WP.26
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (v): New and emerging methods

SEEING THE FOREST THROUGH THE TREES: A GRAPHICAL VIEW FOR EDITING THROUGH E-SPHERE

Invited Paper

Prepared by Paula Weir (Paula.Weir@eia.doe.gov) Energy Information Administration, Department of Energy, United States and Lindolfo Pedraza (under contract to EIA)

I. INTRODUCTION

1. In 2004 the U.S. Energy Information Administration (EIA) implemented the U.S. Census Bureau's Standard Economic Processing System (StEPS), a SAS based system for processing survey data. At this time, 18 petroleum and natural gas business surveys are being processed in this system at EIA. Even though StEPS has multiple types of rule-based edits available, the presentation of edit failures does not allow analysts to see the broader context within which observations fail the edits. For example, while two observations might fail a given edit, their overall effects on publishable estimates may not be the same. Interactive graphical editing is currently being developed to complement the StEPS edits by providing context for the micro-data in order to prioritize failed data, re-evaluate clean data and imputed data, and improve the rule-based edits.

2. A graphical approach to viewing data, along with minimal paradata, is being developed for a number of related surveys Using SAS Insight,. This tool provides survey analysts with a broad view of how survey respondents interact. Interactive graphs have been successful tools in some StEPS Census branches, where analysis focuses on annual snapshots of industry totals and sub-totals. EIA's focus on repetitive snapshots of global economic indicators with complex supply chains and production processes requires additional data dimensions. To this end, cross-sectional and historical-series approaches are being developed to analyze data across surveys and to fully capture geographical, economic and physical constraints.

3. In order to simplify data and graph selection for the users, a simple user interface was developed which allows the user to select the graph type desired from the set listed, while a second layer menu asks the user for the data criteria desired for that selected graph. The interface interactively creates the data set from the user-defined criteria using the StEPS data files, and displays the graph along with the corresponding SAS dataset. All SAS Insight capabilities, such as hiding data points, remain at the disposal of the user interested in further exploration of that selected dataset. In addition, in order to provide further context for an individual survey cell response as it affects the publication tables, graphic representations of the dissemination structure which combines data from multiple surveys in the supply chain are also being developed to aid the analyst in identifying and defining the user data criteria for graph selection.

II. THE PROBLEM

4. The Monthly Petroleum Supply Reporting System (MPSRS) represents a family of nine data collection survey forms that are used to collect detailed refinery, natural gas plant and oxygenated plant operations data; refiner, bulk terminal, oxygenate plant, natural gas plant and pipeline stocks data; crude oil and petroleum product imports data; and data on movements of petroleum products and crude oil between the geographic regions. The integration of the data from the nine forms filed by multiple players in the supply chain allows for the calculation of products supplied. Products supplied are used as proxies for consumption or demand in that they measure the removal of petroleum and petroleum products from primary sources, i.e., refineries, natural gas processing plants, blending plants, pipelines, and bulk terminals.

5. In 2004 the survey processing system for the family of monthly surveys was replaced by the Standard Economic Processing System (StEPS), developed by the U.S. Bureau of the Census. In addition, a number of survey form changes were also implemented that provided new product details in keeping with changes in the industry. Attempts were made within the StEPS environment to reconstruct the previous system's editing rules based on historical consistency, but not all were implemented, and no data were available for editing the more detailed new products. Furthermore, across the nine surveys, the number of data elements or cells is fairly large. For example, seven supply types are collected (beginning stocks, receipts, inputs, production, shipments, fuel uses & losses, and ending stocks) for one product on one survey, such as, conventional gasoline blended with alcohol on the Monthly Refinery Report. Of these seven supply types, four have edits applied at the cell level, and two more edits are applied to the seven supply types as a group, such as the edit which flags a line imbalance. The StEPS edits for the cells were created by survey subject specialists and include: prior value no current value; current value no prior value; part greater than total; first time reported zero in the last 12 months; first time not zero in last 12 months; greater/less than maximum/minimum value reported in the last 12 months; and difference between beginning stock this period and ending stock last period; greater than parameter; negative value; consistency between product; and components balance. On the other hand, one of the nine surveys, the Monthly Imports Report, has no editing at all applied because of the erratic nature of imports. Estimates produced by the MPSRS combine data from the nine surveys, but no editing is performed across surveys at either the micro or macro level.

6. Currently in StEPS, the review and correction of items that fail the edits is performed by examining each respondent's data line by line, and scanning down a screen of responses that flags those items that failed the edits, as shown in Figure 1. A red marker highlights an item that has failed an edit. By clicking on the failed item name, and selecting "view edit failures", the edit failures for that respondent are displayed (Figure 2), along with the number and description of the edit rule that the item failed.

Figure 1. Review and Correction

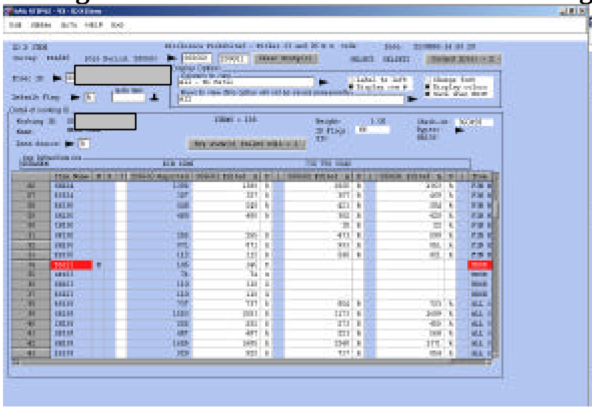
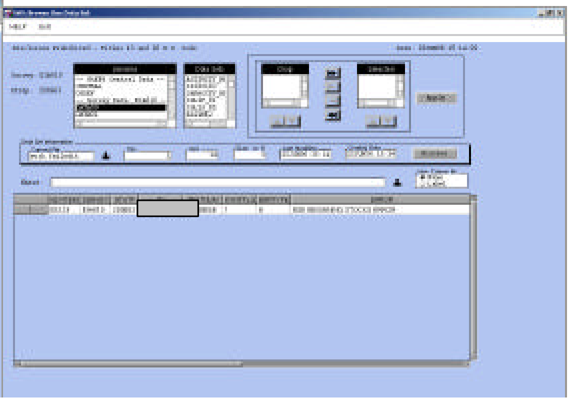


Figure 2. View Edit Failures



7. However, the compilation of components across surveys cannot be performed in StEPs, so aggregates and dissemination-level estimates cannot be examined until very late in survey processing,

making it difficult to identify the impact of potential reporting problems. Response error, or the difference between the true value and the value reported on a survey form, is considered to be the major factor affecting the accuracy of these data. Hundreds of edit rules have been implemented across the surveys. The large number of resulting edit failures, along with the presentation of the edit failures, unintentionally focuses resources on minor data corrections and error failure overrides, rather than focusing on the key errors with respect to the broader data picture. Graphical editing using SAS Insight was examined as an approach to view the data at the cell level, and provide perspective on the reported data and edit status through the context of the respondent's historical data or other respondents' data. To enable the graphical editing approach, a suite of tools and user interfaces were developed.

8. The analytical approach used at the U.S. Census Bureau focuses on annual snapshots of establishments and industry totals and sub-totals. EIA, on the other hand, focuses on capturing the flow of energy through the supply chain and into the economy. StEPS, reflecting the nature of Census data, falls short of the cross-sectional and historical-series approaches needed to analyze data across EIA surveys to fully capture geographical, economic and physical constraints across the supply chain. The data analysis system this paper is based on allows the analyst to drill down to a questionable data point from a distinguishable node in the supply chain by navigating through data from different surveys using the integrated analysis tools such as supply chain maps, data catalogs and interactive SAS-based plotting system. The simple user interface will allow the user to see StEPS-based data in the supply chain context without modifying StEPS itself, by providing an additional window containing the same data within the supply chain context. This is done through the creation of hybrid Oracle-SAS views that allow quick download of real-time data without interfering with StEPs processing. Moreover, analytical procedures and strategies for outlier identification focus not only on the detection of outliers, but, most importantly, on the identification of the most significant outliers within the petroleum supply chain.

III. THE FOREST

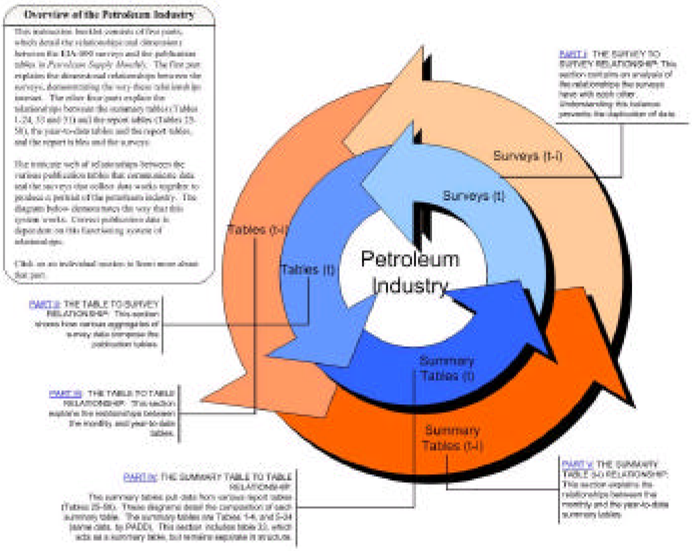
9. Early in the data editing and processing stage, the analyst first needs to see a data item with respect to previously reported periods. StEPS provides this capability, allowing the analyst to view as many as three previous reporting periods for any one data item. It does not, however, provide the analyst with a broader seasonal view of the data, which is equally important in understanding the data context. Additionally, the analyst may need to view a data item with respect to another item on the report. This capability is also provided through the review and correction screen in StEPS. Secondly, the analyst needs to see the data item with respect to other respondents' observations for that data item. This capability is not provided in StEPS. Thirdly, the analyst needs to view preliminary aggregates throughout the processing stage. These aggregates also provide analysts with the ability to determine individual respondent's affect on those aggregates. This capability is also not available in StEPS.

10. StEPS and its data manipulation approach could, to an extent, provide the cross-section and time-series properties in the analytic forms described above. However, its process is limited since StEPS data manipulation is based on independent "data steps" and merger blocks for each time series observation, making consecutive time series displays slow and cumbersome. Instead, the data analyses described here are based on a global notational approach; all programs are based on the same mathematical notation across platforms and intertwined areas of study. Global approaches combined with drill-down graphs and data catalogs, interactive data analysis tools and hybrid data management systems, such as Oracle-SAS views and open-ware software, allowed a more integrated approach that considers survey data as part of a larger survey system. The resulting programs are able to process larger amounts of data for varying periods, and different aggregation levels geographically, across product, or classifications.

11. To account for different ways of viewing the data, bottom-up and top-down, two approaches were developed and integrated. A direct user interface to choosing a graph type, data item, and reporting period(s) were developed, as well as an interactive instruction booklet that maps the final data product to a reported item. Figure 3 depicts the overview of these relationships, and serves as a navigation map.

12. Data used to populate the 58 tables in EIA’s Petroleum Supply Monthly (PSM)

Figure 3. Depiction of the final data product



Overview of the Petroleum Industry

This instruction booklet consists of five parts, which detail the relationships and dimensionality between the EIA-800 surveys and the publication tables in *Petroleum Supply Monthly*. The first part explains the dimensional relationships between the surveys, demonstrating the way these [illegible] interact. The other four parts explain the relationships between the summary tables (Tables 1-24, 33 and 31) and the report tables (Tables 25-58), the year-to-date tables and the report tables, and the report tables and the surveys.

The intricate web of relationships between the various publication tables that communicate data and the surveys that collect data works together to produce a portrait of the petroleum industry. The diagram below demonstrates the way that this system works. Correct publication data is dependent on this functioning system of relationships.

Click on an individual section to learn more about that part.

PART I: THE SURVEY TO SURVEY RELATIONSHIP: This section contains an analysis of the relationships the surveys have with each other. Understanding this balance prevents the duplication of data.

PART II: THE TABLE TO SURVEY RELATIONSHIP: This section shows how various aggregates of survey data comprise the publication tables.

PART III: THE TABLE TO TABLE RELATIONSHIP: This section explains the relationships between the monthly and year-to-date tables.

PART IV: THE SUMMARY TABLE TO TABLE RELATIONSHIP: The summary tables pull data from various report tables (Tables 25-58). These diagrams detail the composition of each summary table. The summary tables are Tables 1-4, and 5-24 (same data, by PADD). This section includes table 33, which acts as a summary table, but remains separate in structure.

PART V: THE SUMMARY TABLE 3-D RELATIONSHIP: This section explains the relationships between the monthly and the year-to-date summary tables.

publication come from nine surveys that collect data at key points of the supply chain such as tankers, pipelines, refineries and blending units. For example, Table 54 reports data from refineries, bulk terminals and natural gas producers. On the other hand, Table 53 reports data from refineries, bulk terminals, blending units, monthly crude oil reports, natural gas producers, and oxygenate producers. Figure 4 shows how the analysis system being developed allows analysts to quickly identify the surveys that provide data to each table. From there, scatter, box, “spaghetti” (stacked time series) and percent change plots are available for each survey.

13. Survey and publication analysts currently only have two perspectives from which to analyze the data--the publication tables or the StEPS screens (Figures 1 and 2). Numbers in the publication tables are composed of data from more than one survey. StEPS, however, only provides the view of the data from a single survey observation, without context to the survey or publication aggregated figures. Data manipulation routines, global programming notation, drill down tables, graphs and interactive tools in the future will allow the analysts to trace survey data from the publication tables and also analyze survey data independently of the publication tables.

14. Figures 4 and 5 depict a “side-ways” view of how hybrid data manipulation processes interact with drill-down web-pages and interactive scatter plot programs using global notation, and presented to the user through an interactive instruction booklet. The objective in using global notation is to bring all survey databases and indexing routines into a distinct set of variables and observations with homogenous properties. Seasonal views, views of data items with respect to other items, and views of data items with respect to other respondents’ data items are possible by the use of homogeneous variable properties and indexing processes. While homogeneous variable formatting and nomenclature are straight forward, the implementation of homogeneous indexing requires a unified super-survey frame to link each survey respondent across survey samples. Ongoing work is addressing homogeneous indexing across survey samples.

Figure 4. Drill-down map of the user interface

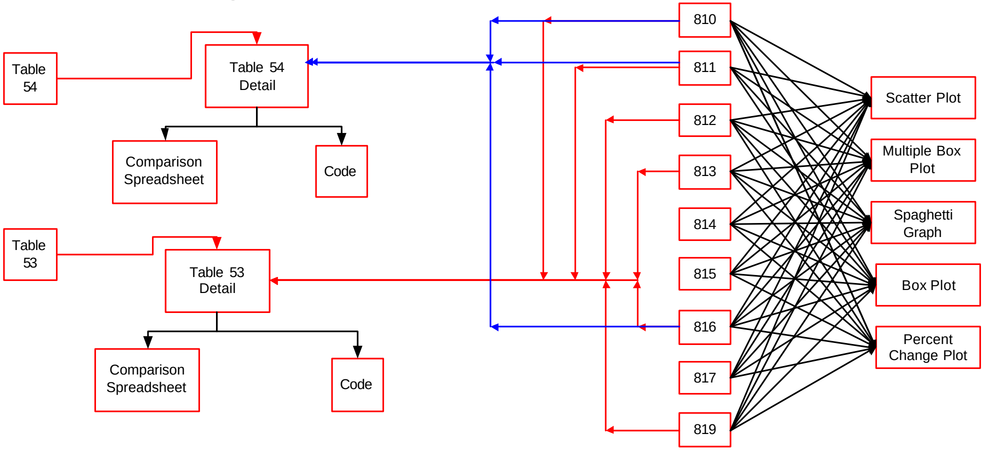
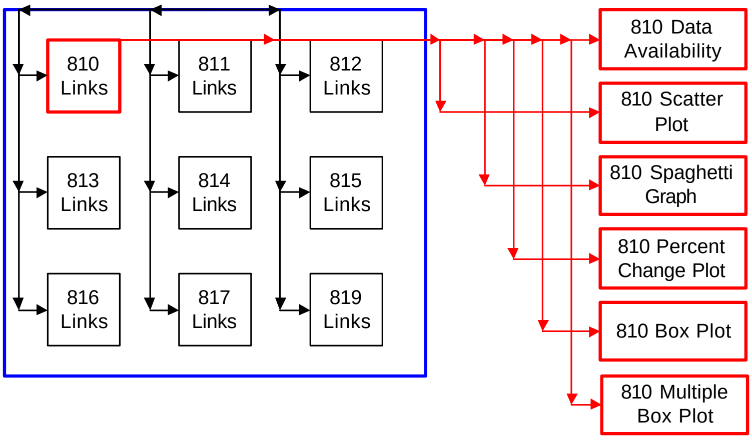


Figure 5. EIA-810 Navigation processes



15. The first part of the interactive instruction booklet uses diagrams to depict the relationship of summary tables in the publication to the more detailed tables as they represent the components of petroleum supply and disposition. The next four parts of the booklet continue the drill down through the publication tables illustrating lower table relationships, as well as year-to-date tables and detailed tables, through to the individual survey components of the reported level data. In particular, Part II of the interactive booklet shows how survey data compose each of the publication tables. Clicking on this section in the overview menu, after Figure 3, brings up the individual surveys, as shown in Figure 6. Clicking on a specific publication table listed in the drop down menu on the left highlights the corresponding surveys used to create the table. Figure 7 illustrates the display for the selection of Table 53. The alternative view of the same process was shown in Figures 4 and 5. All of these graphics were initially designed using off-the- shelf software.

16. The first step of the designed data analysis approach, in this case Table 53, is shown in Figure 8. This table-mapping screen summarizes the individual surveys contributing to the table, and shows the actual aggregate data values in the same format as the publication table. Future versions of the system will use drill down starting from the pseudo publication figures in figure 8 through to the specific multi-survey interactive scatter plots. The interactive scatter plots are then drawn with the use of hybrid data management systems depicted in figures 4 and 5.

Figure 6. Table and Survey Relationship

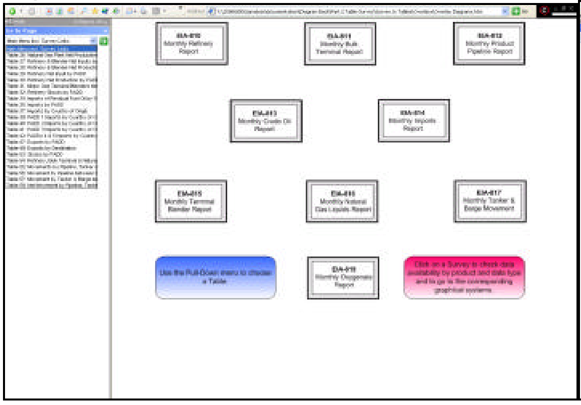
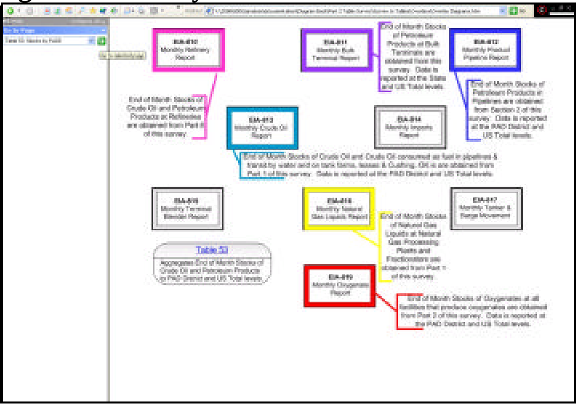

EIA-810
Monthly Refinery Report
EIA-811
Monthly Bulk Terminal Report
EIA-812
Monthly Product Pipeline Report
EIA-813
Monthly Crude Oil Report
EIA-814
Monthly Imports Report
EIA-815
Monthly Terminal Blender Report
EIA-816
Monthly Natural Gas Liquids Report
EIA-817
Monthly Tanker & Barge Movement
EIA-819
Monthly Oxygenate Report
Use the Pull-Down menu to choose a Table
Click on a Survey to check data availability by product and data type and to go to the corresponding graphical systems

Figure 7. Surveys used for Table 53



End of Month Stocks of Crude Oil and Petroleum Products at Refineries are obtained from Part 8 of this survey.

End of Month Stocks of Petroleum Products at Bulk Terminals are obtained from this survey. Data is reported at the State and US Total levels.

End of Month Stocks of Petroleum Products in Pipelines are obtained from Section 2 of this survey. Data is reported at the PAD District and US Total levels.

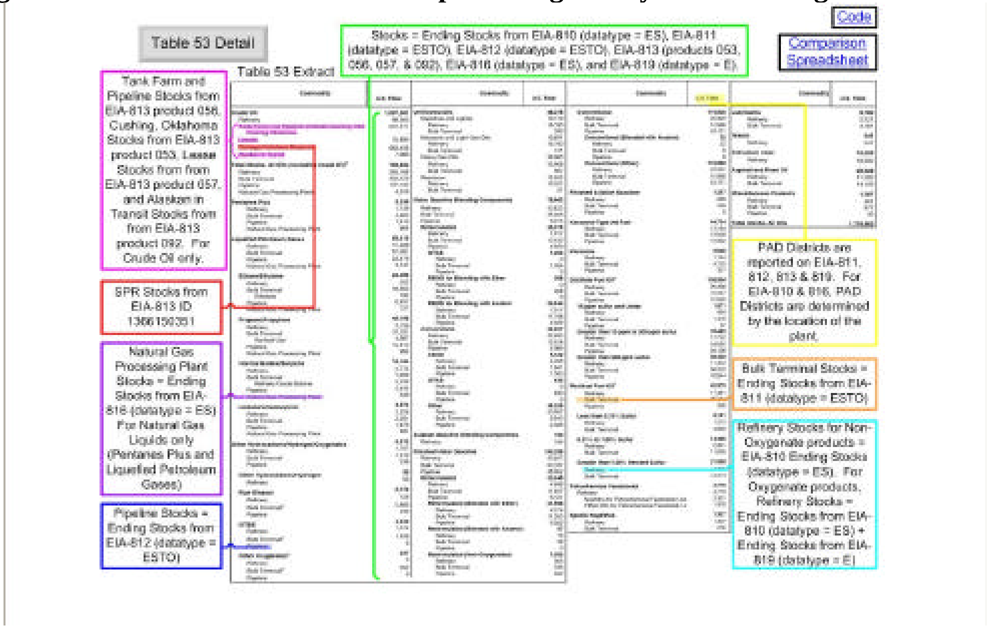
End of Month Stocks of Crude Oil and Crude Oil consumed as fuel in pipelines & transit by water and on tank farms, leases & Cushing, OK is are obtained from Part 1 of this survey. Data is reported at the PAD District and US Total levels.

End of Month Stocks of Natural Gas Liquids at Natural Gas Processing Plants and Fractionators are obtained from Part 1 of this survey

Aggregates End of Month Stocks of Crude Oil and Petroleum Products to PAD District and US Total levels.

End of Month Stocks of Oxygenates at all facilities that produce oxygenates are obtained from Part 2 of this survey. Data is reported at the PAD District and US Total levels.

Figure 8. Publication table marked up showing surveys contributing



Stocks = Ending Stocks from EIA-810 (datatype = ES), EIA-811 (datatype = ESTO), EIA-812 (datatype = ESTO), EIA-813 (products 053, 056, 057, & 092), EIA-816 (datatype = ES), and EIA-819 (datatype = E).

Tank Farm and Pipeline Stocks from EIA-813 product 056, Cushing, Oklahoma Stocks from EIA-813 product 053, Lease Stocks from from EIA-813 product 057, and Alaskan in Transit Stocks from from EIA-813 product 092. For Crude Oil only.

SPR Stocks from EIA-813 ID 1366150351

Natural Gas Processing Plant Stocks = Ending Stocks from EIA-816 (datatype = ES) For Natural Gas Liquids only (Pentanes Plus and Liquefied Petroleum Gases)

Pipeline Stocks = Ending Stocks from EIA-812 (datatype = ESTO)

PAD Districts are reported on EIA-811, 812, 813 & 819. For EIA-810 & 816, PAD Districts are determined by the location of the plant.

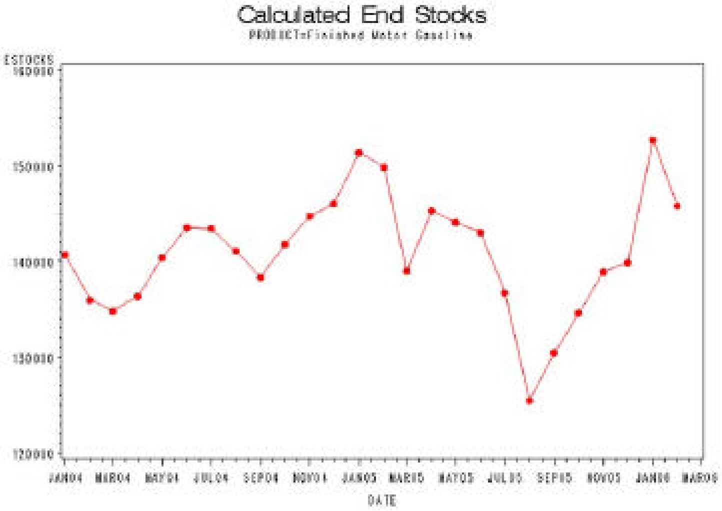
Bulk Terminal Stocks = Ending Stocks from EIA-811 (datatype = ESTO)

Refinery Stocks for Non-Oxygenate products = EIA-810 Ending Stocks (datatype = ES). For Oxygenate products, Refinery Stocks = Ending Stocks from EIA-810 (datatype = ES) + Ending Stocks from EIA-819 (datatype = E)

17.	Information from the marked-up table can, in this example, be used to deduce three facts: 1) Of the available stocks for total finished gasoline, 44%, 33% and 20% of stocks were located in bulk terminals (the EIA-811), pipelines (the EIA-812) and refineries (the EIA-810), respectively; 2) This distribution applies consistently across gasoline types 124, 125, 126, 128 and 130; 3) Gasoline product 130 carries most of the gasoline stocks’ volume with 83%, followed by gasoline product 124 with 14%.

18.	An example of a typical aggregate data concern is depicted in Figure 9, which shows a sharp decrease in stocks from June 2004 to June 2005. Using the facts determined through the marked-up table in Figure 8, the analysis is directed to the particular examination of product 130 in the EIA-811, EIA-812 and the EIA-810 surveys. Additionally, graphs and tables like the ones shown in Figures 8 and 9 are enhanced with drill-down connectivity to the interactive scatter plot tools.

Figure 9. Gasoline stocks since Jan. 2004



IV. THE TREES

19. For each of the nine surveys, a menu has been created. The user can navigate to this menu directly from Figures 6 or 7, having selected a table and survey of interest. These menus (Figure 10) contain options for the graphs available, as well as a helpful summary catalog of what data are available (by survey, product, and supply type), as shown in Figure 11.

Figure 10. Menu options for graphing data.

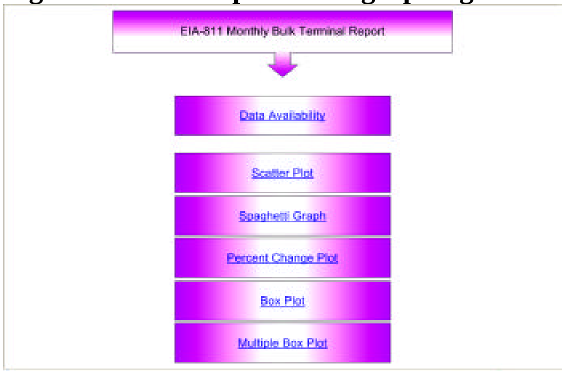
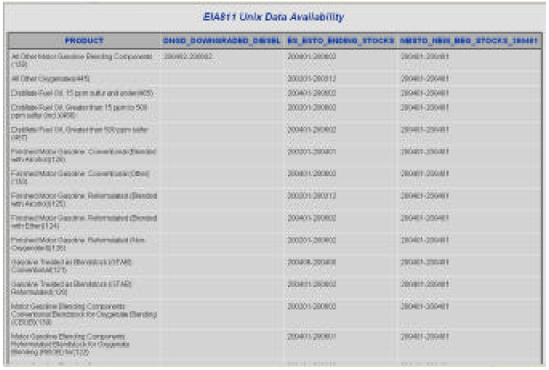


Figure 11. Data Catalog

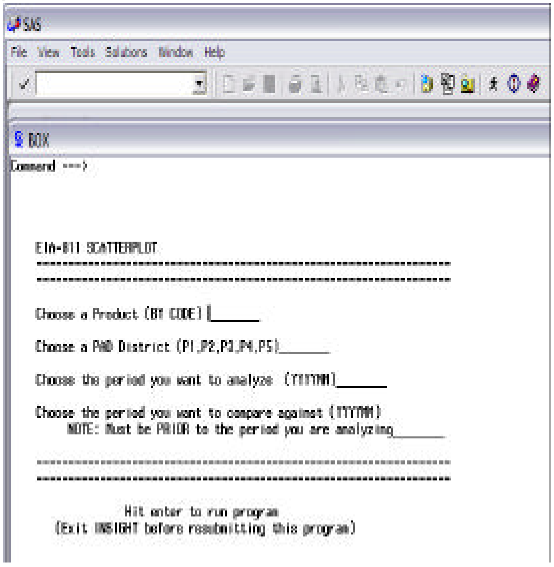


EIA811 Unix Data Availability

PRODUCT	CHSD_DOWNGRADED_DIESEL	EQ_ESTO_ENDING_STOCKS	NESTO_NEW_BEG_STOCKS_200401
All Other Motor Gasoline Blending Components (139)	200402-200602	200401-200602	200401-200401
All Other Oxygenates(445)		200301-200512	200401-200401
Distillate Fuel Oil, 15 ppm sulfur and under(465)		200401-200602	200401-200401
Distillate Fuel Oil, Greater than 15 ppm to 500 ppm sulfur (incl.)(466)		200301-200602	200401-200401
Distillate Fuel Oil, Greater than 500 ppm sulfur (467)		200401-200602	200401-200401
Finished Motor Gasoline, Conventional (Blended with Alcohol)(128)		200301-200401	200401-200401
Finished Motor Gasoline, Conventional (Other) (130)		200401-200602	200401-200401
Finished Motor Gasoline, Reformulated (Blended with Alcohol)(125)		200301-200512	200401-200401
Finished Motor Gasoline, Reformulated (Blended with Ether)(124)		200401-200602	200401-200401
Finished Motor Gasoline, Reformulated (Non-Oxygenated)(126)		200301-200602	200401-200401
Gasoline Treated as Blendstock (GTAB) Conventional(121)		200406-200408	200401-200401
Gasoline Treated as Blendstock (GTAB) Reformulated(120)		200401-200602	200401-200401
Motor Gasoline Blending Components Conventional Blendstock for Oxygenate Blending (CBOB)(158)		200301-200602	200401-200401
Motor Gasoline Blending Components Reformulated Blendstock for Oxygenate Blending (RBOB) for(122)		200401-200601	200401-200401

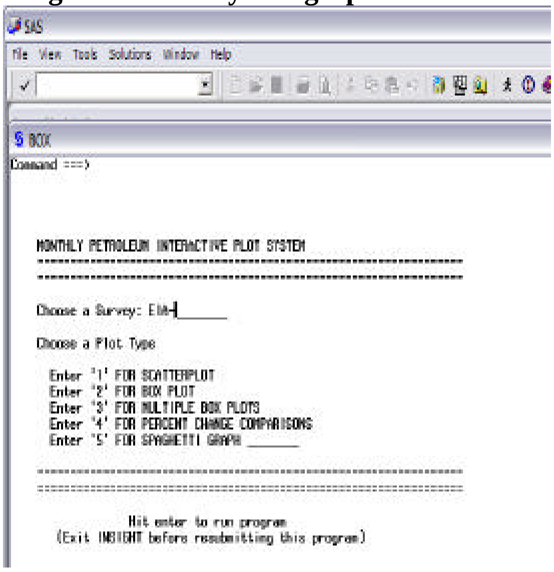
20. Under the data availability option, the user can scroll through the products in the left column, showing the set of products available in this survey. The time periods available for the each product by supply type are displayed, providing the analyst with the data dimensions needed to define the graphs. The types of graphs, the other menu items in Figure 10, include scatter, "spaghetti" (stacked time series), percent change (scatter graph of actual change vs. percent change relative to last report), box and multiple box plots. If the user were to choose a scatter plot, the menu shown in Figure 12 would be displayed. Using the information determined previously (the supply type, and time periods desired), the user would enter that information to define the data to be graphed. In this example, the user might start with the EIA-811 survey (based on its 44% share from above), PADD 1 (based on the severity of the drop in this region compared to the others) product 130, comparing June of 2005 to June 2004. However, the menus also provide the user with the navigational option to directly select the graph and survey, as might be the preference early in the processing cycle. If this is the preference, the user enters the survey of interest and the graph desired in the menu shown in Figure 13, and would then proceed to the menu depicted in Figure 12.

Figure 12. Supply type and time period



```
Command ===>

EIA-811 SCATTERPLOT
==================================================================
==================================================================

Choose a Product (BT CODE) _____

Choose a PAD District (P1,P2,P3,P4,P5)______

Choose the period you want to analyze  (YYYYMM)______

Choose the period you want to compare against (YYYYMM)
    NOTE: Must be PRIOR to the period you are analyzing______

==================================================================
==================================================================

            Hit enter to run program
    (Exit INSIGHT before resubmitting this program)
```

Figure 13. Survey and graph



```
Command ===>

MONTHLY PETROLEUM INTERACTIVE PLOT SYSTEM
====================================================================
====================================================================

Choose a Survey: EIA-______

Choose a Plot Type

  Enter '1' FOR SCATTERPLOT
  Enter '2' FOR BOX PLOT
  Enter '3' FOR MULTIPLE BOX PLOTS
  Enter '4' FOR PERCENT CHANGE COMPARISONS
  Enter '5' FOR SPAGHETTI GRAPH ______

====================================================================
====================================================================

            Hit enter to run program
    (Exit INSIGHT before resubmitting this program)
```

21. The information gathered in the user interface is provided to SAS Insight which views the data directly in StEPS, draws the graphs, and presents a worksheet of the data points in the graph. In this example, the selection of a scatter-plot for the EIA-811, product 130, and ending stocks for June 2005 compared to June 2004 is depicted in Figure 14. As shown on the left side of the figure, the scatter plot of actual values for the two time periods is displayed. Clicking on the data point indicated by the larger black circle invokes a pop-up box with the details of that data point (top right). In addition, the scatter plot of the actual change versus the percent change between the two periods is displayed (bottom right). The corresponding data point is also marked by the largest black circle in this plot, as well as the corresponding row in a spreadsheet of the data. In this example a respondent's ending stock was 156 in June 2004 but dropped to 0 in June 2005, and was the largest contribution from the EIA-811 to the 2005 drop in stocks.

22. Moving on to the EIA-812 survey, the user would do a parallel comparison, as shown in Figure 15. In this case, three data points have a 0 value in 2005, but the largest of these in 2004 had a value of 657. Each of these data points failed the "current no prior" edit for the first month that the value went to 0 in 2005, but there was no priority or context among the edit failures, and the 0 values were each accepted as valid. Similarly, the same type of comparison for the EIA-810 identifies a drop of 274 from 2004.

Figure 14. Comparison of stocks of product 130 June 2005 versus June 2004, EIA-811

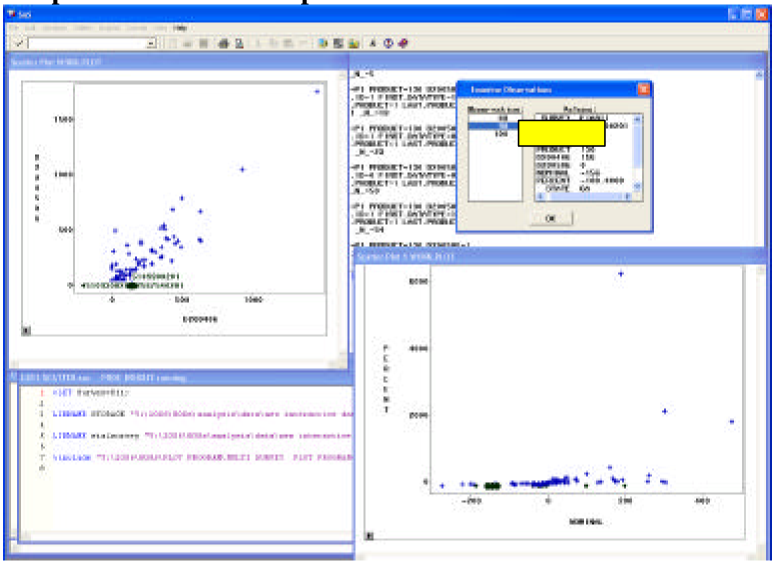
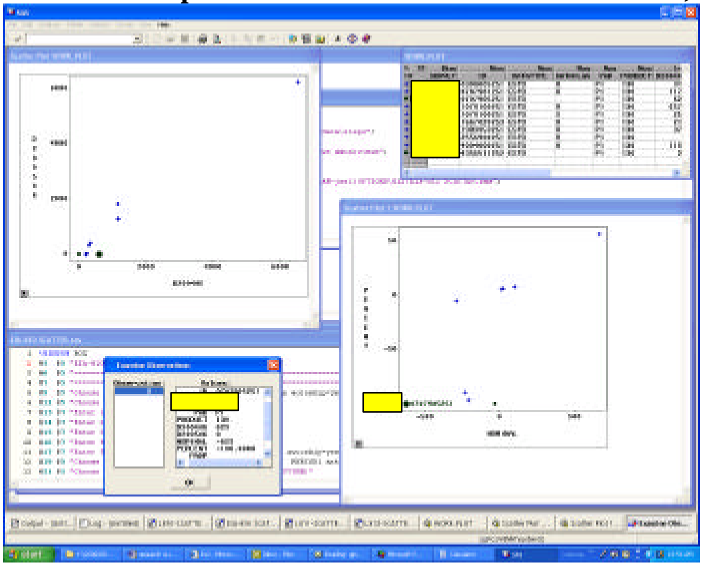


Figure 15. Comparison of stocks of product 130 June 2005 versus June 2004, EIA-812



V. SUMMARY

24. User interfaces to SAS Insight are being developed to provide data analysts with a selection of graphs to view the data reported on nine surveys that are combined in data dissemination. These interfaces define the data to be used in the SAS data sets, view the data in StEPS (the survey processing system), and manipulate the data as needed for the graphs as displayed to the user. These graphs provide the analysts with the data context needed to prioritize edit failures, as well as identify data not flagged by the edits. Higher-level mappings from publication tables to individual surveys contribute to the tables, while links to the graphing menus provide a higher-level context for investigating aggregate data issues. Depending on individual preference and the point in the processing system, the analyst can graphically edit from the top down or the bottom up.
