A POSSIBLE FUTURE DEVELOPMENT FOR THE BLAISE SYSTEM

Gary R. Dunnet
New Zealand Department of Statistics

1. Introduction

In this paper I will suggest that we could improve the editing of data by combining two recent advances in statistics. The first advance is the development of integrated survey processing computer systems, such as the Blaise System. The other advance is the development of simple graphical tools for data analysis, initiated by Tukey's [1] ideas of 'exploratory data analysis' (EDA), coupled with the development of computer technology to display graphics interactively.

Current editing systems, including Blaise, make only limited use of interactive graphics. I believe that interactive graphics will come to have an important place in the editing process. Many of my ideas here are based on the paper "Graphical Editing for Business and Economic Surveys" by Houston and Bruce [2].

My paper explores the importance of interactive graphics for editing data, and discusses how it could be integrated into a system such as Blaise. I will first outline the two different types of editing that are part of the editing process, "micro" and "macro" editing. After explaining the importance of graphics, I will explore the possible implementation of interactive graphics within "macro" editing. As support to my discussion I will outline the New Zealand interactive graphical approach. I will conclude with my thoughts on both where interactive graphical editing could fit into the Blaise System and how the survey cycle may be modified to accommodate this new approach towards "macro" editing.

2. The editing system

Statistical data collection is a complex process involving many stages, in which data on persons, households and businesses are collected by means of surveys and are then transformed into useful statistics. One stage is the process of data editing. The data editing process of any survey processing system may typically consist of :

1. Data entry and input editing,

2. Record Imputation, or Weighting Adjustments,

3. Output Editing & Estimation,

4. Analysis of results.

There is a clear distinction between the editing methods used in the various stages. At the first stage, "micro" editing is used where individual records are checked. The latter stages use "macro" editing involving a complete or near complete dataset. While individual records may be examined in the "macro" phase, the editing is done in the context of the rest of the data.

The philosophy behind editing at the data entry phase is simple : eliminating problems at the beginning of the data collection process reduces problems throughout the survey processing cycle. While any Editing System that follows this philosophy, including the Blaise System (Pierzchala [3]), does give better data quality at lower costs, it does not alleviate the need for editing at other stages. It has been Granquist's [4] and our experiences that "micro-editing may not always detect even serious errors", and that "over-editing" is likely to occur. Granquist [4] also found that "the macro-editing concept is a realistic alternative or complement to micro-editing methods, and can be applied during the processing of the data under the same conditions as computer-assisted micro-editing methods, which reduces the manual verifying work to a considerable extent". I do believe that "macro" editing may be used to complement "micro" editing, and I stress complement, by no means

should an Editing System be solely based on "macro" editing. I believe that any enhancement to the Blaise Editing System, presently solely "micro" editing based, should be directed at the introduction of "macro" editing.

The approach taken to "macro" editing at the New Zealand Department of Statistics is to use interactive graphics. The reasons for the use of graphics is outlined in the following section, after which I will outline how Graphical Editing may be implemented inside "macro" editing. This will lead back to the earlier statement, that I believe that a Graphical Editing approach to "macro" editing may be used to complement the Blaise Editing System.

2.1. The Importance of Graphics

The role of "display" in data analysis is of extreme importance. Tukey [1] stated that "Graphics force us to note the unexpected; nothing could be more important".

Just as the availability of powerful mathematical tools has increased, so has the ability to perform arithmetic computations quickly and efficiently. Graphic techniques are useful to communicate the results of these arithmetic computations to the user. This leaves the user to think about what other things might be done and how results might be interpreted. With the increased power of graphics the process of interpretation is greatly enhanced.

Exploratory data-analysis is necessarily an iterative process, where many "tentative" analysis are performed, promising leads are followed up, some analysis discarded, and where strong, stable and believable patterns are desired. Although "pen and paper" may be used in this analysis it soon becomes apparent, as the number of arithmetic computations becomes sufficiently large, that extensive computerization of the effort is essential.

3. Graphical Editing within the Editing System

As outlined in section 2, the Editing System can be broken down into both "micro" and "macro" editing. This section will outline how Graphical Editing may be implemented inside "macro" editing, which is the approach taken at the New Zealand Department of Statistics. Although Graphical Editing is potentially useful at all stages of the survey processing cycle, I do not see any advantages in using graphics in place of the Blaise System during the "micro" editing stage of the survey processing system.

It should be noted that Graphical Editing can be classified into two modes of use : a "rigid" mode, controlled by menus and buttons, and a flexible "EDA" style, in which different types of plots can be created. The "rigid" mode uses a relatively small number of predetermined plots to uncover certain types of potential errors or problems. The New Zealand Department of Statistics approach is that of a "rigid" mode, but may (depending on the users requirements) incorporate "EDA" ideas at a later date by making it easy to bring data into suitable statistical packages. I will now give a detailed discussion on how Graphical Editing may fit into the three stages of "macro" editing outlined earlier.

3.1. Record imputation

After a sufficient quantity of data has been collected, any non-responses may be imputed using a statistical model. The statistical model may either be survey specific or from a pool of pre-defined imputation methods. Graphical methods would be used in the evaluation of these imputation methods. For example, graphical representations of the distribution of the imputed values may be produced. This sort of analysis would reveal whether imputed values have any unusual or unexpected properties. This type of graphical representation uses the "rigid" mode of editing.

3.2. Output editing

For Business surveys, output editing is the editing of data with the aim to validate the estimates (including the imputations), produced. To date the New Zealand Department of Statistics work in the area of output editing has been concerned with Business surveys, in the near future we hope to examine the area of output editing within Household surveys.

Output editing techniques serve as a backup to input editing techniques. With a reliable output editing system a more focused input editing effort can be achieved. By spending less time on errors which have little or no effect on the final estimates, more time is available to follow-up respondents who have a very large influence.

Occasionally, a sampling unit returns a value which is much larger than those expected at the time of sample selection, leading to an abnormally large weight. Such a sampling unit can have a large volatile effect on the final estimates. One method for handling the problem of large unexpected results is with the use of a robust estimator. Because the robust estimator adjustment may introduce bias, there is still a need for good output editing support. In the paper "Robust estimation and diagnostics for repeated sample surveys" Bruce [5] indicated both the power and relative ease of graphical editing at this stage. Bruce concluded that a fairly rigid graphical editing system along with robust estimators can handle problem estimation weights.

It may well be desirable to deal with "major" errors during the data collection phase, and leave "minor" errors to an automated batch editing procedure [6]. Although this may make a saving on labour costs, a wholesale-automated editing approach can introduce large biases into the estimates. Avoiding such bias would require ongoing validation, requiring the use of similar diagnostics to those outlined above.

Another aspect of output editing is that it can be used as a check on the editing system itself. For example, deficiencies in the input editing system can be analysed by examining the types of errors which go undetected. It

may also be possible to identify when the input editing procedures are having no useful effect on the final estimates, however these types of problems may require a flexible interactive style.

3.3. Analysis of results

The routine analysis of results prior to publication is important. It is important to know whether unusual features are the result of genuine changes in the economy or are simply a reflection of a problem with the survey design or processing procedures. As outlined in section 2.1, routine availability of powerful graphical tools for exploratory data analysis may lead to the detection of previously unnoticed features.

4. The New Zealand approach

A new approach towards output editing based on interactive graphics and data analysis has been developed by Gary Houston at the New Zealand Department of Statistics. The thrust of our development has been towards the Output Editing stage within the "macro" editing system. This is where we saw the most gains could be made with the introduction of graphical editing techniques. The development of the system is based on the ideas of Bruce [5], Houston, and the survey sections themselves. The program is still under development but is being trialed by the business survey sections.

A general overview of the Editing System can be gained from the paper of Houston and Bruce [2]. A brief summary of the advantages of such a system is :

- The use of graphics allows a large amount of information to be displayed at once. "Multiple views" of the data can be simultaneously displayed.

- The use of a mouse allows the user to quickly move through the data identifying any points of concern.

- Buttons and menus can be used to control the user interaction, guiding the user towards consistent application of editing rules. At the same time, options can be provided to permit exploration and investigation.

- "Linked plots" showing the effects of the data value on the estimate can be graphically displayed simultaneously with the plot of the data.

- Historical data of panel surveys is used to show relationships across time as opposed to across variables.

- The system is general purpose.

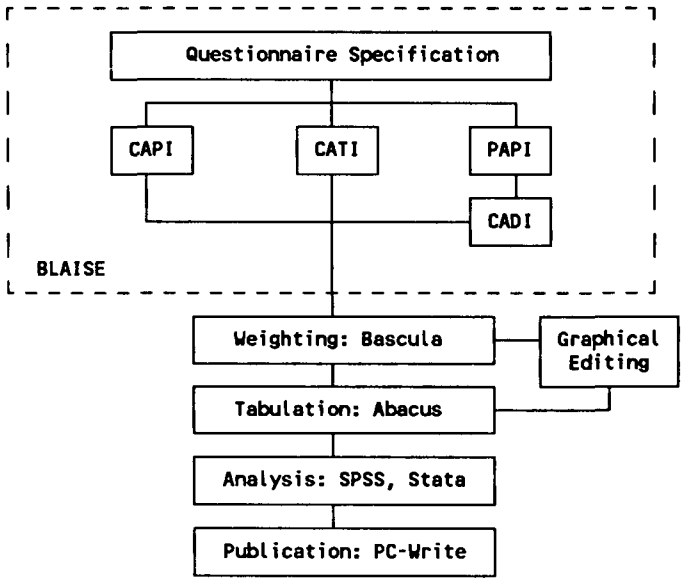
The current system runs on Sun workstations. The data extraction, manipulation and graphics are coded in C (originally Splus) and make use of the X11 window system. It should be noted that the programming involved is not trivial. The system can be applied to reasonably large datasets. It is hoped to construct tools for examining other potential problems, and to examine the usefulness of other types of plots.

5. Graphical editing within the Blaise system

The logical place where Graphical Editing would fit into the current Blaise System is as part of or as an extension to the Estimation System i.e. Bascula. This is the idea in Figure 1. I have deliberately not included graphical editing with Bascula for three reasons. Firstly there may not be a Graphical Editing requirement within the survey process. For example, it is currently unknown whether graphical editing would be useful in a Household Survey. Secondly weighting may not necessarily use Bascula. Bascula allows for post-stratification to correct for non-response, whereas typically for business surveys imputation is used for non-respondents this may use graphics, as outlined in section 3.1. Thirdly I do not believe that the current Blaise System environment can "handle" the graphical requirements. To make full use of Graphical Editing the system may require, say, a 486 with a reasonably advanced operating system,

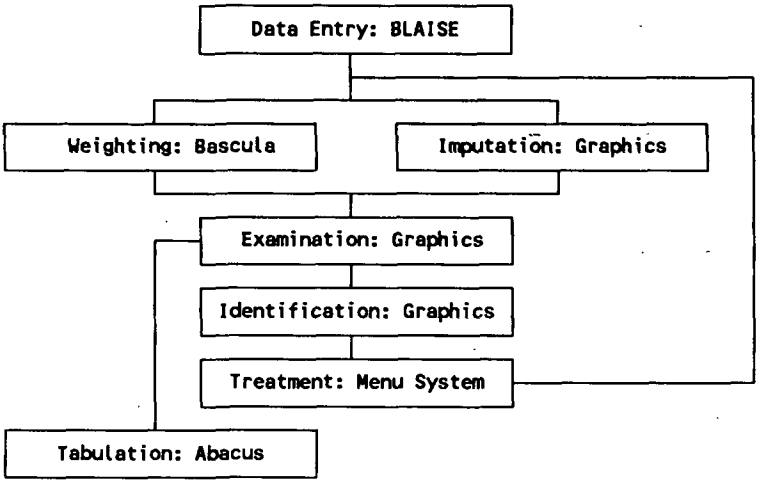
which is solely dedicated to the function of Graphical Editing alternatively, as in New Zealand's approach a Sun workstation, along with Xterminals, linked into the Blaise System.

Figure 1. The Modified Integrated Survey Processing System



It may be more convenient to think of Bascula and Graphical Editing forming a loop with each other. This would lead to a typical Survey Processing Cycle similar to that outlined in Figure 2.

Figure 2. The Modified Survey Processing Cycle



6. Summary

The use of graphics for analysis of survey data shows promise for both survey monitoring and outlier detection. Graphics programs have the potential to develop into a primary tool for "macro" editing.

I believe that the combination of the Blaise System and graphical editing will produce a survey processing cycle with greatly increased data quality and productivity.

References

[1] Tukey, J. W. *Exploratory data analysis*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.

[2] Houston G. and Bruce A. G. *Graphical Editing for Business and Economic Surveys*. Draft paper, submitted to the Journal of Official Statistics 1992.

[3] Pierzchala M. *A Review of the State of the Art in Automated Data Editing and Imputation*. Journal of Official Statistics, Vol 6, No 4, 1990, pp 355-377, Statistics Canada.

[4] Granquist L. *A Review of some Macro-editing Methods for Rationalizing the Editing Process*. Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality, October 1990.

[5] Bruce A. G. *Robust estimation and diagnostics for repeated sample surveys*. Mathematical Statistics Working Paper 1991/1, New Zealand Department of Statistics, Wellington, New Zealand.

[6] Outrata, E. and Chinnappa, N. *General survey function design at Statistics Canada*. Bulletin of the ISI, proceedings of the 47th session, Vol 2, pp 219-238, 1989.