Journal of Official Statistics
Vol. 9. No. 1, 1993, pp. 81–90

gred: Interactive Graphical Editing for Business Surveys

Gary Houston[1] *and Andrew G. Bruce*[2]

Abstract: Despite the prevalence of software for interactive data analysis, current survey processing systems have made limited use of graphics. We believe graphical editing methods show great promise when applied to business and economic surveys. The use of graphics in the context of editing has the potential to improve timeliness and provide greater insight into the data.

At the New Zealand Department of Statistics (NZDoS), we have developed the "*gred*" system for interactive graphical editing of survey data. *gred* is designed to be used in the macroediting stage of the survey processing cycle. While the use of *gred* is still at an early stage, it shows promise for both general survey monitoring by mathematical statisticians and macroediting by subject matter specialists.

Key words: Interactive graphics; macroediting; exploratory data analysis.

1. Introduction

The use of graphics for analysis of data has fundamentally changed the practice of statistics in the past decade. The concept of "exploratory data analysis" (EDA), pioneered by Tukey (1977), has brought forth a collection of graphical tools for gleaning information from a data set (e.g., the boxplot). With the advent of inexpensive modern computer graphics workstations, statistical software systems now offer a powerful and flexible interactive computing environment for analyzing data (Velleman and Pratt 1988; Becker, Chambers, and Wilks 1988; Tierney 1990).

The past decade has also seen impressive advances in the development of generalised survey processing systems, such as BLAISE (Bethlehem and Keller 1989) and GEIS (Whitridge and Kovar 1990). These systems have incorporated new computing technology. Few current editing systems, however, make use of graphical tools.

At the New Zealand Department of Statistics (NZDoS), we have developed the *gred* program for interactive graphical editing of survey data. *gred* is oriented towards the "macroediting" phase in the survey processing cycle. It incorporates many features not present in current editing systems; *gred* is currently being used to edit several business and economic surveys.

[1] Survey Methods Division, Department of Statistics, P.O. Box 2922, Wellington, New Zealand.
[2] Andrew Bruce is currently a research scientist at Statistical Sciences Inc., 1700 Westlake Ave. N., Suite 500, Seattle, WA, 98109, U.S.A.

Acknowledgments: The authors are indebted to Len Cook, whose continuing enthusiasm and support made this project possible, and to many other members of NZDoS who have provided ideas and assistance with implementation. We would also like to thank the JOS editorial board for their helpful comments on the initial version of this paper.

In this paper, we describe some of the capabilities of *gred* and its role in the editing process.

2. Macroediting of Survey Data

Survey processing usually involves two stages of editing: "microediting" and "macroediting" (see Granquist 1984). Microediting and macroediting provide different but complementary approaches towards validating survey data. Microediting occurs at the beginning of the survey processing cycle in the data entry stage, and involves initial screening of individual records with the aim of early detection and correction of gross errors. Macroediting, also called "output editing," comes towards the end of the cycle and is based on checking aggregate statistics. While macroediting can involve examination of individual records, it mainly focuses on detecting errors in groups of records.

Macroediting provides a valuable "top-down" perspective in the validation process. It focuses on detecting errors which affect the published data. Attempting to catch all errors at the microediting stage may not be desirable: many of the suspicious values will have a negligible effect on the final estimates (as shown by various studies, for example, Boucher (1991)). Macroediting may provide a more cost effective way to uncover influential errors.

Macroediting also provides a degree of validation for the survey processing system as a whole. Various problems may only be visible at the macro-level. For example, biases can be introduced through imputation procedures, editing practices, rotation sampling schemes, etc. These biases can only be diagnosed through investigation of aggregate level statistics.

2.1. *Current practice at New Zealand Department of Statistics*

The traditional methods of macroediting used at NZDoS involve one or both of the following techniques:

Detailed listing method: This is a "drill down" approach of identifying suspicious cells in a publication table. Successively more detailed tables are examined in order to determine whether the value is "genuine" or whether it is caused by errors in the data.

Diagnostic method: A diagnostic statistic is calculated for each data value and compared against threshold value. Those values greater than the threshold are flagged as a possible outliers. The diagnostics are typically based on robust measures of location and spread.

There are problems with both of these methods. The main drawback is that they tend to be very time consuming and labour intensive. The drill down approach involves the examination of masses of computer output. In practice, it often evolves into simply exhaustive checking of individual records (i.e., a "drill up" approach).

In theory, use of diagnostics should reduce the manual effort by ensuring that only the important edits are done. However, the thresholds are often difficult to tune, leading to far too many or far too few values to check. Furthermore, some diagnostics have been conceptually too complicated for effective comprehension and use by editing staff. Finally, the diagnostics are never sufficient on their own, and reference to a detailed listing seems inevitable in practice. Hence, use of diagnostics has not led to significant time savings.

The current editing systems have other limitations as well. Outliers may be missed if the effect they have on the estimates is

masked by contributions from other values. Scanning tabular output tends to lead towards biases in the types of unusual values found. Finally, the methods do not put outliers in context: it is not easy to rapidly compare a suspected outlier with other records in the same period and with outliers in previous survey periods.

2.2. *Graphical macroediting*

The use of graphics and exploratory data analysis for problems such as outlier detection is now well established in a variety of application areas. Editing of survey data is no exception: several studies indicate that current macroediting procedures can be improved by incorporating graphical data analysis tools. In particular, Hughes, McDermid, and Linacre (1990) explore the use of non-interactive graphics for the setting of editing bounds, and the detection of outliers is described. They also present ideas for desirable features of an interactive graphical editing system. In a case study of the NZDoS Distribution Quarterly survey, Bruce (1991) demonstrates the importance of exploratory data analysis for validation of survey data. Granquist (1990) describes several macroediting methods, including the "Box Method," in which records are interactively identified for review based on the graphical display of a mathematical function of the data.

3. *gred:* A Program for Interactive Graphical Editing

The *gred* program is a new approach towards macroediting based on interactive graphics and data analysis tools. *gred* has been developed at NZDoS by Gary Houston, arising from ideas put forth in Bruce (1991). While *gred* is still under development, it is currently in operation in NZDoS.

3.1. *An overview of* gred

The use of graphics in an interactive computing environment in the macroediting stage affords a host of advantages. *gred* offers many features not present in the current editing systems.

- Using graphics, a large amount of information can be displayed at once. A single plot can contain both macro and micro information. For example, boxplots of data displayed over time show the trend of all observations as well as identifying outliers over an extended period of time. By contrast, the equivalent information from printouts can be difficult and time consuming to obtain.
- The data values are "mouse-sensitive" allowing the user to quickly and naturally browse through the plot to identify influential and unusual points.
- "Multiple views" of the data can be maintained simultaneously on the display. For example, relevant aggregate statistics can be plotted together with the survey data, allowing the user to view the data from both the macro and micro perspectives.
- Using the concept of "linked plots," outliers and other data values can be simultaneously highlighted in multiple plots. This allows for easy identification of the influence of a specified observation on the aggregate statistic of interest.
- Use of colour and other highlighting techniques extracts key information. Unusual values stand out immediately and can be compared with other values.
- A graphical user interface is used to control the user interaction, guiding the user towards consistent application of editing rules. At the same time, options are provided to permit exploration and investigation.

- The effect of manual (and other) adjustments to estimator weights can be displayed clearly, keeping their use more consistent and objective.
- An abstract data interface is provided so that *gred* can be applied to many different surveys.

3.2. *A typical* gred *session*

To illustrate *gred*, we will run through some typical operations using an artificial data set: the "Not Fruit Survey." This example is based on a prototype version of *gred*. The production version of *gred* is similar, but has an improved graphical user interface.

The constructed data represents a simple stratified random sample with six variables and about $N = 400$ observations. It was generated from actual survey data, with modifications to protect confidentiality. Sampling weights for the observations range from 1 to 100.

Let x_{it} denote the data for the ith firm in time period t. Let w_{it} be the corresponding sampling weight. Then the unadjusted estimate of the total in period t is given by

$$\hat{X}_t = \sum_{i=1}^{N} w_{it} x_{it}.$$

The estimate of the interperiod changes is given by

$$\hat{\Delta}_t = \sum_{i=1}^{N} (w_{it} x_{it} - w_{i,t-1} x_{i,t-1}).$$

To determine influential observations on the estimate of the total $\hat{X}_t$, we need to look at the weighted observations $w_{it} x_{it}$. Similarly, outliers in change are detected by examining the weighted interperiod change $w_{it} x_{it} - w_{i,t-1} x_{i,t-1}$.

3.3. *The canonical display: Loganberries example*

The canonical *gred* screen display consists of boxplots of the weighted survey data and a plot of the corresponding estimates for a particular variable. Figure 1 gives the canonical display from the constructed data for the variable "loganberries." The top plot gives a series of boxplots corresponding to weighted survey data $(w_{it} x_{it})$ for a given period t. Note that the data set displayed consists predominantly of zero observations, so that the "boxes" are actually flat. The boxplot on the right-hand side, separated by a solid vertical line, displays the changes between the most recent periods as a percentage of the total estimate $100\ ((w_{it} x_{it} - w_{i,t-1} x_{i,t-1})/\hat{X}_{t-1})$.

3.4. *Interacting with the display*

Using the mouse it is possible to select any of the points on the boxplots. This causes the program to identify the other points for that firm with a line of a different colour. The firm's reference number is displayed adjacent to the data value which was selected and the sampling weight is displayed at the bottom of the plot. In Figure 1, firm number 20350 was selected by clicking on the top point in the right-hand boxplot. As indicated on the right-hand axis, the increase in the reported value for this firm was almost 10% of the total estimate. As of the December month, the sampling weight for firm 20350 is 1, and the value does not appear to be unusual with respect to other periods. Hence, no action for this firm would be called for.

3.5. *Linked plots*

The solid line in the bottom plot of Figure 1 gives the aggregate statistics for the loganberry data. Selecting a data value in the top plot causes the plot of the aggregate statistics to be updated as well. In the top plot of Figure 1, we highlighted firm 20350. The aggregate statistic excluding firm 20350 is

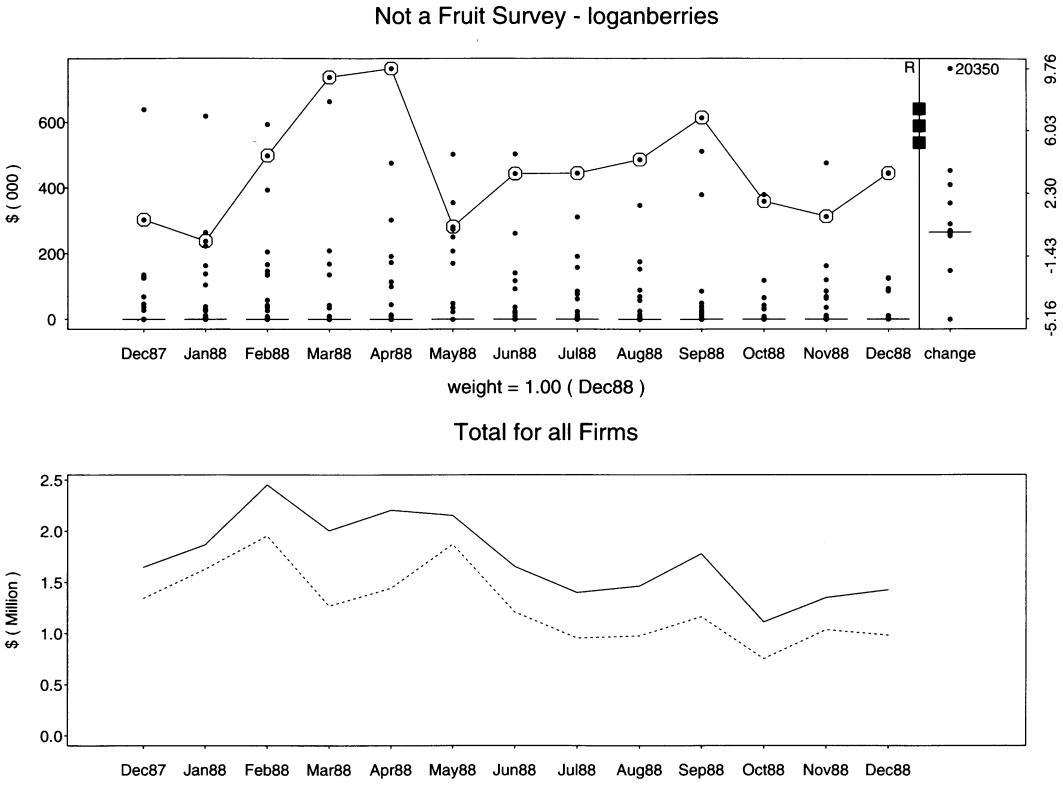

Fig. 1. A typical display from the gred *program. The boxplots correspond to weighted survey data for loganberries. The boxplot on the right-hand side (separated by a solid vertical line from the other boxplots) displays the month to month changes between the most recent periods as a percentage of the total estimate. The data associated with firm 20350 has been highlighted using the mouse. The bottom plot displays the total estimate for loganberries both with and without firm 20350 (solid and dashed line respectively).*

displayed as a dashed line in the bottom plot. As we would expect, this estimate is considerably lower than the original estimate.

3.6. *Controlling the interaction*

The square blobs on the line separating the change boxplot are control buttons. These can be selected with the mouse. Two of these buttons permit stepping through the firms in order of decreasing or increasing effect on the interperiod change. In a typical editing session, one might step through the most significant changes for each variable. The third button ends the editing of the current variable and continues with the next.

3.7. *Reweighting data: The blueberries example*

Figure 2 gives a display for blueberries from the same survey data. Firm 20616 is identified in this plot, and is highly influential. Since it has a very high sampling weight of 100, it is a good candidate for manual reweighting. Using one of the mouse buttons, we can display the firm's data for all

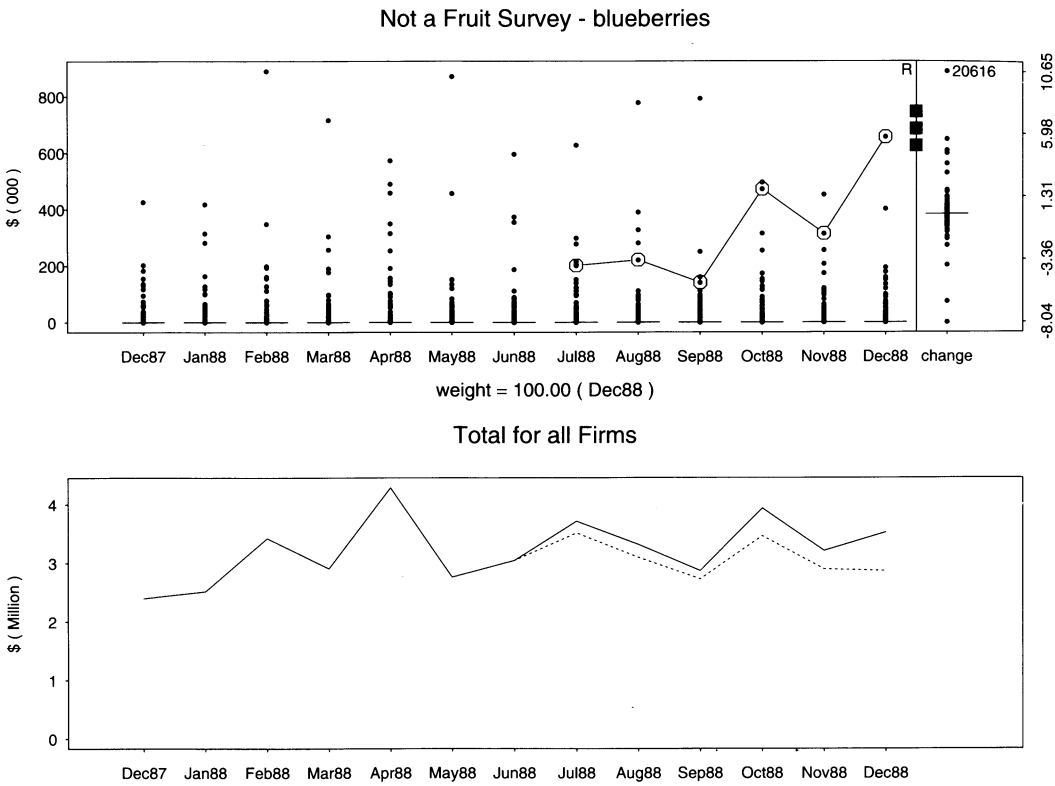

Fig. 2. The gred *display for blueberries. Firm 20616 is highly influential. Since it has a sampling weight of 100, it is a possible candidate for weighting adjustment.*

variables in a separate text window on the screen. Using another mouse button, we can modify the sampling weight to a smaller value (in this case to 10).

Figure 3 shows *gred* display after reweighting firm 20616. In this case, the firm 20616 no longer stands out as an unusually large observation. On screen the reweighted firms will be highlighted with a different colour or symbol. The plots can also be redisplayed with the original weights if desired.

An alternative treatment in this case could have been to temporarily remove the firm from the data set (i.e., assign a weight of 0). This capability is also provided by *gred.*

3.8. *Other features*

gred has a number of other useful options and features. It is possible to some extent to customise the display for particular users and surveys. For example, *gred* can be configured to plot points instead of using boxplots. The boxes in the boxplots of Figures 1 to 3 indicate the location of the median of the data. When we know that the data is very long tailed, as is often the case for business surveys, the boxplot may be an unnecessary complication.

As another example, instead of displaying the relative changes in the right-hand boxplot, *gred* permits users to display the actual changes. These may be easier to interpret by survey staff.

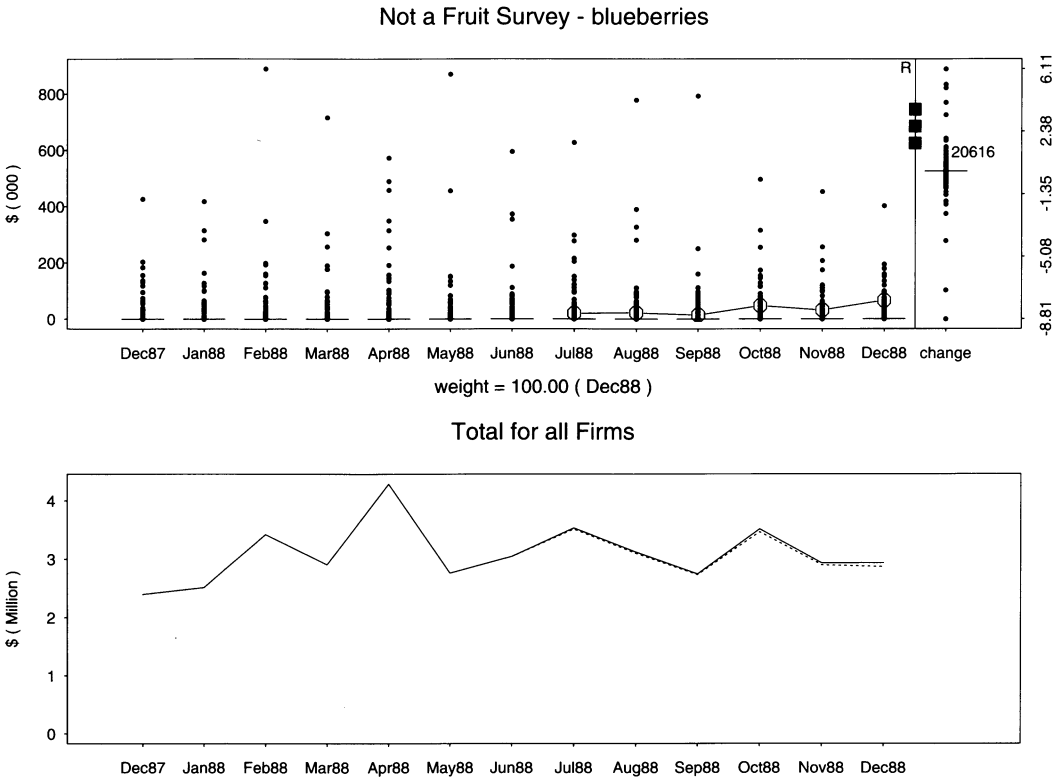

Fig. 3. The gred *display for variable 5 after manually reweighting firm 20616 from 100 to 10.*

Some of the other options built into *gred* include:

- selecting a new variable,
- displaying a sub-group of the data, e.g., a stratum or storetype,
- display of a particular firm, given a reference number,
- displaying multiple firms simultaneously, and
- displaying additional data in text form below the plots.

3.9. *Applications of* gred

Although *gred* is still in the development phase, it is being used by the Survey Methods Division, which comprises mathematical statisticians whose work includes designing surveys, for the monitoring of surveys and identification of firms for weight adjustment. It has also been used by survey staff on a limited trial basis and initial feedback had been very positive. Experience is also being gained through the use of *gred* on a relatively simple survey on a production basis as an alternative to traditional macro-editing methods, and its use in other areas is being considered.

3.10. *Who uses* gred?

As we integrate *gred* into the normal survey processing cycle, we will need to address such issues as who is to use the system, and what training and guidelines are required.

To date our trials have generally involved the heads of the survey sections concerned, who have much subject matter experience

and who are responsible for the quality of the final datasets. These are the people for whom graphical editing is likely to give the greatest gain, by allowing access to the data which would otherwise be buried in large printouts.

The use of *gred* often requires a confidence building period, during which it can be demonstrated that graphical editing does indeed meet the requirements for macroediting, and the reduction of time spent on checking does not lead to missed errors. During this period it is important to note the reservations of the survey staff, particularly regarding deficiencies of *gred* relative to the current editing system.

gred is designed to be used by staff with some knowledge of the data or subject matter. The philosophy of the system is not to attempt to diagnose whether particular firms are unusual. Rather, *gred* presents a powerful data viewing tool allowing the staff to decide which observations are "peculiar" and should be checked. Allowing ultimate control of the checking to remain with the survey staff is seen as important for increasing the acceptance of the new techniques and encourages involvement by staff in improving the quality of the data.

The precise details of how the system is used in practice is likely to vary, depending on factors such as the size and complexity of the survey and the depth of subject matter expertise, as well as the degree of development of other editing systems.

3.11. Current implementation

The current version is written in C and makes use of the X Window System (Massachusetts Institute of Technology) to provide the user interface and graphics. It has been used successfully on Sun SPARC (Sun Microsystem, Inc.) workstations with both colour and monochrome displays, including an ELC with 8MB of RAM memory. Porting to other workstations supporting the X window system should be possible.

3.11.1. Efficiency

The current version of *gred* has been successfully applied to reasonably large surveys. The program allows data to be divided into "groups," for example, retail trade storetypes or regions, and each group can then be graphically edited separately. The number of variables on the data set is not a limiting factor, but the time required when starting up the program and changing variables will increase with the number of observations in a group. The program has been used with 12 periods of data from a survey with 21,500 observations, which could not be divided into groups, but a more reasonable limit for responsive use would be 5,000 observations per group. The speed of the system may be constrained more by the time required to read data from disk files than by CPU speed.

When working with moderately sized data sets (a few thousand observations per group) it is expected that a single Sun workstation with sufficient memory would be able to support several simultaneous users with X-terminals or PC emulators.

3.11.2. Data transfer

Many of the current implementation issues are closely connected with the transition from mainframe computing to LAN based client-server systems. Data transfer between systems has been problematic, and this has slowed the practical application of the system. The changing and more varied technology also presents problems for support of systems and software, when resources and expertise are limited.

Data and meta-data are transfered to the Sun from various microediting systems using a machine-independent file layout. It

will often be necessary to transfer data for previous periods as well as for the current period, to allow for late returns and revisions. To allow effective use of graphical editing it is useful to have as many as 12 or more periods of data stored on the Sun. Meta-data are stored individually for each period, to allow the system to be used when data sets are not completely compatible.

The data on the Sun is generally treated as read-only. If errors are detected they must be noted and corrected using the microediting system. The corrections will then appear the next time the data are transfered from the micro system. An exception to this is that "special treatment" weights can be assigned directly using the graphical editor, in which case the final version of the data must be sent from the Sun to the publication systems. This will also be the case when other survey functions are performed on the Sun, such as imputation of non-respondents using the past data sets.

3.12 Prototyping versus production

The prototype version of *gred* was based on the S system for data analysis and statistics (Becker et al. 1988). As S offers a powerful and flexible programming environment, it makes a good prototyping language for graphics applications. However, S proved too slow and consumed too much memory for use as the basis of a production system with large data sets. Hence, *gred* was reimplemented directly in C.

A rigid graphical editing system (such as *gred*) does not allow the advanced user to expand beyond the confines of the existing system. Hence, its use for applications which were not foreseen during the design of the system would be limited. In an ideal framework, we would tightly integrate *gred* with a powerful interactive data analysis environment. This would deliver a powerful general graphical editing facility, combining an efficient and well focused graphical editor (such as *gred*) with a general data analysis environment (such as S). In such a system, new ideas which arise from general analysis can be rapidly integrated into the graphical editor.

We need a programming environment suitable for general data analysis and prototyping a graphical editor without sacrificing any efficiency. McDonald and Pederson (1988) argue that the Lisp programming language meets these demands with powerful interactive graphics as well as the potential to develop very efficient implementations. Unfortunately, data analysis environments built on top of Lisp are still in their infancy. Several Lisp-based systems, in various stages of development, include Lisp-Stat (Tierney 1990), Arizona (McDonald and Sannella 1991), and Quail (Hurley and Oldford 1991).

3.13. Further developments

In the immediate future we intend to continue the development of *gred* along current lines, with new features added as the needs are identified from surveys using the system. The system is also likely to be applied to more surveys on a production basis.

So far, the graphical editing in *gred* is mostly focused on identifying problems with outliers in the weighted data. We hope to construct tools for examining other potential problems such as the effects of the imputation procedure or rotation sampling plan. In addition, the usefulness of other types of plots may be explored, such as linked scatterplots of pairs of variables or of a single variable across time periods. The ability to produce different types of tables showing unit record and summary data as required on screen is also likely to be a useful feature, as would the ability to attach

notes to unusual data values which have been investigated.

4. Summary

Use of graphics for analysis of survey data within the NZDoS is currently at an early stage. The *gred* program shows promise for both survey monitoring by mathematical statisticians and outlier detection by subject matter specialists. We anticipate that *gred* will develop into a primary tool for macro-editing systems in business and economic surveys.

Development on graphical editing will continue using experience gained from using *gred* as part of the routine survey process. Increasing use of EDA tools in the validation of survey data will generate new ideas and broaden the use of data analysis. The use of graphical editing tools throughout the survey processing cycle will lead to both improvements in data quality and an increase in productivity.

5. References

Becker, R.A., Chambers, J.C., and Wilks, A.R. (1988). The New S Language: A Programming Environment for Data Analysis and Graphics. Wadsworth.

Bethlehem, J.G. and Keller, W.J. (1989). The BLAISE System for Computer Assisted Survey Processing. Bulletin of the ISI, Proceedings of the 47th Session, 2, 239–257.

Boucher, L. (1991). Micro-Editing for the Annual Survey for Manufacturers: What Is Value Added? U.S. Bureau of the Census Annual Research Conference 1991, Proceedings, 765–781.

Bruce, A.G. (1991). Robust Estimation and Diagnostics for Repeated Sample Surveys. Mathematical Statistics Working Paper 1991/1, New Zealand Department of Statistics, Wellington, New Zealand.

Granquist, L. (1984). On the Role of Editing. Statistisk tidskrift, 22, 106–118.

Granquist, L. (1990). A Review of Some Macro-Editing Methods for Rationalizing the Editing Process. Proceedings of Statistics Canada Symposium 90, 225–234.

Hughes, P.J., McDermid, I., and Linacre, S.J. (1990). The Use of Graphical Methods in Editing. In Bureau of the Census Annual Research Conference 1990, Proceedings, 538–550.

Hurley, C. and Oldford, R.W. (1991). A Software Model for Statistical Graphics. In Buja, A. and Tukey, P., (eds.), Computing and Graphics in Statistics, IMA Volumes in Mathematics and Its Applications, 36. New York: Springer–Verlag.

McDonald, J.A. and Pederson, J. (1988). Computing Environments for Data Analysis, Part 3: Programming Environments. SIAM Journal of Scientific and Statistical Computing, 9, 380–400.

McDonald, J.A. and Sannella, M. (1991). Arizona Overview and Notes for Release 0.0. Technical report, University of Washington, Department of Statistics, GN-22, Seattle, WA 98195.

Tierney, L. (1990). Lispstat. New York: John Wiley.

Tukey, J.W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley Co.

Velleman, P. and Pratt, P. (1988). DataDesk Professional 2.0. Odesta Corporation, Northbrook, Illinois.

Whitridge, P. and Kovar, J. (1990). Applications of the Generalized Edit and Imputation System at Statistics Canada. Proceedings of the Section on Survey Research Methods, American Statistical Association, 105–110.

Received March 1992
Revised January 1993