16
Chapter

Software Demonstrations

Software Demonstrations

[Alphabetically, by first author]

> ***Editors Note:*** Many of the Workshop presentations were accompanied by software demonstrations. Also, in some cases, the presenters chose to do only an exposition. All of the demonstrations are listed below. If a presentation was also included as part of the regular sessions, please see the appropriate chapter for the abstract or paper. If only a demonstration was given or if the presenters provided a separate paper to describe the exposition, that material is included in this chapter.

Richard J. Bennof, M. Marge Machen, and Ronald L. Meeks -- Data Editing Software for NSF Surveys

Ronald S. Biggar -- Generic Editor Developed with Powerbuilder and a Relational Database

Joel Bissonnette -- Generalized Edit and Imputation System for Numeric Data

Dale Bodzer -- PEDRO

Richard Esposito and Kevin Tidemann -- Extensions of ARIES at the BLS

Glen Ferri and Tom Ondra -- Towards a Unified System of Editing International Data

Robert Hood -- Improving the Quality of Survey Data Through an Interactive Data Analysis System

Michael Horrigan, Polly Phipps, and Sharon Stang -- On-Line Edits in the Survey of Employer-Provided Training

Givol Israel -- Automatic Transferring from Paper to ASCII Code

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

Stanley E. Legum -- A Computer-Assisted Coding and Editing System for Non-Numeric Educational Transcript Data

Sharon Mowry and Jason Bulson -- Distributed EDDS Editing Project (DEEP)

Mark Pierzchala, Roberta Pense, and Arnie Wilcox -- The 1995 June Area Frame Instrument for CAPI and Interactive Editing

Anne Rhodes, Kishau Smith, and Peter Goldstein -- Electronic Data Collection: The Virginia Uniform Reporting Sytstem

Sylvie Rivest and Mike Bankier -- Imputing Numeric and Qualitative Variables Simultaneously

Peter Sailer, Terry Nuriddin, and Gary Teper -- Editing Occupations from IncomeTax Returns

Janet Sear and Peter Garneau -- PERQS (Personalized Electronic Reporting Questionnaire System)

Linda Simpson -- A Statistical Edit for Livestock Slaughter Data

J. Tebbel and T. Rawson -- CDC Edits: Tools for Writing Portable Edits

Paula Weir -- Graphical Editing Analysis Query System (GEAQS)

Demonstration Abstracts

❒ Generic Editor Developed with Powerbuilder and a Relational Database -- ***Ronald S. Biggar, National Center for Health Statistics***

An interactive demonstration of both editing National Ambulatory Care Survey data and imputation of nonresponse. The software was developed for use by statisticians and may be modified as survey requirements change. Powerbuilder was chosen as the development tool so that statisticians could interact with the data from their PC's, while the data are stored on a file server. In addition, the Windows-based graphical user interface improves the user friendliness of the process.

❒ PEDRO -- ***Dale Bodzer, U.S. Energy Information Administration***

PEDRO is an electronic data collection product that facilitates the fast, accurate, and efficient transmission of data from the respondent's remote site to the EIA computer facility. Using a PC for data entry, PEDRO provides the user with an image of a printed survey form. Users can enter information through the keyboard or by importing data from another computer system. PEDRO performs numerous quality checks comparing the data entered with established ranges, lists of accepted values, or criteria derived from data entered in the past. PEDRO automatically transmits the information via modem to EIA' computer facility. Security of the transmission is protected by passwords and all data are encrypted. Accuracy is ensured by several levels of error detection. PEDRO is available to respondents.

❒ Improving the Quality of Survey Data Through an Interactive Data Analysis System -- ***Robert Hood, National Agricultural Statistics Service***

The National Agricultural Statistics Service (NASS), an agency of the U. S. Department of Agriculture, conducts surveys in order to provide accurate and reliable agricultural forecasts and estimates for a variety of commodities. NASS has recently begun the implementation of an interactive data analysis system based on SAS/AF and SAS/EIS software to ensure the quality of its survey data.

Currently, a lot of time is spent editing incoming data with little time devoted to analysis before data are summarized. Present analysis tools are limited to data listings and outlier printouts. While useful, they are somewhat limited in the problems they flag, and resolution of problems generally involves time-consuming review of paper questionnaires or data files. The time between data collection and summarization is very limited and must be used as efficiently as possible.

The authors have developed a SAS-based application to interactively analyze survey data. This system identifies potential "risky records" during the data collection period. Users are able to more efficiently analyze the data and resolve problems in a more timely manner. This Interactive Data Analysis System (IDAS) is an easy to use mouse-driven system that requires little knowledge of the SAS system. Pushbuttons, icons, list menus, and program entries provide easy selection of options. This paper gives an overview of IDAS and its development.

❒ On-Line Edits in the Survey of Employer-Provided Training -- ***Michael Horrigan, Polly Phipps, and Sharon Stang, U.S. Bureau of Labor Statistics***

The Survey of Employer-Provided Training 2 (SEPT2) was conducted from May to October, 1995 by the Division of Special Studies of the Bureau of Labor Statistics (BLS). The survey data were collected by BLS regional field economists using laptop computers during a personal visit to establishments. Employer and employee representatives were interviewed for SEPT2, and several instruments were administered to each respondent. Due to multiple survey instruments, the decentralized nature of the survey and the desire to avoid telephone callbacks to respondents to clarify inconsistencies, on-line edits with edit-error messages and summaries were included in the SEPT2 instruments. In addition, the SEPT2 laptop system included real-time reports on the status of establishment cases, a function for leaving case notes, and it incorporated Windows standards, such as a graphical-user interface and help system.

Edits involved logical, range or other consistency checks within an instrument.

During an interview when an answer to a question triggered an edit error, a pop-up window appeared with a message describing the error, what questions it involved and how to resolve it. The field economist could move on to other questions, but needed to resolve all errors before the establishment could be transmitted as complete. To resolve an error in situations where data items were not available from the respondent (missing) or when the datum was correct, field economists could select a comment for the question, such as data not available, verified by respondent or employment growth and provide case notes to describe the inconsistencies. When exiting an instrument,an edit error summary was automatically displayed, listing the remaining edit errors by question and the respective comment code. The field economist could simply double click on the question to return to it and resolve the edit error before exiting the instrument.

❒ Distributed EDDS Editing Project (DEEP) -- ***Sharon Mowry and Jason Bulson, Federal Reserve Board***

A key resource supporting the monetary policy-making and open market operations of the Federal Reserve is data representing the daily balances of deposits, borrowings, and reserves for the largest 7,700 depository institutions in the United States. The data require a high level of confidence, and identifying and resolving errors in quality must be done under increasingly stringent deadlines.

The Distributed EDDS Editing Project (DEEP) was initiated with the objective to improve the data analysis effort through the utilization of graphical user interface tools in conjunction with the presentation of statistically significant data edits. The DEEP system is a Windows-based client/server application designed to provide analysts with a sophisticated tool to access both raw data and data that have fallen outside of the data model forecasts based upon five different types of data edits or forecasts. Our presentation will detail the manner in which features of the DEEP application are employed for the editing and analysis of these critical data.

❒ Imputing Numeric and Qualitative Variables Simultaneously -- ***Sylvie Rivest and Mike Bankier, Statistics Canada***

At the Bureau of the Census 1996 Annual Research Conference, a presentation entitled "Imputing Numeric and Qualitative Variables Simultaneously" will be given by Mike Bankier. It describes the New Imputation Methodology (NIM) that will be used in the 1996 Canadian Census to impute the basic

demographic variables: age, sex, marital status and relationship to person 1. The NIM allows, for the first time, minimum change hot deck imputation of numeric and qualitative variables simultaneously.

As a follow-up to the presentation at the Annual Research Conference, it is proposed to give a software exhibit of the mainframe implementation of the NIM which will be used in the 1996 Canadian Census. Mike Bankier, the Senior Methodologist responsible for NIM and Sylvie Rivest, the System Analyst, who implemented it, will be present. They will give short presentations during the 3 hour slot given to demonstrate the software. A brief outline of the presentation is given below.

A small slide show (PC-based) will explain the generalized nature of the NIM program and the input data required. The slide show will also present the User Edit Interface used by NIM in the mainframe environment (3-5 minutes).

A PC version of the NIM program will then be used to demonstrate the functionality of the imputation engine (5-10 minutes). The methodology supporting NIM can be demonstrated interactively as the program is being executed.

Finally, a short (3-5 minutes) slide show will demonstrate the effort of applying the NIM methodology on a large volume of data. Statistics from the NIM testing done so far will be used to show the improvements in Data Quality that the NIM methodology offers on a large scale basis. These statistics will open informal discussions with the other participants at this Software Exhibit.

❒ Editing Occupations from Income Tax Returns -- ***Peter Sailer and Terry Nuriddin, Internal Revenue Service, and Gary Teper, Information Spectrum, Inc.***

Special tabulations, such as selected tax return data classified by Standard Occupational Classification (SOC) codes, are produced by the Statistics of Income (SOI) Division of the Internal Revenue Service. To facilitate coding, a computerized occupation-coding dictionary has been developed. However, because the manual updating process is time-consuming and increases errors, a computer utility program was created to automatically research and edit or replace the occupational entries.

The program compares occupational titles (from tax returns) to similar occupational titles already used in the dictionary. If similar entries are suggested, the user can replace the original entry with the best suggested entry. If no entries are suggested, the user can edit the original entry and retry or determine the record uncodable and bypass. Display will include a computer with the editing software installed; hands-on experience will be possible.

Extensions of ARIES at the Bureau of Labor Statistics

Richard Esposito and Kevin Tidemann,
U. S. Bureau of Labor Statistics

Abstract: A PC-based demo of prototypes that build on and extend the existing ARIES graphical approach to editing Current Employment Statistics Data will be shown. These prototypes include an updated version of ARIES, which incorporates visual representations of statistical measures of sample standard deviations to be used as aids to outlier detection and a DOS version of ARIES adapted to statistical exploration of universe data.

ARIES (Automated Review of Industry Employment Statistics) is a graphical and query PC-based data review system which has significantly enhanced the sample screening and estimation review procedures in the Current Employment Statistics (CES) program of the U. S. Bureau of Labor Statistics. During the time that ARIES has been in use, a number of areas of possible improvement have been suggested by the industry analysts who are responsible for reviewing the CES sample data and resulting estimates. As a result of studying these areas of possible improvement, we have developed two further prototypes to extend the capabilities of ARIES, which were shown at the Data Editing Conference, and outline these prototypes in this paper. Fuller treatments of ARIES and the first prototype are given in the *Journal of Computational and Graphical Statistics*, June 1994, and in the ASA 1994 *Proceedings of the Section on Statistical Graphics.*

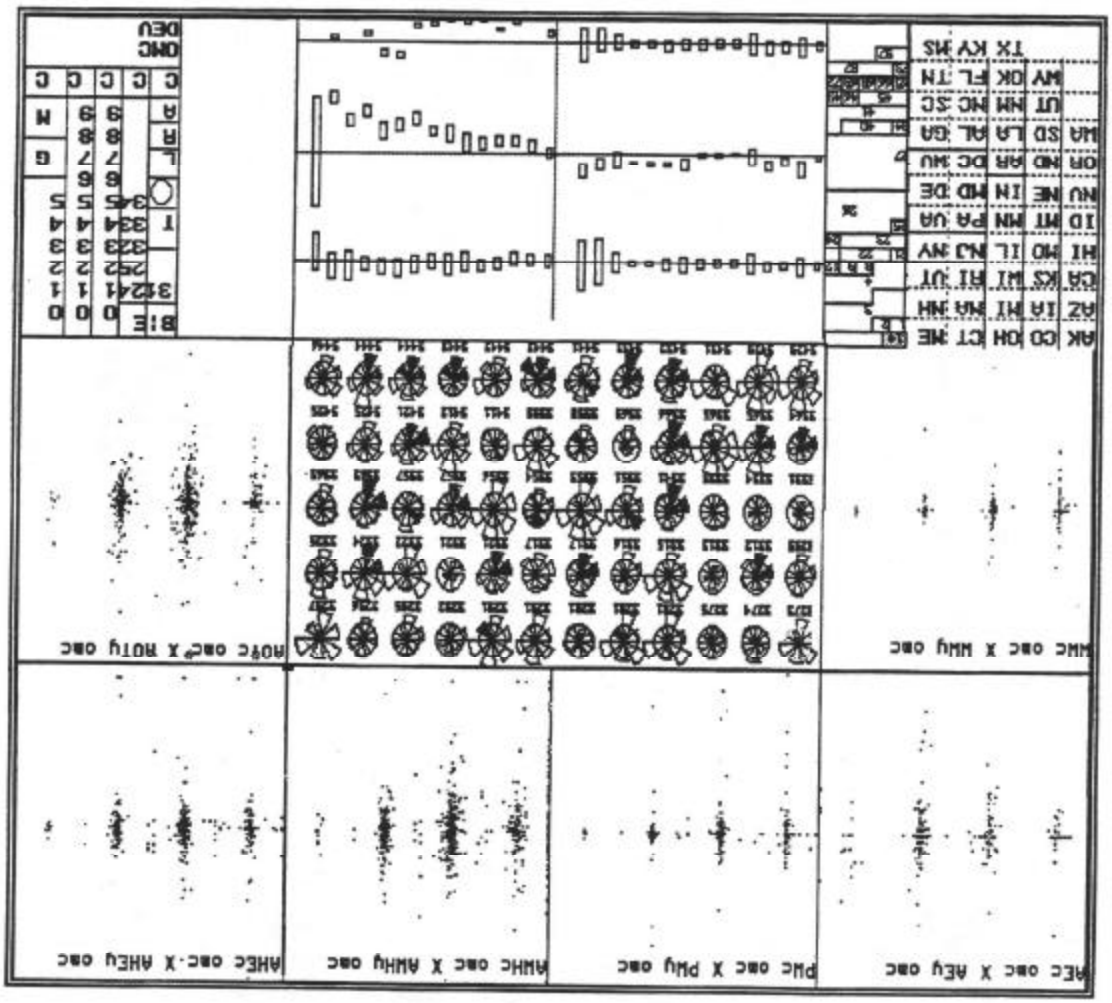
The prototypes have been designed to meet the specific needs of the CES data review process. As the prototypes move closer to production use, it is to be expected that their design and underlying principles may be suitably modified. Perhaps the most important principle evident in the prototypes is "put as much information on a single screen as reasonably possible." Figure 1, on the following page -- in which ARIES has been adapted to simultaneously show all 6 data variables -- shows the results of this idea.

Underlying this principle is the desire to have more layers of information immediately accessible, without having to page through an extensive hierarchy of screens. In Figure 1, rather than showing the scattergrams of the different variables on separate screens, all six variables are shown simultaneously (the upper row and left and right center row boxes). In this way, cross-comparisons among data variables are facilitated. Additionally, in the option shown in Figure 1, each scattergram has been subdivided by size of establishment (4 strata), so that the relative importance of sample establishments can be immediately seen. The actual PC screen uses colors liberally, as indicators of various characteristics and measures, including, somewhat redundantly, size of establishments.

Just as in ARIES, when one selects points in the scattergrams using a mouse, the corresponding establishment information will appear (not shown here), overriding the center and bottom row screens.

A second principle followed in this prototype is to make every element on the screens both informative and interactive. That means that each object on the screen should be designed to provide statistical information, as well as be used as an index to further information or further actions. This is true of the scattergrams, which both show statistical movement of the sample, as well as allowing retrieval of detailed individual or groups of individual establishment information.

Figure 1.--Scattergrams for 6 Variables, Daisy Variations, and Rough Tukey Box Plots



O.K., what are the strange objects in the center of Figure 1? First, each "daisy" of the array of 60 daisies in the center of the screen functions as an index to a specific SIC industry. When a specific daisy is selected by the mouse, the scattergrams corresponding to that industry immediately appear. In addition, rough tukey box plots corresponding to each of the six variables appear at the bottom row center. Each sequence of 15 vertical bars represents a monthly time-series of the sample standard deviations for that variable, with the current month's sample standard deviation represented by the rightmost bar within each of the six sequences. In our review, we concentrate on the new current month sample data, since we presume that we've already done an O.K. job cleaning up bad data for previous months. One possible indicator of outliers in the current month might be an especially large standard deviation for the current month as compared to previous months. Such a case is shown in Figure 1 for several of the variables.

Following the 2nd principle above, besides using the daisies as indexes to select statistical information, we have also made the daisies statistically informative: each of the 6 petals of each daisy represents one of the six variables we produce monthly estimates for in the CES, and the shape of each daisy

functions as an informative index to the rough tukey box plots just mentioned. If the size of a petal is larger than normal, that represents that the current month's standard deviation for that variable is larger than the average standard deviation of the previous 14 months. That average point is represented by the circle or "flowerpot" within each daisy. Each analyst will have approximately 250 industries to review, so by selecting those daisies that flop outside the circle, the analyst can pinpoint industries with unusually large standard deviations, and immediately show the associated scattergrams and box plots.

The following prototype shows ARIES adapted for universe data and for sampling and estimation capabilities:

Figure 2.--One Variable, 11 Months, with Establishments Selected in Scattergrams and Corresponding Sample Establishment Time-series Shown in the Center

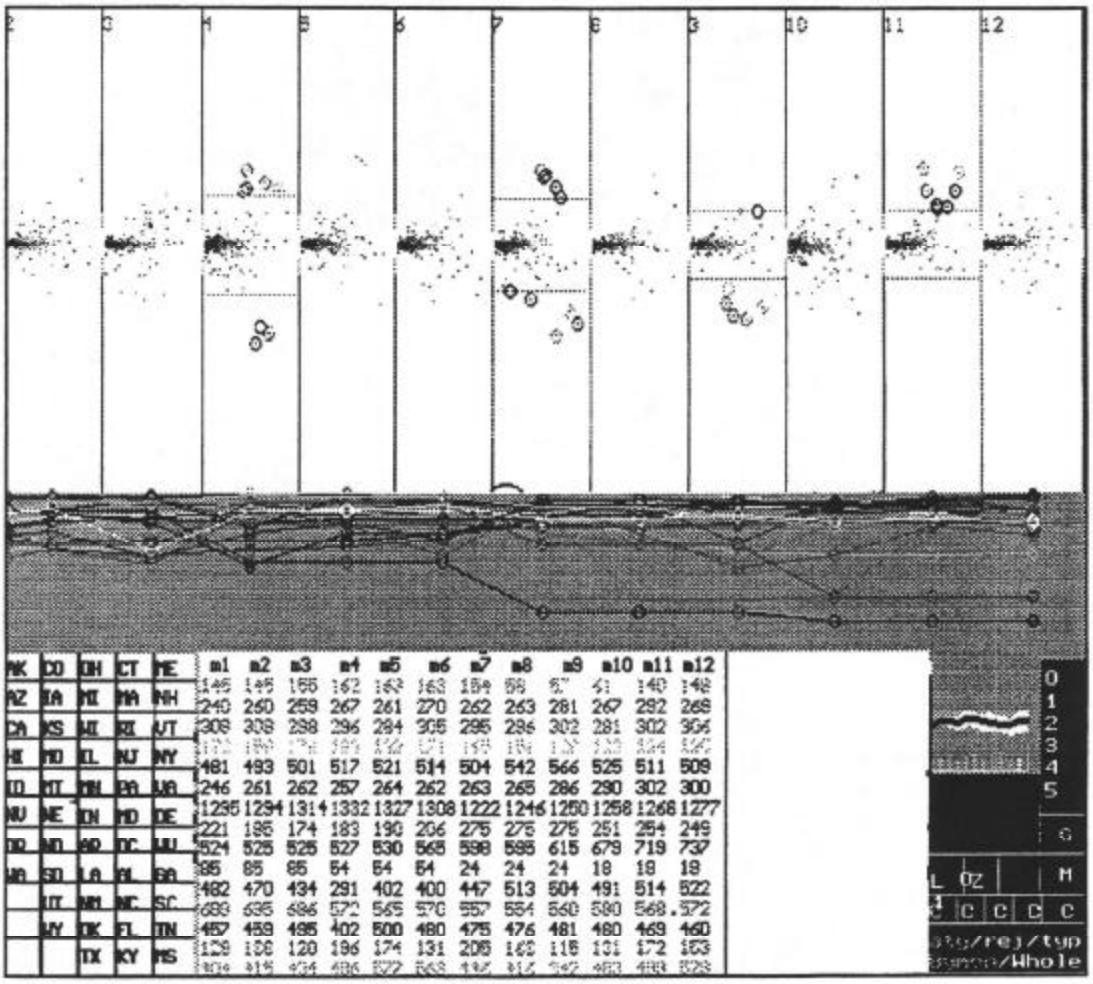
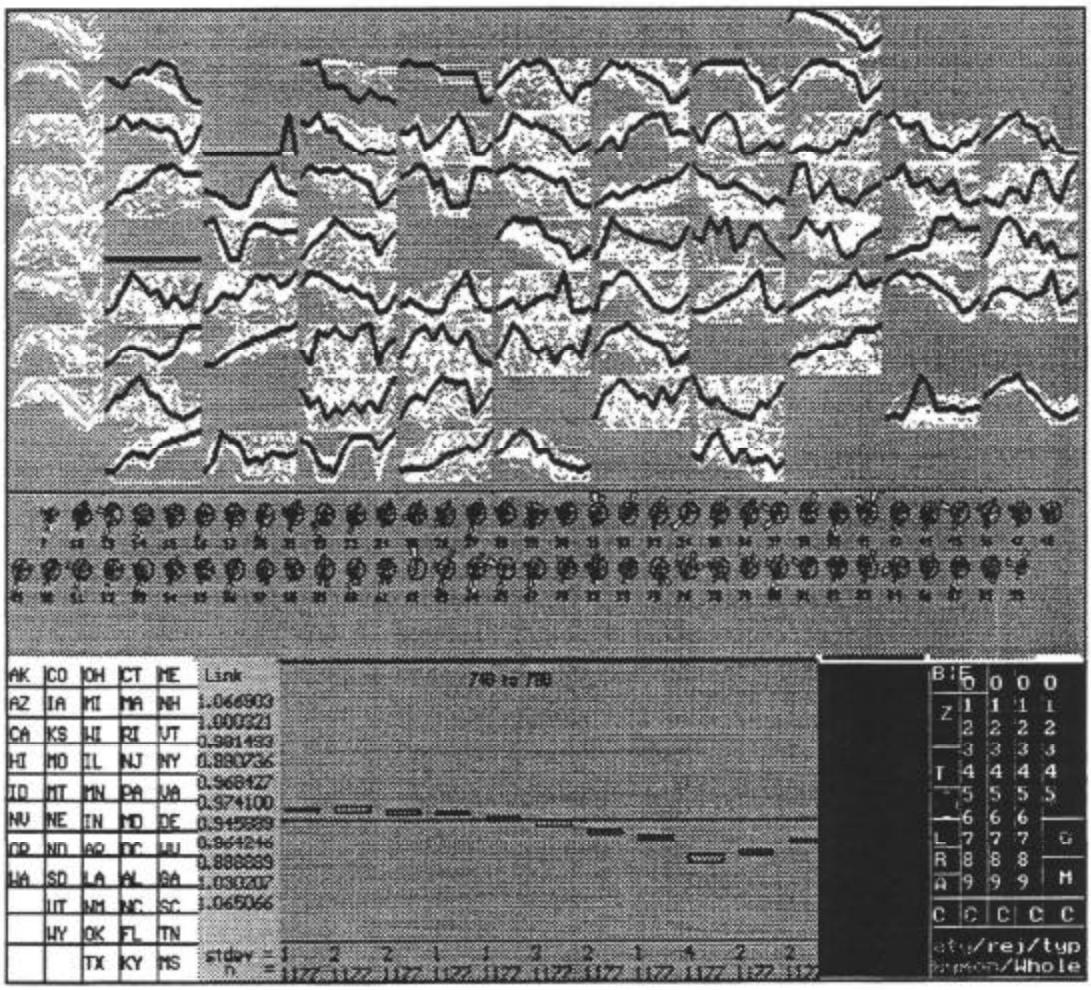


Figure 2 represents a simple adaptation to handle 11 months of a single data variable's sample change from month to month, rather than the six variables of Figure 1; we view more months but viewer variables. In this case, however, we show universe data rather than sample data, so this prototype is suitable for viewing and editing the frame.

Figure 3 moves the whole process in the direction of sampling, editing and estimation research. Figure 3 is an option of Figure 2. The middle row contains reduced size 2-digit SIC daisies for the 12 months of universe data for the industry employment variable, and the daisies work analogously to those of Figure 1. The top window of the screen graphically represents the results of selecting 30 random samples from the universe data for each 2-digit SIC industry, and computing the 30 independent time-series of monthly estimates for each 2-digit industry. The actual universe time-series is shown as the solid black line, with each distinct sample-based time-series as a white line. One can immediately see how well the estimates fared for each 2-digit industry. The left-most column represents the sum (i.e., the 1-digit SIC estimates) of each row.

Figure 3.--Estimates for 30 Random Samples as White Time-series, Universe as Solid Black



This type of comparison between estimates and universe can of course only be done once one has universe data, which in our program are fortunately available some months after we produce our sample-based estimates. However, with the capability to instantaneously select many samples, compute estimates and visually (and also numerically) portray the results, we can test different editing practices, different sampling techniques, and different stratification patterns, to determine an optimum combination of techniques to improve the estimates we produce.

References

Esposito, Richard; Fox, Lin; and Tidemann, Kevin (1994). ARIES: A Visual Path in the Investigation of Statistical Data, *Journal of Computational and Statistical Graphics*, Volume 3, Number 2, 113-125.

Esposito, Richard; Fox, Lin; and Tidemann, Kevin (1994). ARIES: Visual Techniques for Statistical Data Investigation at the Bureau of Labor Statistics, American Statistical Association, *Proceedings of the Section on Statistical Graphics*, 7-10. ■