Chapter 6

Graphical/Interactive Systems

Chair: Cynthia Z. F. Clark, National Agricultural Statistics Service

Bill Goodman ◆ Laura Freeman ◆ Mike Murphy ◆ Richard Esposito

Paula Weir

Mary Kelly

Chapter 6

Experiences on Changing to PC-based Visual Editing in the Current Employment Statistics Program

Bill Goodman, Laura Freeman, Mike Murphy, and Richard Esposito, Bureau of Labor Statistics

Abstract

A demo and talk were presented on the ARIES system, as used in the Current Employment Statistics Program at the Bureau of Labor Statistics. The system uses a top-down graphical and query search technique to identify outliers in estimate-level data, and then isolate and treat outliers in the corresponding sample data. This talk discusses the experiences of both developers and users in adopting ARIES, including the problems encountered and suggestions for future development.

Experiences on Changing to PC-based Visual Editing in the Current Employment Statistics Program

Bill Goodman, Laura Freeman, Mike Murphy, and Richard Esposito, Bureau of Labor Statistics

Background

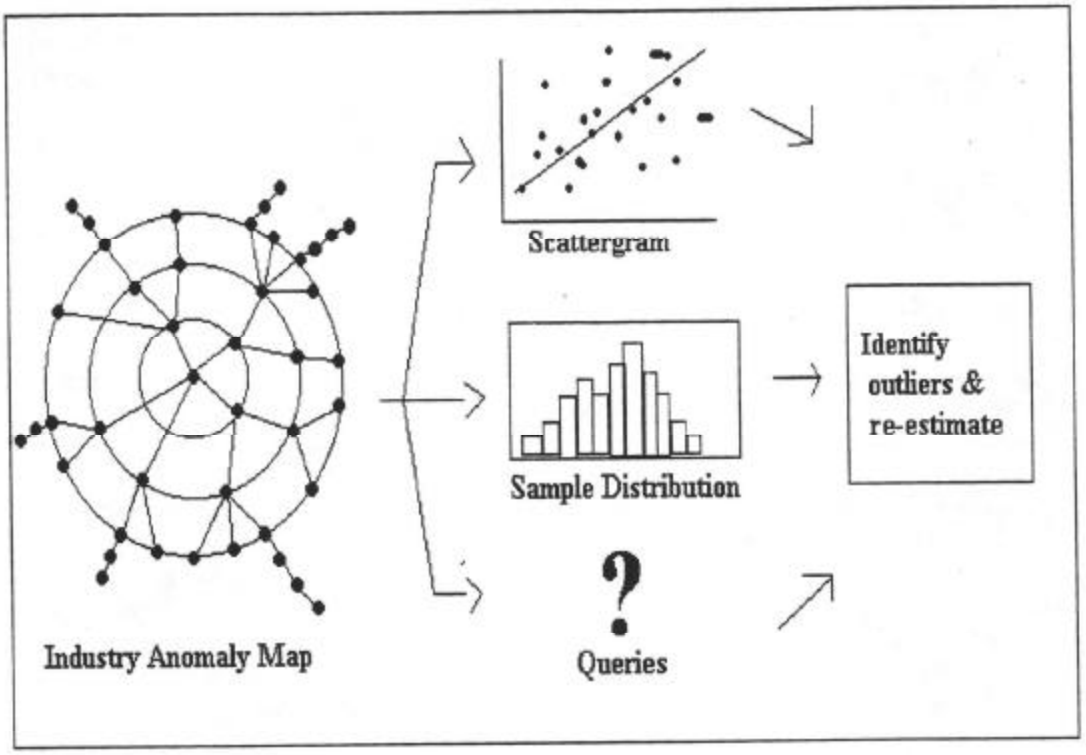
To convey the advantages of the ARIES system, some background on the survey in which it is used and on the preceding system will be necessary. The Current Employment Statistics survey is the largest monthly survey of employers in existence and one of the first major economic indicators each month. It samples nearly 400,000 establishments employing nearly 40 percent of all payroll workers, with over 45 million workers in the sample. It produces data for over 600 industries, including major divisions and more detailed industry levels. Our data were primarily collected by mail until we recently increased the use of electronic media. Now, phone collection, touch-tone self response, computer-assisted interviews, fax technology, and voice recognition are being used to obtain higher and faster response rates. The Current Employment Statistics survey of employment, hours, and earnings entails estimates of six to seven fields in 1,700 strata, for a total of approximately 10,800 estimates, monthly. Because of demands for quick responses, three sets of estimates -- two preliminary sets based on partial samples available earlier and the third set using the entire set of usable responses -- are produced each month.

The previous system, designed in the early to mid-seventies and written in COBOL, relied heavily on the use of listings, which were reviewed by nine industry analysts. Once questionable estimates were identified, it was generally necessary to review all individual reports for that stratum, as listings were sorted only by industry, state, and reporter identification number, so that finding all outliers entailed a complete review of the reports for the stratum. While certain preliminary screening of responses was performed, the screening did not catch all errors.

The ARIES System

ARIES is a graphical and query-based PC review system, designed to more easily identify and treat both estimate outliers at the macro level and sample outliers at the micro level. Review using ARIES consists of a winnowing through successively more micro-level data to isolate the questionable individual sample reports among the approximately 400,000 reports which may be responsible for unusual movements in estimates for the current month. For each data analyst, review starts with an anomaly map, which is a graphical tree-structure of the industries and estimates for which that analyst is responsible. Colors on the nodes of the map indicate the location of questionable industry estimates. Once significant problems are identified at this macro level, review continues through individual industry

scattergrams of reports representing sampled establishments, distribution graphs, and interactive queries, to isolate the individual sample members most likely to have contributed to the questionable estimate. Decisions on data weighting or rejection are made at the individual report level, and estimates are automatically re-generated on the PC and displayed on the PC screen. (A more complete description of this process can be found in the *Journal of Computational and Graphical Statistics*, Vol. 3, Number 2, June 1994.) The process is represented in the following schematic:



Loosely termed the "cell garden," the anomaly map is used to perform data validation at the macro level during the estimate review process. An excellent tool at each "closing," the anomaly map is especially helpful during the initial preparation of a given month's first preliminary estimates, or "first closing."

During the first closing, analysts must check their estimates at the macro level, decide which cells should be reviewed, and then search for reports that may be causing the abnormal estimates. Abnormal estimates may be due to typos in reported data and are often corrected. But if the estimate appears to be correct the analyst should become familiar enough with the situation such that it can be readily explained to interested parties. Finally, necessary changes are implemented, and the data are uploaded for merging into the production database. At first closing, the analyst has about three hours to perform these tasks and any others related to preparing and understanding the estimates for that month.

The anomaly map does two great things for the time-sensitive production analyst. The first is to identify cells which are out of range compared to the historical data for that particular month. The second is to make possible rapid review of all cells, regardless of whether ARIES found the cell's estimate abnormal. The ARIES anomaly map provides the historical over-the-month change data for each estimate. Estimates, or cells, can be selected and reviewed based on this history. The cell garden serves to facilitate the analyst's review of each cell in a rapid manner, and without the use of the somewhat cumbersome paper over-the-month change book.

Using the scattergrams in ARIES, one can select perhaps ten outlying individual reports for review, instead of reviewing all reports in the stratum, perhaps several hundred. The selected reports can immediately be reviewed in the ARIES system. An additional advantage of ARIES is its capacity to present, immediately on user demand, the latest sixteen months' data for a selected reporter. The old listings showed only the three latest months' data and the same three months a year earlier. With more data for the reporter, one can make a more informed decision about the validity of its latest data.

The Advantages of ARIES

One of the advantages of using the ARIES system is that members of our staff can take data editing a step further. For example, the Query function allows the analyst to investigate estimation problems more thoroughly by asking specific questions pertaining to the sample. For example, you could ask to see all the reporters that increased their employment by a certain percentage. To do so was impossible using the old paper method without going over thousands of pages of data, risking mistakes made on a calculator and using up valuable time. Another advantage to the query function is that it adds a new realm of research as to why estimates came out the way they did. It gives solid facts, for example, on reporting trends by state and comment codes. We are now able to tell in a matter of seconds how many of our reports came in coded, for example, with an effect of bad weather. So ARIES has not only improved our efficiency, but has also improved our analysis of the data. Our analysis of the data with the assistance of the ARIES system has become more accurate, more efficient, and more thorough. ■

Chapter 6

Graphical Editing Analysis Query System (GEAQS)

Paula Weir, U. S. Energy Information Administration

Abstract

The Graphical Editing Analysis Query System (GEAQS) is a software tool developed by the Energy Information Administration to graphically edit respondent level data using a top down approach with drill down capability. GEAQS combines exploratory data analysis and visualization techniques with the directional power of the anomaly map concept of ARIES, within an object oriented Windows application using PowerBuilder.

Graphical Editing Analysis Query System (GEAQS)

Paula Weir, U. S. Energy Information Administration

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 findings that median editing cost as a percentage of total survey costs was 40 percent 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 for the Current Employment Survey is also described. The Automated Review of Industry Employment Statistics (ARIES) system helps to identify true errors quicker and results in fewer man-hours to edit the data. Graphics, particularly screen graphics, were found to be a preferable approach by the data processors and greatly reduced the amount of paper generated during the survey cycle. The recommendations of the subcommittee included the need for survey managers to evaluate the cost efficiency and timeliness of their own editing practices and the implications of important technological developments such as microcomputers, local area networks, and various communication links, as well as the expertise of subject matter specialists.

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. Examination of each case individually allowed the analysts to neither see the bigger industry picture nor see the impact of the individual data point on the aggregate estimate. The analysts, therefore, examined more cases than necessary. Thirdly, edit parameters or tolerances were derived from previous surveys which implied the relationships were constant 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 group considered box plots, scatter plots and some fitting methods, as well as transformations. This graphic approach could also be combined with batch-type edits while simultaneously evaluating dynamically set parameters or cutoffs. The working group concluded that a successful system requires that the system be acceptable to the people who use it. This requires training and incorporating the tools into the production environment and system. Two other systems, the Graphical Macro-Editing Application at Statistics Sweden, and the Distributed EDDS Editing Project (DEEP) of the Federal Reserve Board, have further demonstrated the efficiency of graphical editing.

The Concept

The Graphical Editing Analysis Query System (GEAQS) is being developed by 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-Whiskers graphs 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 level data for the current period versus the last period for the select aggregate. Actual reported data is distinguished from imputed data by the use of circles and triangles. This allows the user to pursue different follow-up procedures accordingly. The additional benefit of different symbols for respondents and imputed data is the visualization of the distribution of imputed data with respect to reported data and confirmation of whether respondents are similar to nonrespondents. Data points with high influence are indicated by color. High influence points that 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 contribute to over editing. Batch edit flags can be passed to the system to further prioritize the failures, as well as evaluate and help determine parameters or cutoffs. 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-Whiskers, as well as a data point on the scatter graph. The user can take advantage of tool bars, dialogue boxes, and icons. Resizing and zooming are 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. An icon for a legend is also provided to assist the analyst in distinguishing colors, shapes, etc. In order to maximize the usefulness of GEAQS to other surveys, additional time and effort was taken to make GEAQS object oriented. This allows for minimal 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 the data processing system.

GEAQS will also build on the work done for the DEEP system of the Federal Reserve Board by capitalizing on time series information. It allows the analyst to view the respondent data over an extended period of time. What may appear as an anomaly with respect to other respondents in that cell may be consistent with that respondent's historical reporting. This capability supplemented with pull down text comments helps the analyst determine if the respondent's reporting difference has been verified previously. Like the Federal Reserve System, GEAQS was developed in PowerBuilder and uses Pinnacle graphics server to help generate the graphs. The use of PowerBuilder and Pinnacle resulted in quicker development time and less cost. In addition, in order to capture the recommendation of the Census Working Group that the system is 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. This made the product more useful to the analysts by allowing their direct input throughout the process.

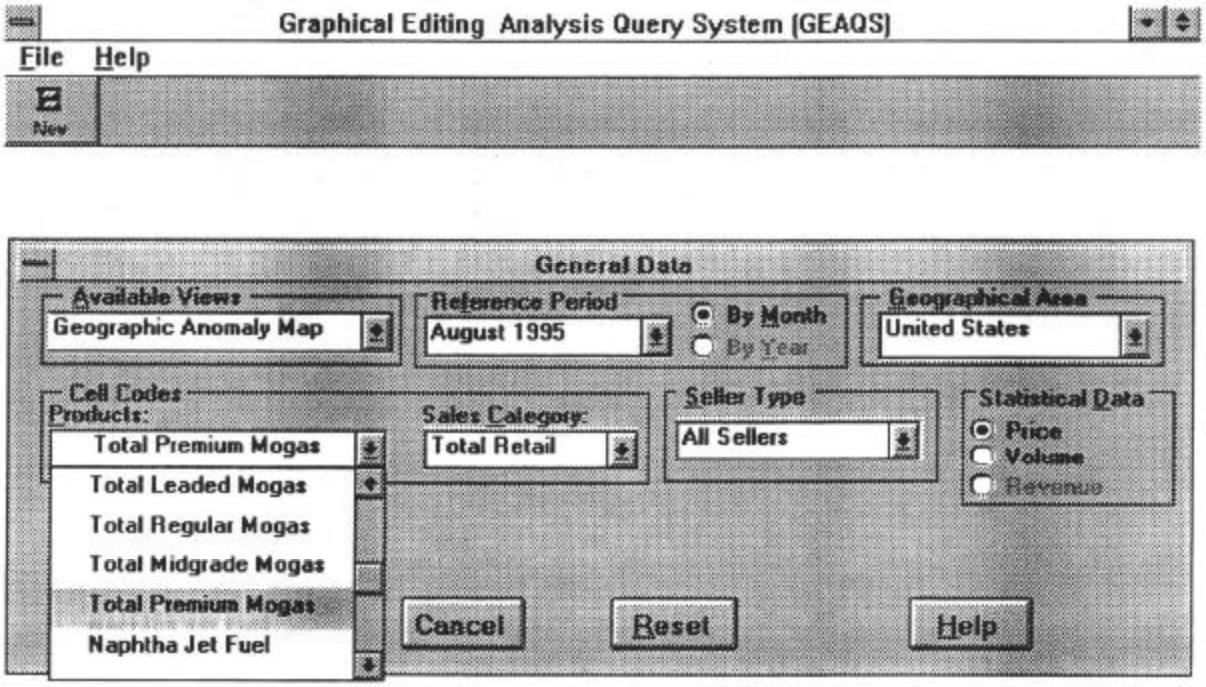
GEAQS also incorporates many of the visualization techniques described by William Cleveland. The top-down approach is an iterative process. Edit failures are not just listed prioritized or ranked by some predetermined variable. The analyst discovers which aggregates deviate the most, which next level aggregates directly contribute, and then which respondents are outliers and which have a high impact on that aggregate. Circles are used for data points to minimize darkening with the exception that triangles are used for imputed data. 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 would be 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. The fit 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. Transformations, particularly power transformations, of the data may also be necessary to uncluster the data, reduce the spread of the data, or reveal an underlying linear relationship. Logarithms make the data more symmetric and reduce skewness, monotone spread and multiplicative effects which make it difficult to visually determine the true outliers. To further assist in unclustering and identifying individual responses, zoom and resize capabilities are provided by a mere click on the respective icon. Tiling of the windows is also possible, allowing the analyst to keep the bigger picture in mind or a road map of where the analyst is in the process. The scatter graph automatically brings up the data table/spreadsheet into the right half of the window. Clicking on individual data points 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. Similarly, the data table will reflect only those points.

The pilot survey used in the development of GEAQS was chosen because of its complexity. It was felt that if graphical editing could be successfully accomplished for this survey, it would be a small task to modify the system for other surveys. 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" 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-Whisker 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

Figure 1

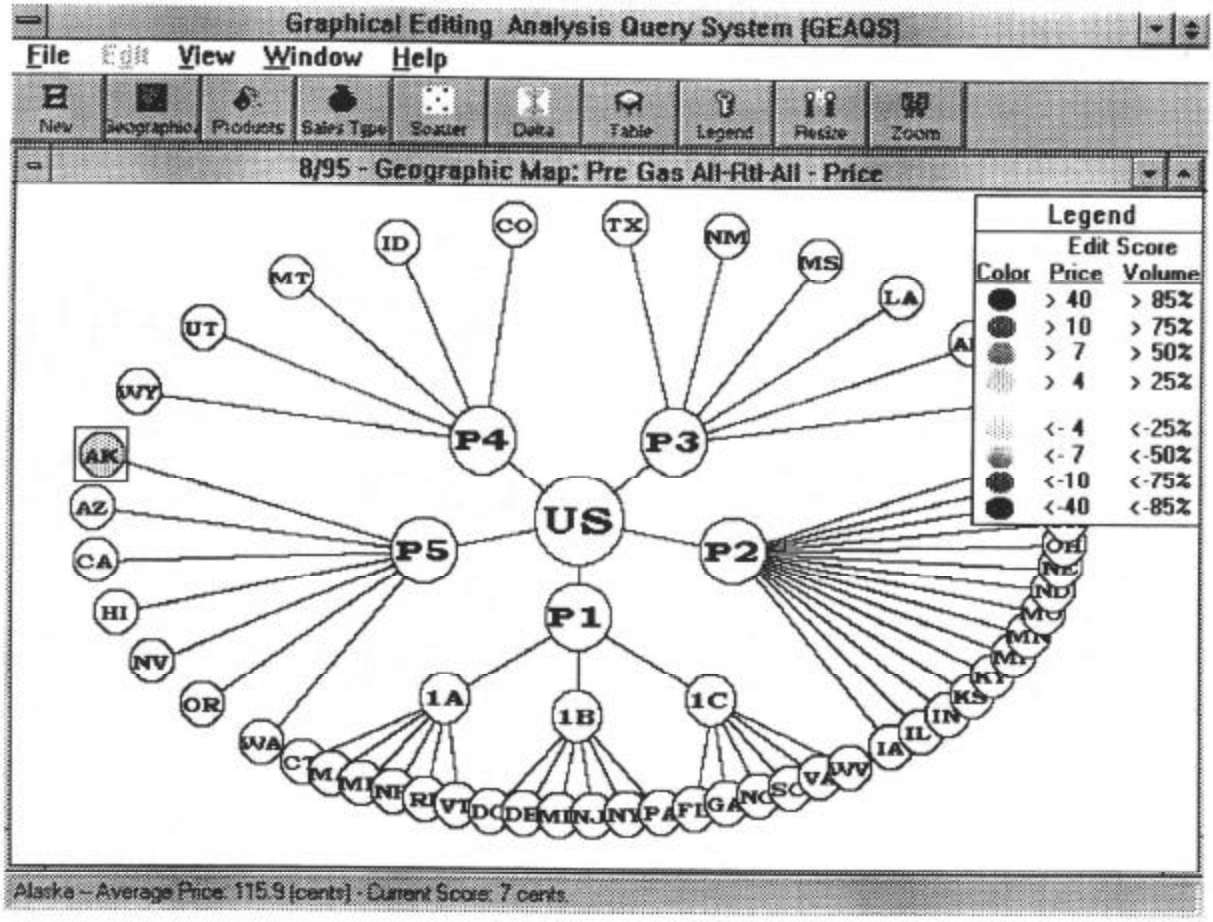


possible sales categories is adjusted accordingly. Once all selections have been made, the user clicks the OK button. The graphic is then displayed (Figure 2). The geographical anomaly view graphically represents aggregate cells of the selected data by placing a node for the highest level aggregate, 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. One PADD 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 or subPADD level calculated without including that particular state; the edit score for state k, at time period t is:

$(P_{k,t} - P_{k,t-1}) - (P^*_{.,t} - P^*_{.,t-1})$, where $P^*_{.,t}$ 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 or 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

Figure 2



(PADD or U.S.) by 4, 7, 10, or 40 cents as the darkness of the color 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 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 will show the name of the state, subPADD, PADD, or U.S. 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's 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 anomaly map (Figure 4) 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.

Figure 3

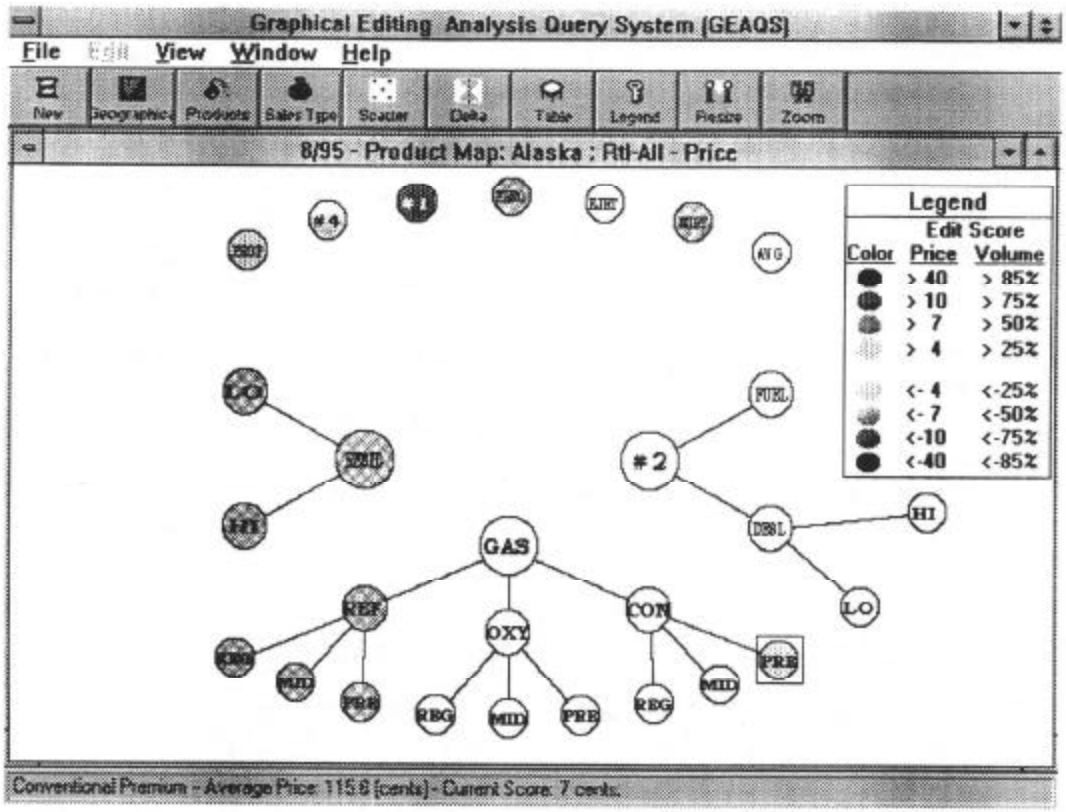
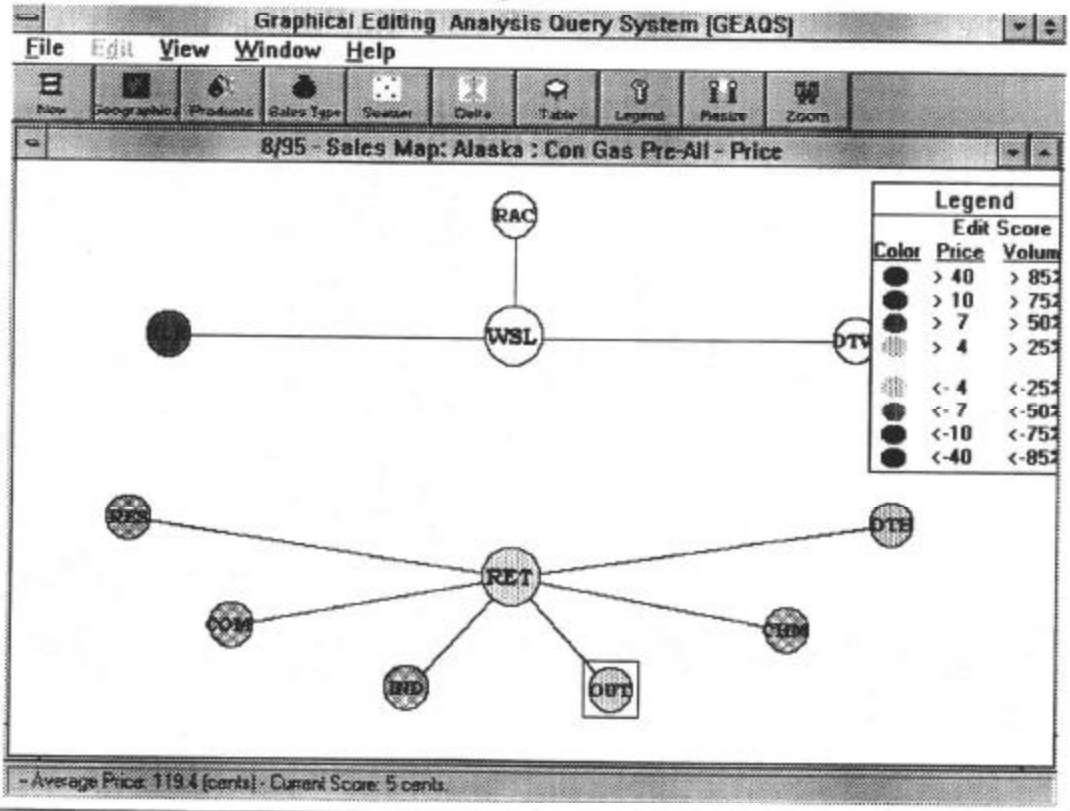
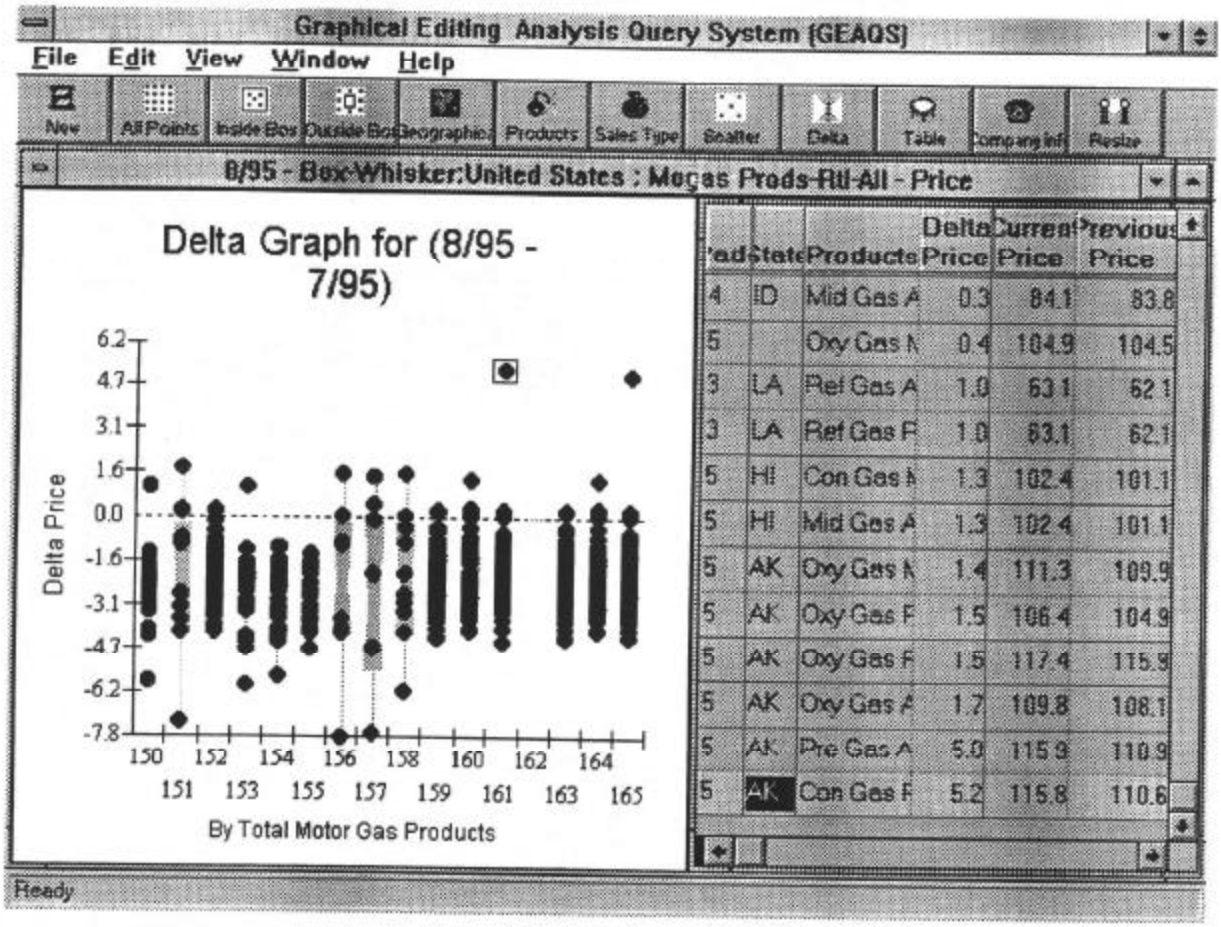


Figure 4



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-Whisker graph of change -- price, volume or revenue -- between reference periods. In the opening dialogue box, the user selects delta graph under the 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 5) 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 percent of the values for the geographic changes, within the whiskers, and outside the whiskers, which are called outside values.

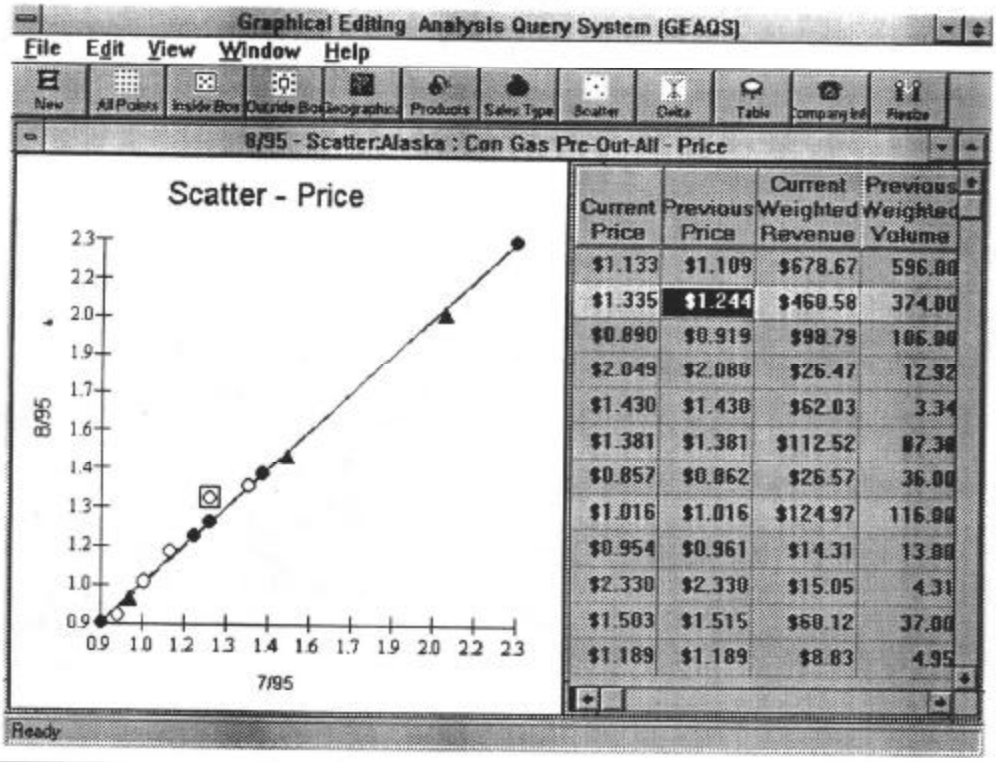
Figure 5



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 Box-Whisker is the summary of the distribution of change. 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 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 window's arrow buttons. Columns in the spreadsheet can be rearranged by the usual click, and drag method, clicking on the column title at the top of the column. 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 by clicking on the shaded area left of the arrow button at the bottom of the screen and dragging it to the end of the last column to be held fixed. An icon is also provided for the Box-Whiskers graph. After an analyst has identified a particular product and sales category through the anomaly maps, the analyst can click on the Box-Whiskers' icon to see a single box plot representing the distribution of change across geographical areas between the previous and current periods. Regardless of the path chosen, at this point the analyst has determined the lowest level aggregate(s) that contributed the most to the higher level aggregate anomaly.

The analyst can further drill down to the respondent level by clicking on the scatter icon. For the activated geographical, product, and sales type, a scatter graph of the data will be displayed in the left half of the window (Figure 6). The y-axis is the coordinate for the current period and the x-axis is the

Figure 6



coordinate for the previous period. Each respondent-level 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 percent or more are depicted by red, values that represent 5 percent or more, but less than 50 percent, are yellow, and the remaining values, less than 5 percent share, are blue. A dashed line is provided that indicates no change; 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 regression line is also provided, represented by a solid line. The analyst can draw a box around points of interest by clicking to the left and above the respective points, holding down the button, and dragging to the bottom right of the respective points and releasing the button (Figure 7). The user then clicks on the "inside box" 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 8). The "inside box" icon allows the analyst to uncluster points and focus on particular values. The "outside box" icon allows the user to examine the scatter without certain points. The original graph can be obtained by clicking on the "all points" icon. The right side of the window shows the information relating to each point on the graph. Each row of this spreadsheet represents a respondent. The spreadsheet contains the values 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 will appear around the corresponding point on the scatter graph. Alternatively, clicking on a point in the scatter graph, which boxes the value, will result in highlighting the corresponding row in the spreadsheet associated with that value. Further information for contacting the respondent 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-Whiskers plot. The combination of the scatter graph and the spreadsheet provide the user the tools needed to identify the specific respondent(s) causing the aggregate cell to be an anomaly.

GEAQS was designed to be interactive with the data base of the processing system. Once a particular respondent value has been identified, the analyst could change the response directly in the spreadsheet if so desired. The analyst would then be able to reexamine the newly computed aggregates to determine if it were still an anomaly. At this time, because GEAQS is not tied in with the processing system, and the pilot survey's estimation system is too complex to duplicate within GEAQS, the changing of respondents' data and recalculation of the aggregates cannot be demonstrated using the pilot's downloaded Watcom SQL database. It should be clear, however, that changes could be made, even temporarily, to determine the effect of the change.

Future Enhancements

Additional enhancements are still to be made in GEAQS. Work is ongoing to incorporate a more sophisticated measure of each respondent's contribution to the aggregate change. In particular for the pilot survey, because price is the ratio of revenue to volume, a respondent's contribution can be measured by the shift in the respondent's market share of revenue between months multiplied by the difference in price between that respondent and the respondent who inherits (or gives up) the majority of the market share in the corresponding month. This contribution to the change would be an improvement over a simple market share measure for influence which only indicates potential for contribution to the aggregate change. The other major enhancement to GEAQS is called "bubble up." This functionality provides the user anomaly information at the highest levels of aggregates concerning the associated

Figure 7

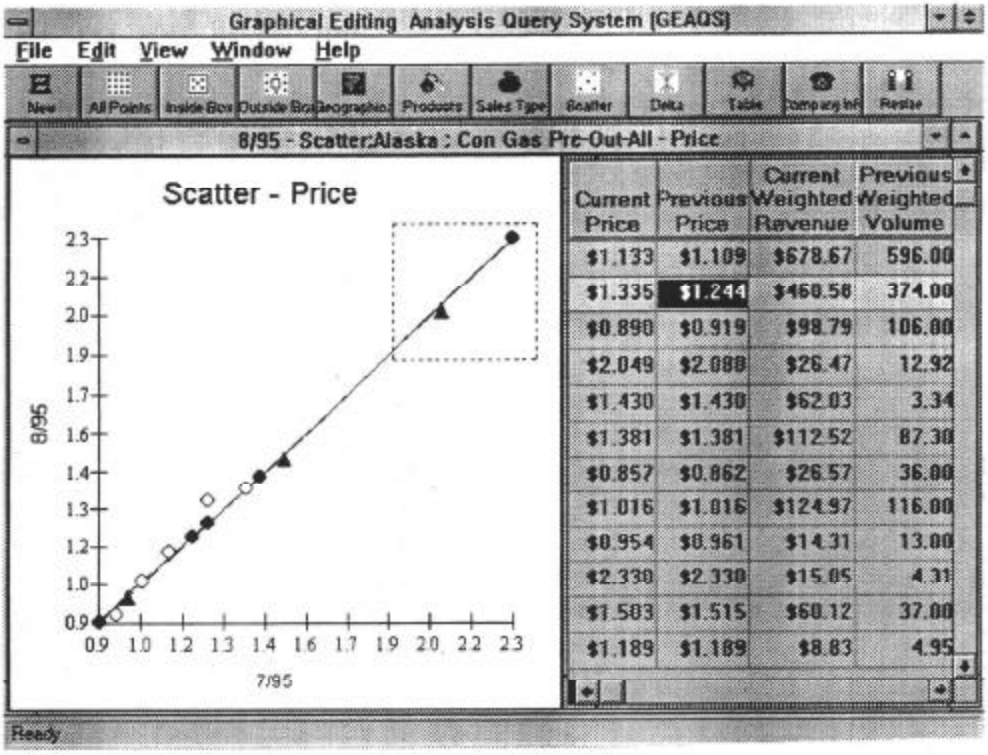
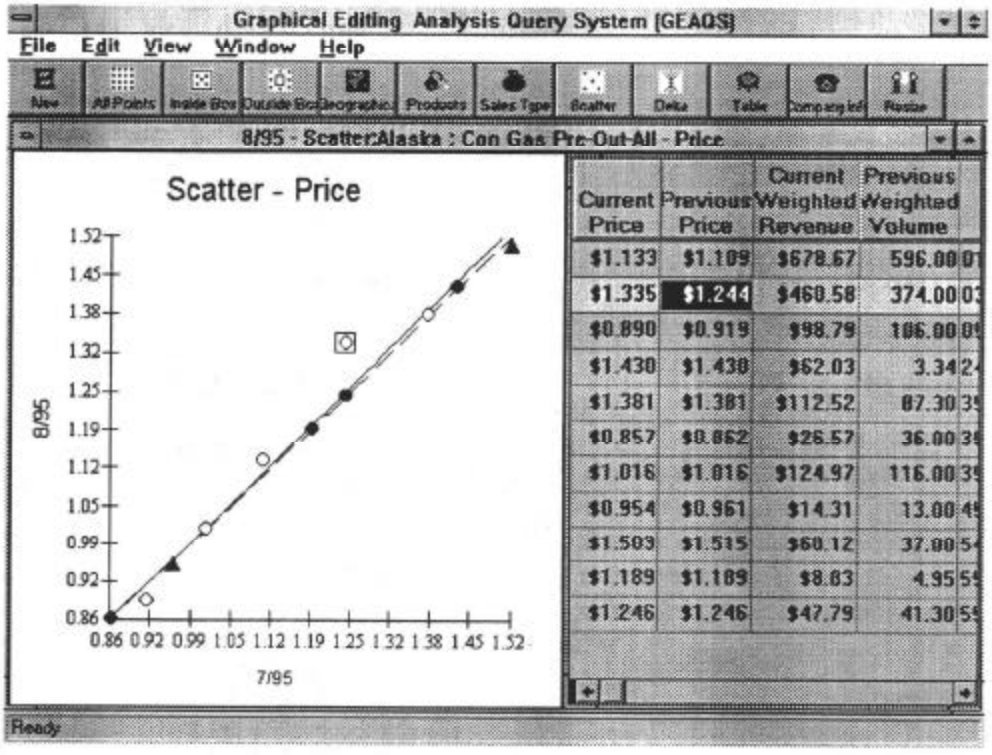


Figure 8



lower levels of aggregation. It graphically signals the user that even though the current aggregate is not anomalous, a component of that aggregate is anomalous. The user would immediately see where drilling down was necessary. This 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. It is expected that when GEAQS is incorporated into the processing system other variables in the data base will be available to it. Respondents who failed edits in the batch process can be flagged in the spreadsheet and scatter gram. Transformations such as logarithms and roots will also be possible. Recorded comments obtained by contacting respondents will be accessed by clicking on the "company" icon. Time series data for aggregates and respondents will also then be possible. Standard errors of aggregate estimates will also be incorporated.

Summary

The Graphical Editing Analysis Query System (GEAQS) 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 outliers that impact the published aggregates. GEAQS provides anomaly maps and Box-Whiskers plots to identify aggregate level outliers. The anomaly maps summarize the relationships of various levels of aggregates 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 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 respondents' influence 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, Working Paper 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, Working Paper No. 14.

Esposito, R.; Lin, D.; and Tidemann, K. (1993). The ARIES Review System in the BLS Current Employment Statistics Program, *ICES Proceedings of the International Conference on Establishment Surveys*, Buffalo, New York.

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

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. ■

Chapter 6

Integrated Data Capture: A System for All Office of Compensation and Working Conditions Surveys

Mary Kelly, U. S. Bureau of Labor Statistics

Abstract

The Office of Compensation and Working Conditions (OCWC) conducts surveys of wages and other compensation by occupation and industry. Until recently, the office collected these data via three surveys: the Employment Cost Index, the Employee Benefits Survey, and the Occupational Compensation Survey Programs.

Each survey was completely independent of the other, with its own method of sampling, set of field collections staff and survey administrators, and its own set of computer systems for data entry, editing, transmittal and publication.

Approximately one year ago, the Office began a major reconceptualization of its survey programs, effectively collapsing the three sets of surveys into one consolidated survey program, along with a single Integrated Data Capture (IDC) system.

This presentation and associated exhibit will highlight the Integrated Data Capture system. The presentation will be built around answers to the following questions:

- Why was the IDC built?
- How was IDC designed and tested?
- What does IDC do?
- What is unique about IDC?

We offer the following overview of IDC -- The system was designed to consolidate all aspects of data capture for the three sets of surveys within OCWC. The data capture portion of IDC is a Windows-based Power Builder

Abstract (Cont'd)

system with data being fed into a relational SYBASE database residing on a Unix server. Prior OCWC computer systems were primarily mainframe platforms built as large batch systems requiring significant resources to implement modifications. IDC is a modular database system, with each component (e.g., data capture and edits, database, transmittal) residing independently of the others. Such a system has obvious advantages with regards to updates to the software.

Prior OCWC computer systems were designed so that many of the edits occurred at the same time as data entry. Thus, the system would "stop" and query the user if an important data element was missing or "out of scope" of an edit.

Such a design rendered the system unusable within a "live" in-person collection environment. IDC permits the user to enable the edits at the end of data entry, thus making the system viable for use during in-person collection.

Finally, IDC is compatible with the latest programmatic priority -- moving away from centralized data review and towards schedule review and edits under the greater control of individual field staff.

Integrated Data Capture: A System for All Office of Compensation and Working Conditions Surveys

Mary Kelly, U.S. Bureau of Labor Statistics

Introduction

The Bureau of Labor Statistics' Office of Compensation and Working Conditions (OCWC) conducts surveys of wages and other compensation data by area, industry, and occupation. The data for the surveys are collected across the country by about 160 field staff located in 8 regional offices and 9 smaller outstations. Survey data are collected by the data collectors, reviewed at the regional offices by the review staff, and sent to the main office in Washington for final processing. Until recently, the office collected data via three surveys: the Employment Cost Index, the Employee Benefits Survey, and the Occupational Compensation Survey Program. Each survey was completely independent of the other with its own method of sampling, set of field collection staff and survey administrators, and its own set of computer systems for sampling, data entry, editing, transmittal, and publication.

Approximately one year ago, the Office began a major reconceptualization of its survey programs and is in the process of integrating the three surveys into one. With the development and testing of new survey procedures, the time seemed right to develop new systems for the new survey.

This paper is a report on the new system that was designed. The Integrated Data Capture System (IDC) is a Windows-based Power Builder application that accesses a SYBASE database located in Washington. The application can also access a similarly designed Watcom database loaded on the hard drive of a PC or laptop. The System development started in July 1995 and was used for the first time in a test survey in February 1996.

Structure of IDC Development Team

The Office-wide redesign was brought about by the exigency, due to downsizing, of combining the three surveys into one new survey. Collapsing the three surveys necessitated a review of all aspects of survey sampling, data collection methodology, review, and processing. Deciding which of these systems to revise and/or develop first presented the initial challenge. Given the publication criteria already in place within OCWC, it was determined that the sampling methodology would be similar to one of our current surveys, and thus, we could use one of the existing survey sampling systems with only slight modifications. In addition, it was felt that although we knew what new data we wanted to publish we were unsure just how much of the data collected from of our test surveys would meet the publication criteria. Given this uncertainty, we decided to take a wait and see attitude regarding making changes to the existing tabulation systems. It became quickly evident that the greatest change was within the data collection methodology -- we were combining three compensation surveys, each survey measuring a distinct aspect of compensation, and we wanted to retain, as much as possible, all the individual data

elements. Changes to data collection methodology implied that data entry and data editing system development had the highest priority. Furthermore, coordinating the redesign of these systems with the redesign of survey procedures would allow us to conduct a dual test of both the new survey procedures and the system at one time and one location.

To develop the IDC, we convened a small group of individuals, each of whom knew the content of the new survey and had worked on a variety of the Office's previous compensation systems and special projects. This small group served to coordinate IDC development. In addition, and perhaps more importantly, it was decided to split the systems staff who conducted the software development into specialties. In past efforts, the systems staff would work on the entire "System" starting at the beginning, and usually running short of time at the end. The systems staff patterned themselves after the System they were developing; they each became small "modules" developing discrete units of the System. From the beginning, people were placed within the larger framework and one person was identified to provide the interconnection between the small work units, and ensure that what was needed from each was available at the critical time. The person selected for this job was someone who was experienced on compensation systems projects and had been successful in coordination roles in the past.

IDC System Decisions

The first decision in designing the system was choosing the software in which to write the application. Traditionally, the Bureau of Labor Statistics used mainframe systems, even if the up-front data collection occurred on a PC. For the new system we wanted a system that would function solely in a PC/server environment. The data capture system is a windows-based Power Builder system with data being fed into a relational SYBASE database residing on a UNIX server. We chose Power Builder because Power Builder is a powerful and flexible graphical user interface (GUI) development tool. In addition, Power Builder has powerful database interface tools that work with both Watcom and SYBASE. This made it possible to develop the IDC without having to write separate data access routines for our similarly designed SYBASE and Watcom databases. SYBASE was chosen because it is a sophisticated relational database management system that meets our requirements for relational integrity, security, performance, and reliability. Finally, because we have staff collecting data all over the country, we needed a database tool that would work well on laptops. Watcom was chosen for this because it runs in DOS rather than UNIX and therefore is a better match for the PC/laptop environment.

IDC is a system that is used for data collection, data editing, and data transmission. It is not a sampling or data processing/publication system. These systems will be developed later. Four separate modules comprise the system:

- The data capture module which allows for the entry of establishment and occupation data.
- The transmission module lets the user move schedules to and from the main server.
- A utilities function is used to backup and restore data, and
- An assignment module is used to make data collector assignments. Each of these modules accesses either a local PC Watcom database or the central SYBASE database.

What Does IDC Do?

IDC is a windows based system, and thus, has recognizable features such as the ability to open multiple windows, maximize and minimize windows, and cut and paste portions of text. In addition, it is a visual package that makes use of many icons and pull-down menus that let the user navigate from screen to screen.

After the sample is selected, some basic information about each establishment is loaded into IDC. This includes a unique identification number for each unit in the survey, the name of the company, and any information from the sample that will help during data collection or data processing. The remaining data are entered by the data collector during or after the interview.

The data capture module of IDC is composed of 5 main sections: establishment information; an occupation selection calculator; occupation information; leveling information; and wage information. Each section is briefly described below.

- The establishment data section of the system is designed for the capture of establishment data such as the usability of the unit, Standard Industrial Classification (SIC), and employment. In addition, the administrative aspects of the unit -- such as date of collection, interview time, and method of collection -- are captured on this screen.

- Probability selection of occupations is a disaggregation technique for selecting the occupations within each establishment for which we will be collecting data. Because this is a random process, the data collector must perform a calculation to determine which occupations to select based on the establishment employment and the number of occupations that are required (statistically) for an establishment of its size. In the past, the data collector performed the calculation on paper using a hand-held calculator. In IDC, the system will perform the calculation. In most cases the only required entry is the total employment count for the establishment.

- The occupation section of the system captures the list of occupations that have been selected, along with certain occupational characteristics -- such as occupational employment, the census occupation code (a code that the Census Bureau uses to categorize the occupations of individuals as recorded by the decennial Census), whether the occupation is full-time or part-time, and whether it is union or non-union.

- One objective of the OCWC compensation survey is to determine the level of duties and responsibilities of each occupation. The office has developed a set of leveling criteria patterned after the Federal Point Factor Evaluation System. The technique, called Generic Leveling, is a way of easily leveling any job that is (randomly) selected in each establishment. To level a job, there are ten "factors" that must be measured for each job. Knowledge, complexity, supervisory controls, and physical demands are examples of factors. Each factor is further subdivided into a certain number of levels and corresponding number of points. The point total defines a particular generic level. The IDC generates a separate screen on which to record the level of every job. This screen has ten tabs that correspond to each of the ten factors. When a folder tab is chosen, a list of levels is presented. When the user makes a selection, the System generates the corresponding number of points. When all ten tabs are complete, the System totals the points to determine the corresponding Federal General Schedule (GS) level of the job and it displays the GS level on the screen.

- IDC is designed to accept wages in any form provided, including hourly, weekly, monthly, or annual wages. The System then changes the data entered into the form needed for publication -- typically hourly rates. To perform the transformation, the user enters hours and earnings. The System will calculate an hourly rate for each worker. In addition, the system will calculate an average hourly rate for each occupation. In many cases a worker receives a base rate of pay as well as some addition to the wage such as a commission. The system can also handle this type of wage calculation and not only calculates an hourly rate for the worker, but also produces a base hourly rate for each worker.

Unique Features of IDC

The final section of this paper highlights certain features of the Integrated Data Capture system that we in OCWC are particularly proud of. We feel these features make IDC a unique Federal government data capture system.

In the past all of the Office of Compensation and Working Conditions systems, even if PC based, were designed in such a way that data entry was difficult to key during an interview. The systems were designed to edit data at the same time as data entry was occurring necessitating that all fields be complete as the user progressed from screen to screen. Because the surveys often dealt with complex compensation data that the respondents did not typically have at their fingertips, the systems were difficult to use during actual interviews. IDC was built to overcome this problem. We wanted a system that could be used during the interview. While the Integrated Data Capture system is not an Expert System in that it does not lead nor help the data collector through the interview, it is designed for data capture during an interview. IDC permits the user to enable the System edits when convenient, rather than operate them automatically.

Because each occupation selected must be assigned a census occupation code, an electronic copy of the census codes and census titles is included in the system. By clicking on the field where the user enters the census occupation code, the system brings you to the census code list. Selection of a code simultaneously enters the code into the correct field and brings the user back into the census field. In the future the hope is to not only include census codes and titles on this list but to also include complete census definitions.

As a final example of unique aspects of IDC, in order to assist the user in leveling the chosen occupation, the system contains a prototype screen for generic leveling. These prototypes were developed using data from the typical factors for a given Federal General Schedule level. Once the user completes leveling all ten factors, they select a prototype tab. The system generates a graphical representation of what the user choices were versus what the prototype expected. Edits are generated when the user choices do not match the expected prototype.

Conclusion

While the Integrated Data Capture system was just a first step in the long process of developing all new and interconnected systems for OCWC, it went a long way in teaching us about successful systems development. From all indications the development process and design of the system were a success. OCWC was able to produce a system that functions well in terms of accuracy, usability, and speed for the data collection, editing, and transmission systems and were able to produce it on time. For this reason, we plan to continue to use the same development process, that is, splitting staff based on specialty, for our new sampling and processing systems as well as for the many additional data collection features still to be added to IDC. ■