphic while operating on the next graphic of the same drill down effort. In order to maximize the usefulness of GEAQS to other surveys, GEAQS was developed as an object oriented PowerBuilder application that uses Pinnacle

graphics server to help generate the graphs. This reduces costs to modify or enhance GEAQS to operate on surveys other than the survey originally piloted. It will also allow for ease of integration with the rest of thesurvey's system.

In order to capture the recommendation of the Census Working Group that the system be acceptable to the people who use it, the development of GEAQS emulated the iterative user feedback process used by the Federal Reserve Board through testing by users at various stages of development. Unidentified requirements were quickly discovered and modifications made. By allowing analysts direct input throughout the process, this resulted in a product that is more useful to the analysts.

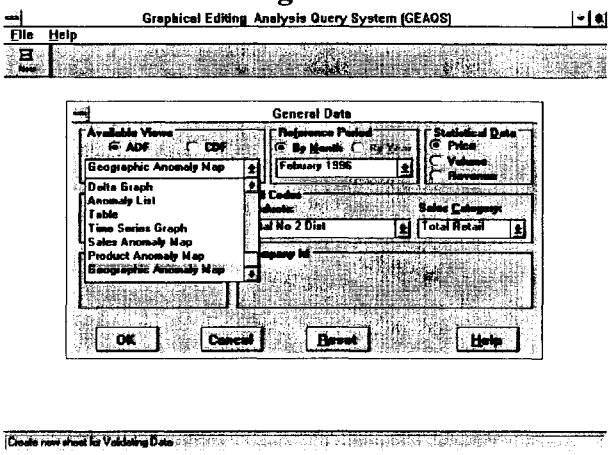
GEAQS also incorporates many of the visualization techniques described by William Cleveland. The top down approach is an iterative process. Edit failures are not just prioritized or ranked by some predetermined variable and printed. The analyst discovers which aggregates deviate the most, which next level aggregates directly contribute, and then which respondents are outliers and which have a high impact on that aggregate. Only two colors, limited to four shades each, are used in the anomaly maps, while the scatter graphs contain only three colors. Colors are used to distinguish different levels of severity. Even though legends are provided, the limited number of colors allows for "effortless perception." That is, it lessens the need to use the legends which is a cognitive process. Limiting the number of shades allows for clear distinction between shades within a color. In addition, visualization in scatter graphs of data also requires fitting the data which may not be immediately apparent. GEAQS displays a least squares regression line in addition to the no change or current-equals-prior line for orientation. The box and whisker plot and scatter graph automatically bring up the data table/spreadsheet into the right half of the window. Clicking on individual points on the graphs highlights the data in the spreadsheet and vice versa. Analysts can choose to focus on certain parts of the graph by drawing a box around the points of interest and then selecting either the inside box or outside box icon. The graph is then redrawn showing only the chosen set of data points with less clustering. Similarly, the data table will reflect only those points.

The pilot survey used in the development of GEAQS was chosen because of its complexity. The survey chosen collects state level prices and volumes of petroleum products sold monthly from a census of refiners and a sample of resellers and retailers. Volume weighted average prices are published at the state, Petroleum Administration for Defense District (PADD), and U.S. level for a variety of sales types and product aggregation levels. Volume totals and volume weighted average prices for refiners are also published. Approximately 60,000 preliminary and final aggregates are published each month.

THE APPLICATION

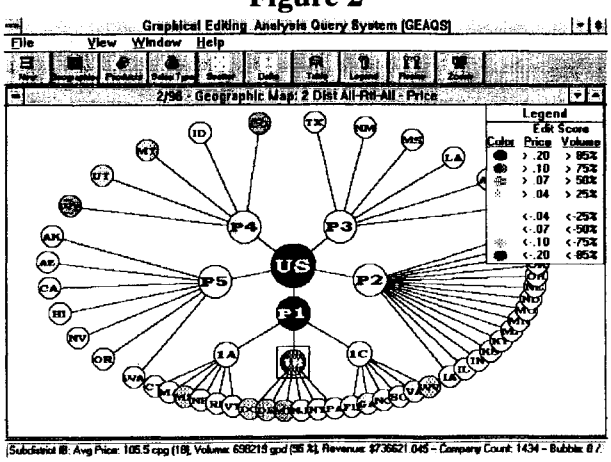
The user of GEAQS is provided the flexibility to decide where in the system to start. After clicking on the "new"

Figure 1



icon, the opening dialogue box (figure 1) allows the user to choose from various views. Four of these views are associated with aggregates, three anomaly views and a delta graph (box plot on change). The anomaly views are available for geographical, product, and sales type, the three main dimensions of the pilot survey, in addition to time. The geographical view requires the user to also select a product, sales category, seller type, statistical data type, and reference period from the drop-down lists provided by clicking within the respective boxes. As the user makes the view selection, the system adjusts the possible product selection, according to the combinations of aggregates calculated by the survey's processing system. Similarly, as the user selects the product, the list of possible sales categories is adjusted accordingly. Once all selections have been made, the user clicks the OK button and the graphic is displayed (figure 2). The geographical anomaly view

Figure 2



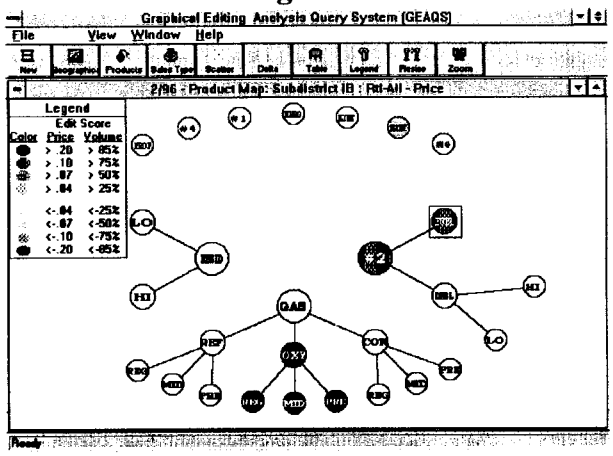
graphically represents aggregate cells of the selected data by placing a node for the highest level aggregate, the U.S., in the center of the map. Orbiting out from the center are nodes for the next level of aggregates, five regions of the country called PADDs. PADD I, the East Coast, is broken out into three more nodes for subPADDs. From each PADD or subPADD node, state level nodes are used to represent the lowest level of geographic aggregate. Each node, regardless of the level, is colored according to its current edit score. For price data, the current score is the difference between the price change (current price minus the previous period price) at the state level and the price change at the PADD/subPADD level calculated without including that particular state; the edit score for state k, at time period t is:

$(P_{k,t} - P_{k,t-1}) - (P^*_{\bullet,t} - P^*_{\bullet,t-1})$, where $P^*_{\bullet,t}$ is the PADD average price excluding state k.

Volume and revenue current scores are similar, but use the difference in percent change between the state and the PADD/subPADD. The current scores for the U.S. and PADDs are just the price change between the previous and current period. Four shades of blue are used to represent scores that indicate the price change is greater for that area (state or subPADD) than the more aggregated geographical area (PADD or U.S.) by 4, 7, 10, or 40 cents with increasing darkness of the color as the score increases. Similarly, four shades of green are used to represent area price changes that are less than the more aggregated geographical area by 4, 7, 10, or 40 cents. Areas where data do not exist are shaded grey. The analyst may click on the legend icon to clarify the color distinctions. The legend may be moved around the window or turned off as the user desires. If the user had chosen volume or revenue, rather than price for the statistical data selection, the shades of blue and green would represent different levels of percent change.

A user may click on any node of the map, which places a square around the node, to activate a geographical area colored to indicate a large price increase or decrease relative to the PADD. The tool bar at the bottom of the window shows the name of the geographic node activated, along with the weighted average price and the score for the area. The user may drill down by either clicking on the products or sales type icon. If the user had previously selected a product that can further be broken down to the reported product level, the user would choose the product icon. The window would be replaced by a new graphic, a product anomaly map (figure 3), that shows for the activated geographical area node all products broken down to the reporting level component products. The nodes are shaded the same way as the geographical anomaly map to indicate the levels of the edit score. The user can click on the appropriate component product to activate the reporting level product and then click the sales type icon to further drill down. The screen is then replaced with the sales type

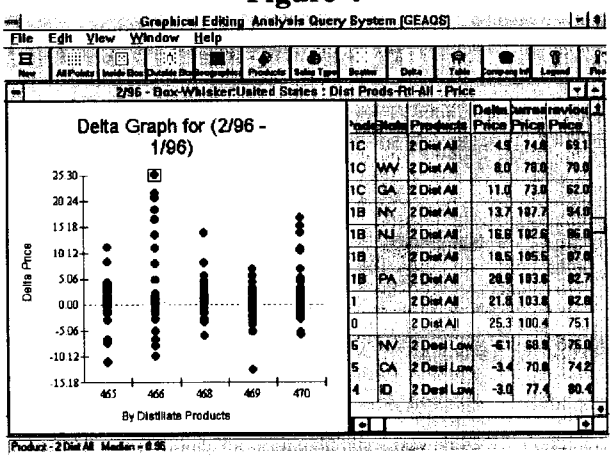
Figure 3



anomaly map which shows retail and wholesale sales type components for the activated state and product. Colored nodes are again used to signify the levels of relative change for the various sales types.

GEAQS allows the user to determine the path for drilling down. A user can start with a product or sales type anomaly rather than a geographical anomaly. The procedure is the same, only the order changes. An alternative procedure for drilling down is provided through the delta graph, a box and whisker graph of change -- price, volume or revenue -- between reference periods. In the opening dialogue box, the user selects delta graph under the

Figure 4



available views. The user would next select a group of related products through a product selection preceded by "all," a high-level sales category, total retail or total wholesale, and all sellers for seller type. Once all selections have been made, the user clicks the OK button. On the left side of the window, the box-whisker graphic (figure 4) displays a box plot for each individual product in that

product group, allowing the user to compare the spreads of the changes across those products. The vertical axis represents the change (price, volume or revenue), positive and negative, between the current and previous reference periods. Each box plot is labeled at the bottom by the product code associated with it. The "waist" of the box signifies the median for that product across geographical areas, including the aggregate areas of subPADD, PADD, and U.S. Individual circles plot the change for the geographical areas within the box (the middle 50% of the values for the geographic changes), within the whiskers, and outside the whiskers for each reported level and aggregate level product. The values within the whiskers are those values less than or equal to (greater than or equal to) the upper quartile plus (lower quartile minus) 1.5 times the distance between the upper and lower quartiles. Values beyond the whiskers, outside values, may not exist if the largest (smallest) valued geographic area within the whisker is the maximum (minimum) of the changes of the geographical areas. If outliers exist, they would be outside values. The additional information gained from the multiple box plots is the summary of the distribution of change, within and between products. If the distance between the top of the box, the upper quartile, and the median is very different from the distance between the bottom of the box, the lower quartile, then the distribution of change is skewed. The right side of the window contains a spreadsheet of the information for each circle on the plot. The analyst can click on any circle on the plot and the associate row of information for that value will be highlighted in the spreadsheet. The change for that aggregate cell, the current period's actual value, the previous period's actual value, as well as the label of the cell's state and/or PADD/subPADD and other relevant data are provided in the highlighted row. Utilizing windows' functionality, the analyst can scroll across, up or down the spreadsheet by clicking on the appropriate arrow buttons at the bottom and top right side of the spreadsheet. Columns in the spreadsheet can be rearranged by the usual click and drag on the column title. Column size can be changed by clicking and dragging the line that separates the columns. Leading columns can be held fixed while scrolling across the rest of the spreadsheet. An icon is also provided for the box and whisker plot. A single product box plot can also be selected. Regardless of the path chosen, through anomaly maps or box plots, the analyst at this point has determined the lowest level aggregate(s) (the reported level product, sales type, and geographic area) that contributed the most to the higher level aggregate anomaly. The analyst can further drill down to the individual respondent level by clicking on the scatter icon. For the activated geographic area, product, and sales type, a scatter graph of the data is displayed in the left half of the window (figure 5). The y-axis is the coordinate for the current period and the x-axis is the coordinate for the previous period. Each respondent's price (volume or revenue) is plotted using a circle and each nonrespondent's imputed value is plotted using a triangle. Data values whose contribution to the aggregate are 50% or more are depicted in red, values that represent more than 5% , but less than 50%, are yellow, and the remaining values, less than 5% share, are blue. A dashed line is provided that indicates no change between the previous and current period; the current period's value equals the previous period's value. Data falling above this line indicate increases in the current period, while data below represent decreases in the current period. In addition, a least squares unweighted regression line is also provided, represented by a solid line. These lines allow the analyst to visually perceive the overall trend of the data. Data values with a current but no prior value lie on the y axis with a (0,0) origin. Similarly, data values with a prior but no current value lie on the x axis with a (0,0) origin.

The analyst can draw a box around points of interest by clicking and dragging the mouse to draw a box around the respective points. The user then clicks on the "inside box"

Figure 5

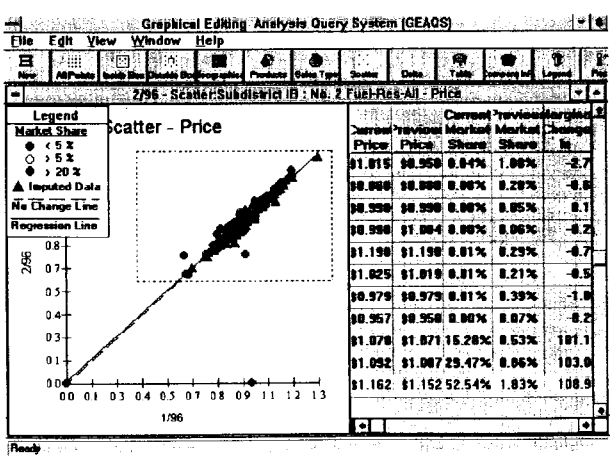
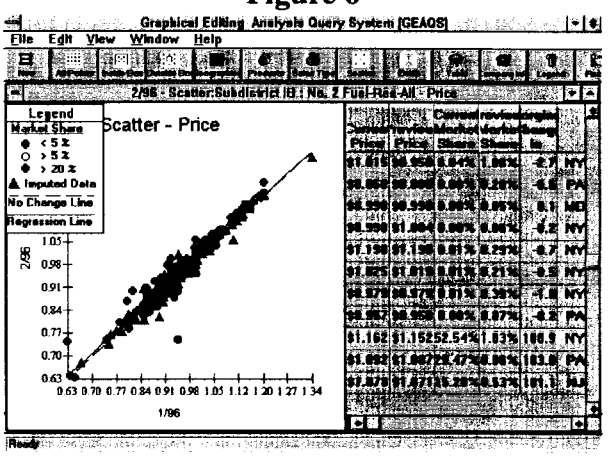
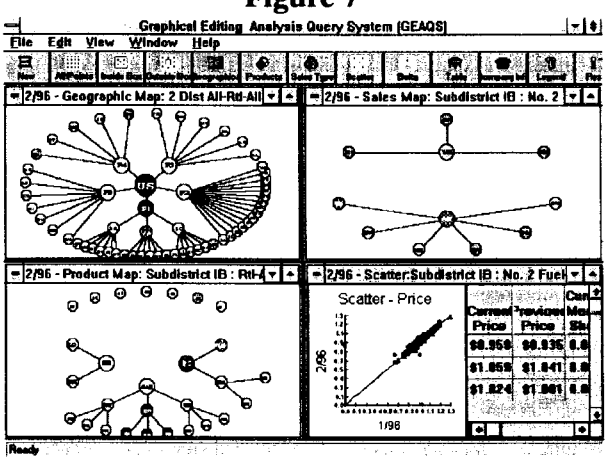


Figure 6



or "outside box" icon to have the graph redrawn according to the selection, using only those points in the box or those points outside the box (figure 6). These icons allow the analyst to uncluster points and focus on particular values. The original graph can be obtained by clicking on the "all points" icon. The ability to regraph only select data points is also provided in the box and whisker plot. The spreadsheet is simultaneously adjusted to display only the selected points. The right side of the window shows the information relating to each point on the graph as a spreadsheet . Each row of this spreadsheet represents a respondent. The spreadsheet contains the values or coordinates of each point, respondent identifier information, sample weights and volume weights, and other relevant information. The analyst can click on a row in the spreadsheet, highlight it, and a box appears around the corresponding point on the scatter graph. Alternatively, clicking on a point in the scatter graph, which boxes the value, results in highlighting the corresponding row in the spreadsheet associated with that value. Further information for contacting the respondent or determining if the respondent has been contacted can be obtained by clicking on the "company" icon. The analyst can scroll up, down, or across the spreadsheet and rearrange columns as previously described for the spreadsheet associated with the box and whisker plot. The combination of the scatter graph and the spreadsheet provides the user with the tools needed to identify the specific respondent(s) causing the aggregate cell to be an anomaly. Two other features, tiling and resizing, provide the analyst with a map of the remaining related

Figure 7



anomalies/outside values (figure 7). The analyst can close windows as outliers are resolved and return to higher level windows to determine the next aggregate to investigate.

FUTURE ENHANCEMENTS

Additional enhancements are still being made to GEAQS. Work is ongoing to incorporate a more sophisticated measure of each respondent's contribution to the aggregate change for price. This measure takes into account two time periods and the two response variables, price and volume, that affect the change in the volume weighted average price. This respondent contribution to the change in price will be an improvement over a simple market share measure for influence which only indicates potential for contribution to the aggregate change, based on volume for one time period.

The other major enhancement to GEAQS being examined is called "bubble up." This functionality provides the user anomaly information at the higher levels of aggregates concerning the associated lower levels of aggregation. It graphically signals the user that even though the current aggregate is not anomalous, a lower aggregate which is a component of that aggregate is anomalous. This early identification would remove from the user the burden of having to bring to the screen lower level published aggregates to determine if there are outliers at that level.

SUMMARY

The Graphical Editing Analysis Query System (GEAQS) was built upon the concepts developed in four other systems. A top down approach to data editing and validating, macro-editing, enables the analyst to efficiently focus on outlier respondents that impact the published aggregates. GEAQS provides anomaly maps and box and whisker plots to identify aggregate level outliers. The anomaly maps summarize the relationships of various levels of aggregates' change, and highlights outliers through color as determined by the current edit score. In comparison, the Box-Whiskers plot summarizes the distribution of change across geographical aggregates, allowing comparison of distributions within product groups, and highlights outliers as the outside values, outside the whiskers. Either path that is chosen directs the analyst to drill down to the lowest level aggregate. The scatter graph of the lowest level aggregate depicts the respondent and nonrespondent level data that contribute to the aggregate. Outliers are identified by their position relative to the other respondents' values and the fit line, while color is used to emphasize individual respondents' degree of impact on the aggregate estimate. The split window with the spreadsheet mapping to the scatter graph provides immediate identification of the values.

REFERENCES

Bienias, J., Lassman, D., Scheleur, S. and Hogan, H., (1995), "Improving Outlier Detection in Two Establishment Surveys," ECE Work Session on Statistical Data Editing, Athens 6-9, November 1995, WP No. 15.

Cleveland, William S., (1993), Visualizing Data, Hobbart

Press, Summit, New Jersey.

Engstrom, P. and Angsved, C., (1995), "A Description of a Graphical Macro Editing Application," ECE Work Session on Statistical Data Editing, Athens 6-9, November 1995, WP No. 14.

Esposito, Lin and Tidemann (1993), "The ARIES Review System in the BLS Current Employment Statistics Program," ICES Proceedings of the International Conference on Establishment Surveys, June 27-30, 1993, Buffalo, New York.

Mowry, S., and Estes, A. (1995), "Graphical Interface Tools in Data Editing/Analysis," (1995), Washington Statistical Society Seminar presentation, March 10, 1995.

Subcommittee on Data Editing in Federal Statistical Agencies, Federal Committee on Statistical Methodology (1990), Data Editing in Federal Statistical Agencies, Statistical Policy Working Paper 18, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.