

IMPLEMENTATION OF THE GRAPHICAL EDITING ANALYSIS QUERY SYSTEM

Ruey-Pyng Lu, Paula Weir, and Robert Emery
Energy Information Administration, Science Applications International Corporation

The Graphical Editing Analysis Query System (GEAQS) has been developed by the U.S. Energy Information Administration (EIA) as a tool to reduce survey costs and reduce the amount of paper generated. It builds on the concepts and features of four other systems - the ARIES system, the Bureau of Census working group prototype, the Graphical Macro-Editing application of Statistics Sweden, and the Distributed EDDS editing Project (DEEP) of the Federal Reserve Board. GEAQS constructs an anomaly map to summarize the relationship of various levels of data aggregates, flags questionable aggregates through the use of color, makes use of the tools of exploratory data analysis to identify outliers and suspicious responses, and capitalizes on time series information through graphs.

To significantly reduce data collection and processing costs, and reduce software lifecycle costs, EIA is currently developing a new Common Collection and Processing System (CCAPS) and Master Universe Database (MUD). It is highly desirable for an efficient data editing system to be adopted in CCAPS. As a result, GEAQS is being piloted on the Natural Gas Monthly survey and tested for its effectiveness and adaptability for use by other EIA surveys. The current system's edit rules and parameters were transferred into the GEAQS to evaluate the efficiency of the system and its ability to replicate the results of the current system. This paper describes the results of the comparison of the current production hard copy output process to the graphical approach using the same edit rules.

KEYWORDS:

validation, macro-editing, micro editing, anomaly map

1. Introduction

The use of graphical approaches were cited in the Statistical Policy Working Paper No. 18 "Data Editing in Federal Statistical Agencies" released by the Data Editing Subcommittee of the U.S. Federal Committee on Statistical Methodology. Graphics, particularly screen graphics, were found to be a preferable approach by the data analysts, and screen graphics greatly reduced the amount of paper generated during the survey cycle by the previous batch system. The recommendations of the Data Editing 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. It also recommended that more attention is given to the future roles of subject matter specialists and the tools needed to perform their jobs. The use of graphics promoted a rethinking about the possibilities for improving the data review capabilities of the analyst through simultaneous views of multiple levels of analysis. This led to the development of a new graphical editing system that took advantage of the concepts and features of several graphical systems. The Graphical Editing Analysis Query System (GEAQS) merged the ideas of: (1) the anomaly map of the BLS ARIES system which summarizes the relationship of various levels of aggregates, and flags questionable aggregates through the use of space and color; (2) the tools of exploratory data analysis, such as box-whisker plot combined with subject matter specialists' expertise to identify unusual cases, as described by the Census Working Group prototype; (3) the tool bars, dialogue boxes, and icons of the Graphical Macro-Editing Application at Statistics Sweden, which allow the GEAQS user to point-and-click on an aggregate in the anomaly map, the box-whiskers, or a data point on the scatter graph; (4) the Distributed EDDS Editing Project (DEEP) of the U. S. Federal Reserve Board, which displays multiple times series graphs of individual respondents with the ability to point at an observation and obtain text comments regarding that observation. Like the Federal Reserve System, GEAQS was developed in PowerBuilder, but also uses Pinnacle graphics server to help generate graphs. The use of PowerBuilder and Pinnacle resulted in quicker development time and less cost.

GEAQS also incorporated many of the visualization techniques described by William Cleveland. The top-down approach is an iterative process. Micro 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. 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. The detailed concept of GEAQS was thoroughly discussed in Weir (1996). In

this report we evaluate the efficiency of GEAQS when it mimics the existing edit system.

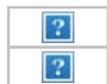
2. Implementation

To significantly reduce data collection, processing, and software development and lifecycle costs, EIA is developing a new Common Collection and Processing System (CCAPS) and Master Universe Database (MUD). It is highly desirable for an efficient data editing system to be adopted in CCAPS. The surveys used to produce the Petroleum Marketing Monthly, the EIA-782A and B, were the pilot surveys used in the development of GEAQS. Prior to initiating the development of CCAPS, this pilot project demonstrated the potential of graphical editing. The surveys chosen collect 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.

More recently, another survey, the EIA-857: "Monthly Report of Natural Gas Purchases and Deliveries to Consumers", was selected to further evaluate the benefit of GEAQS, the use of microcomputers and Window's applications. In this survey, a sample of 376 natural gas companies, including interstate pipelines, intrastate pipelines, and local distribution companies report. The data collected monthly at the state level include the volume and cost of purchased gas, and the volume and cost of natural gas consumed by sector. Approximately 50 percent of the responses are received by the due date. A series of manual and computerized edit checks using a mainframe computer are used to screen the data. Six variables are the basis for the input to the primary checks in the edit: the price of residential, commercial and industrial sales; the volume of residential, commercial and industrial sales.

For this project, we implemented GEAQS using three rounds of editing, at different points in the production process within a month/reference period for each of three consecutive months. A total of nine rounds of edit were evaluated using GEAQS and the existing edit system, imposing the identical, historical editing rules.

The editing rules for the EIA-857 are based on changes in individual respondent's market shares and percent change from previous period for volumetric data and price data. A respondent's market share is calculated as:



In the graphical presentation, all error flags (E0, E1, E2, and E3) for price or volume are displayed in red and warning flags (W1) are displayed in yellow. All other prices and volumes are displayed in blue.

For the purposes of presenting aggregate data, the same rules were applied at the aggregate rather than the respondent level. However, for volumetric aggregates the market share is always 100, so the only flag possible is E1. For price aggregates, both E1 and W1 are possible.

3. Results

To exploit the time series capabilities within the GEAQS, a total of 14 months of data were read and edited in order that the times series graphs displayed in each round of editing were sufficient. The current system's data, however, were stored on the mainframe in ADABAS. Data were, therefore, downloaded and preprocessed for the PC environment. At this stage, the speed and capacity of the microcomputer played a significant role in the preparation of the data. At first, a 200 MHz Pentium with 64 MB RAM microcomputer was used, and the process of downloading and preprocessing data took about 100 minutes to complete. A 450 MHz Pentium III with 128 MB RAM microcomputer was then tried which finished the download and preprocessing around 20 minutes.

Generally, the current production editing process requires 40 man-hours to screen over 400 pages of computer printout to identify any unusual respondents in each round of edit, then link to the company list to do the follow-up. With GEAQS, it took only 8 man-hours of computerized edit and to get the company information within the system.

The following is an example of the tables used to summarize the data flagging after the edits in GEAQS and the existing edit system for one round of editing for one month:

Table example of data flagging summary

		857 Current Editing Procedure			
		Warning	Error	Non-error	Total
GEAQS Current Edit	Warning	19	0	0	19
	Error	12	94	0	106
	Non-error	0	0	203	203
	Total	31	94	203	328

For the six variables involved in the primary checks, the tables were gathered, summarized, and the efficacy of GEAQS in replicating the existing edit system examined. These results were then reported to the survey manager for further investigation. It appears that for this table, there was only one cell that showed a deviation between the two systems. For twelve occurrences, variables that were identified as warnings in the batch hard copy output system produced errors in the graphical system. These records will be further examined to determine if any other criteria were applied in the current system or were misprogrammed in the graphical system.

GEAQS was intended to be interactive with the database of the processing system. Once a particular respondent value has been identified, the analyst would change the response directly in the spreadsheet, if so desired. The analyst would then be able to re-examine the newly computed aggregates to determine if it remained an anomaly.

Other Results

The intent of this pilot was to examine the feasibility of GEAQS from a more logistic point of view of getting the data to the system and providing flagged data that reflect the current system's output of flagged data. However, because the flagged data are presented graphically, GEAQS also provided a bigger picture of the editing in terms of the efficiency of the edit rules themselves.

For example, for a particular round of editing, the anomaly map for a particular sector (commercial) showed the aggregate (the published weighted average price for Maryland), to have failed the edit. GEAQS displayed the time series of published commercial prices for Maryland, as shown in figure 1 below. Errors were displayed for the last month, July, and for December. Warnings were shown for January, April, May and June. The individual respondent level data used to form the aggregate for each of those time periods are shown in figure 2.

Figure 1. Aggregate published prices for 14 months



Figure 2. Respondent level price data for 14 months



Figure 1 shows that the average price of commercial natural gas was 8.19, up from 6.65 the previous month, a 23% increase. This change exceeded the 15% limit, causing an E1 error flag, represented in the graph by red. The graph of the time series of aggregates, however, depicts the variability of the aggregates over the longer period, not just with

respect to the last month. Figure 2 provides the detailed data that form the aggregate. It appears from the comparison of July to June that prices overall are up, so it is only the magnitude of the price increase in the aggregate that is questionable.

It can be seen from this graph that the two highest priced respondents have error flags, and the third respondent has a warning flag in the last month, July. The distribution of the respondents' prices for July can be compared to that of the previous month and month year ago. Data for a particular respondent can be tracked across the series by clicking on the data points, showing the respondents history of flagged data, and their price position relative to the other respondents. This data display takes on a box-whiskers like graphic image, minus the whiskers and box. In this figure, the highest priced respondent in July had both a price error, E1, and a volume/market share warning, W1. The W1 is the result of the volume decrease of 65%, but a change in market share of less than 4.99%, indicating that this price change did not have a large impact on the aggregate. This same respondent's price is shown to only be fourth highest in June with no flags, but fifth highest in May with both a price error and a volume/market share warning.

Figure 3 shows the detailed respondent volume table associated with the prices in figure 2.

Figure 3. Respondent level volume data for 14 months



This graph shows that in July the highest volume respondent maintained his high market share at roughly 71% and did not have a price flag in July. More importantly though, the second largest respondent who also did not have a volume/market share flag, did have a 25% market share and a price flag of E1. This respondent's price change had the largest impact on the aggregate price. By combining information, rather than keeping volume/market share and price edit information separate, figure 2, which shows respondents prices, could be modified to show the second point in July as the significant data point, based on their marginal contribution to the aggregate change (Weir, 1996), rather than just the potential for impact.

It is through this type of finding that GEAQS can provide a new way to evaluate the current edit rules, and potentially test alternative rules, to reduce the amount of correctly reported data failing the edits and producing Type II errors. Type II errors, data flagged that result in no change upon follow up, are costly and may even contribute to Type I errors, incorrect data that are not changed.

4. Future Enhancements

One main additional enhancement is yet to be made in GEAQS. This major enhancement, called "bubble up" would provide the functionality that anomaly information at the point of the highest levels of aggregates concerning the associated lower levels of aggregation is provided to the user. 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 were outliers at those levels.

For this pilot of GEAQS, some data transformation capability was incorporated so the user could visualize intrinsically linear data. Also, a second set of editing rules were programmed and provided as an option in the opening dialogue box for graphic selection. This allowed the user the ability to compare the two sets of rules.

It is expected that when GEAQS is incorporated into the generalized processing system, other variables, criteria, etc., in the database would be available to display visually with the data. Respondents who failed edits in the batch process would be flagged in the spreadsheet and scatter gram. Clicking on the "company" icon could access recorded comments obtained by contacting respondents. Standard errors of aggregate estimates would also be incorporated as a data type that could be viewed graphically and through spreadsheets. Changes in the standard errors of aggregates between survey periods could help to identify reporting errors, similar to the box plots, while consistently high standard errors would provide information for design issues and sample sizes.

5. 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, Working Paper No. 15.

Cleveland, William S., (1993), Visualizing Data, Hobart 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, Working Paper 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.

Weir, P. (1996), "Graphical Editing Analysis Query System (GEAQS)," Data Editing Workshop and Exposition, Statistical Policy Working Paper 25, pp. 126-136. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

Weir, P. and Emery, B., Walker, J. (1996), "The Graphical Editing Analysis System Developed by EIA ", 1996 ASA Proceedings, Section on Survey Research Methods, Chicago, Ill.