

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

IMPROVING QUALITY BY MODERN EDITING

Submitted by Statistics Sweden¹

Invited paper

Abstract

The paper treats two principal aspects of improving data quality by means of editing: continuous improvement of the whole survey, and the design of edits. The Continuous Improvement Approach is a direct consequence of the new editing paradigm that emphasises identifying and eliminating error sources ahead of cleaning up data. Related issues discussed are collection of data on error causes, the need for and the requirements of a high qualitative Process Data Subsystem (see section II.2), and standardisation of editing process by developing and implementing Current Best Methods (see section II.3). Design of the query edits is considered a crucial issue in quality improvement. Of particular interest for improving data quality by editing is inlier edits. A useful technique for developing inlier and outlier edits is using Exploratory Data Analysis methods in a graphical environment.

I. INTRODUCTION

1. Traditionally, editing is considered the procedure for detecting, by means of edit rules, and for adjusting, manually or automatically, individual errors in data records resulting from data collection or data capture. It is considered a necessary survey operation because errors in survey data may distort estimates, complicate further processing, and decrease user confidence. It should be noted that here editing is considered a tool also for improving quality beyond cleaning up data.
2. Here we discuss the most common setting of editing: The computer identifies erroneous or suspicious data by means of a large number of edits provided by subject-matter specialists; the flagged records are manually reviewed, very often by follow-ups with respondents, (see Granquist and Kovar 1997 for details).
3. Practically all published studies of traditional editing processes indicate that many originally reported values are being changed by insignificant amounts and that few errors are responsible for the majority of the total change. The studies present data like: 10 to 15 percent of the changes contribute to more than 90 percent of the total change; 5 to 10 percent of the changes bring the estimate within 1 percent of the final

¹ Prepared by Per Engström and Leopold Granquist.

estimate. The hit-rate (the proportion of flags that result in changes) lies between 20-30 percent in the few studies where hit-rates are estimated. These facts suggest two things: 1) the entire set of edits should be designed to identify errors more efficiently; 2) many errors could be left unattended or subject to automatic treatments. Many statistical agencies are aware of these problems and devote considerable efforts to raise the productivity of editing systems.

I.1. Increasing Productivity

4. During the last decade a number of selective editing methods have been developed, that can decrease the number of unnecessary flags and order the errors (or the suspect data) with respect to their (potential) impact on estimates either prior to or during survey processing, without having examined all the cases. Selective editing includes any approach which focuses the editor's attention on only a subset of the potentially erroneous micro-data items that would be identified by traditional editing methods. Those methods are known as macroediting, aggregate editing, top down editing, graphical editing, Hidioglou-Berthelot bounds, score function editing (identifying data records that need to be followed up), and others (see e.g. ECE 1994 and ECE 1997, Hidioglou and Berthelot 1986, van de Pol and Bethlehem 1997). We have now a continuous ongoing research on refining this kind of edits and on new types of edits and editing procedures. Furthermore, interactive editing procedures at the data entry stage or when collecting data have proven to be good ways of rationalising the editing of survey data.

5. It is empirically shown that selective editing methods put together in a system, where rework and recontacts to respondents are minimised can increase productivity by 50 percent and more (Granquist and Kovar 1997). But for improving quality we have to go further!

I.2. Quality issues

6. One quality benefit of selective editing is that over-editing is prevented or essentially limited as the editors' work is directed to influential records and items in priority order. Further quality improvements will depend on how the gains in productivity are invested. If the editors are not given more time and higher capacity for solving questionable data further quality improvements should not be expected.

7. Many authors claim that suggested selective editing methods improve quality, but do not provide data to show it. Note that the suggested selective editing methods are evaluated only against the current editing. Thus concerning quality, it can only be stated that the new method is practically as good as the current one, and in data where the new method was evaluated. Certainly, the new method will detect erroneous data that the current method misses, but we do not know whether quality is significantly improved or even improved.

8. Suppose for example that reported data contain both negative and positive errors. Then the edits have to address both types of errors equally. Otherwise, a bias will be introduced by the edits, irrespective of the skill of the reviewers in finding accurate data to replace the flagged data. Thus, edits still play a crucial and determining role for improving quality. Note that generally selective editing primarily means that priority is built into the edits to make them focus on influential items or records. Therefore, programmed edits have to be continuously evaluated and improved to get them to identify all erroneous data that have impact on quality. Special attention has to be paid to parameter based programmed edits. They depend heavily on how well the parameters are estimated and whether underlying assumptions really hold in the data to be edited. Generally, the populations we are surveying are subject to dynamic changes, why we cannot rely on edits and edit bounds used in passed periods. We will return to this issue in the next Chapter.

II. CONTINUOUS IMPROVEMENT - A PROCESS PERSPECTIVE

9. A new editing paradigm or wider definition on editing has been suggested. It is focused on identifying and collecting data on errors, problem areas, and error causes to provide a basis for a continuous improvement of the whole survey.

10. The key objective of the new paradigm is that quality should be built into the processes to prevent errors rather than identify errors once they have occurred and replace them with more accurate data. This is advocated in recent editing literature, for example in Granquist (1995), Jong (1996), Granquist and Kovar (1997), Weir (1997), and Nordbotten (1998). A successful way of doing it is to apply the concept of continuous quality improvement to the whole survey process, where editing is but one process, Linacre (1991). Note that editing under this new paradigm is a key process, in that it will furnish data on errors as a basis for measures to eliminate root causes of errors from the survey. This role of editing will probably improve the quality much more than the editing of data per se. The less error prone the survey process is the higher the resulting quality. It will also reduce the cost of editing substantially provided the editing process is effective in finding and removing the errors that still occur. Thus continuous improvement also applies to the editing process.

11. The process perspective in surveys is described in for instance Linacre (1991), Morgenstein and Marker (1997) and Lyberg et al. (1998). It is based on Deming's Total Quality Management (TQM) principles, in particular Deming's Plan-Do-Check- Act (PDCA) procedure. It implies a shift from mass inspection to controlling the survey processes, because product quality is achieved through process improvement. Lyberg et al. (1998) give an example from experiences at the Research Triangle Institute (RTI), where the coding error rate was reduced by about 75 percent in an application to industry and occupation coding. Costs were reduced and at the same time quality improved.

II.1. Collecting Data on Error Causes

12. The new paradigm imposes a new and probably rather heavy and difficult task to the editors. They have not only to verify flagged data and find acceptable values, but also they have to identify and register quality indications of the new data, error causes, respondent problems, and possible problem areas of the survey. It will require deep subject matter knowledge of and insight into the survey design. Furthermore editors have to understand that this task is substantially more important in recontacting respondents than verifying suspicious data. This contrasts to the common comprehension that flagged data should be changed to pass the edits (creative editing). To build quality into the survey also means that recontacting respondents includes educating the respondents in answering the questions in continuing surveys.

13. Engström (1997) presents a study from the 1995 Swedish European Structure of Earnings Survey (SESES), where data collection on error causes was integrated in the survey process. The editors had to identify and code error causes like misunderstanding, questionnaire problems, typing problems etc. Furthermore, they had to indicate whether respondents were contacted to solve flagged items. Engström found that the edits were rather efficient. The error cause data for the most erroneous item (4000 cases out of 16000) showed that 90 percent of the errors were due to respondent misunderstanding. It was judged that most of these errors could be avoided by changing the wording of the question and improving the instructions to the respondent. However, the coding was burdensome and the editors had problems in referring the error cause to the erroneous item. Engström (1997) concludes that error cause data are extremely useful and that the system has to be carefully designed to facilitate the work of the reviewers.

14. Linacre and Trewin (1989) indicates that rates for item nonresponse and form/system design errors are both about 30 percent of the errors in business surveys and concludes that improving questionnaire design would improve the quality of incoming data. The example given by Engström (1997) emphasises that a significant number of errors can be prevented by improving the questionnaires and that a tight co-operation between questionnaire designers and survey managers would be extremely beneficial for the organisation.

Of course, questionnaires should always be carefully tested. Australian Bureau of Statistics established a forms design unit, when facing the results of a number of evaluations of editing processes as pointed out in Linacre (1991). The cited paper states that the quality of statistical estimates is largely influenced by the respondent's ability to understand questions unambiguously and to have relevant data available. If respondents do not have data for a particular item in their accounting systems, the strategy of collecting data of that variable has to be revised. Note that respondents are likely to deliver the data they have irrespective of any difference in definitions.

II.2. Statistics on the Survey Process

15. In addition to final product quality indicators, the continuous quality improvement approach requires data on the applied editing system architecture as background data and on the performance of the process including interactions with respondents and others to evaluate the process. The editing architecture data are results of the design of the system, while performance data have to be collected and stored during the editing. The product quality, the editing system architecture and the performance data have to be collected and stored in a well-designed system, here called the process data subsystem (PDS). Cost and timeliness constraints particularly for short period surveys exclude post evaluations for this purpose. The data have to be analysed and measures have to be taken to improve the current editing.

16. A PDS has many purposes. Performance measures are needed during the editing process for monitoring and regulating the process while maintaining quality goals, and for improving future system designs as to quality and performance objectives (Weir 1997).

17. The effectiveness of the query edits has to be continuously evaluated, because query edits have expiration dates, although unknown, due to rapid changes in the surveyed populations. For example Hogan (1995) reports that the edit bounds of a used edit did not cover the median. Furthermore, note that (at least initially) bounds are often set on purely subjective grounds. Weir (1997) and Engström (1996) discuss and suggest performance measures and how they can be displayed in user-friendly graphics that easily can be understood by the editors and the survey manager.

18. A PDS should give data on quality for both the user and the survey manager. The users want data about quality to evaluate whether the supplied statistics are suitable for their needs, while the survey manager need data on quality to analyse alternative production strategies (Nordbotten 1998).

19. Editing processes have to be described in a uniform way, making it possible for a statistical agency to compare the effectiveness of the editing between surveys. The top level managers need data in order to allocate their methodological resources, select surveys for revisions, and see the effect of research and development efforts.

20. Key issues associated with a process perspective are definition of key process variables (Morgenstein and Marker 1997), how the measurements should be made to assure that data are collected with high quality, and how statistics should be presented to the different users.

21. Research is needed on designing a PDS beyond as a means for improving individual surveys. The PDS must become an integrated part of a general meta data system permitting research in the origins of errors, improved survey design in general and of improved editing systems in particular. Nordbotten (1998) presents a Process Data Project outline for systematic collection and storing of data on editing architecture, quality and performance for individual surveys which combined with other meta data provide a basis for survey design improvements.

II.3. Standardising Editing Processes

22. Lyberg et al. (1998) state that probably the most effective way to improve quality is to develop Current Best Methods (CBM) for its major recurring processes, to have them implemented and continuously updated as new knowledge is generated. The role of CBMs in the improvement of survey quality is discussed in detail in Morgenstein and Marker (1997). Granquist (1997) presents the development of a CBM on editing, Edit Efficiently, that is used at Statistics Sweden since April 1997.

23. Agency manuals on editing, papers on editing strategies, and generalised software may have similar effects as CBMs in getting sound, recommended practises communicated and used within the agency. The advantage of CBMs is that they are supported by the top level management and developed by the agency's experts together with a number of carefully selected users (here statisticians responsible for editing processes) to assure that each CBM will reflect the organisation's apprehension of what are best practises.

24. A first important step in continuous improvement is to understand the process. Morgenstein and Marker (1997) stress that the staff actually working with the processes under study should use flowcharts to visualise all processes related (in this case) to the editing of the survey. Editing activities are usually carried out in several phases of the survey implying a high risk of lots of rework, see for example Linacre and Trewin (1989). Generally, avoiding rework is one of the most efficient ways of rationalising processes. Thus identifying rework should be important in revising editing processes. At Statistics Sweden process changes are organised as Total Quality Management projects. The CBM on editing (Granquist 1997) has been appreciated by project teams working on improving the editing process. In general, the data editing should be carried out in the early stages of processing, preferably when the respondent is still available (Granquist 1995, Lepp and Linacre 1993). There should always be an output editing subprocess to identify serious errors that were missed in earlier subprocesses. When incoming data have few errors one may rely fully on output editing, in particular when graphical editing is applied (e.g., Granquist 1995, van de Pol et al. 1997).

III. DESIGN OF EDITS

25. We have established that the edits play a crucial role in improving data quality, and that they have to be continuously evaluated in repetitive surveys to control that their good qualities are maintained. As pointed out in the introduction (Chapter 1), edits should not only be considered as constraints on data but powerful means for detecting serious errors. Analysts or methodologists should be involved in the design of edits. The situation has changed the last ten years and suggestions of improved methods appear on conferences and in the editing literature rather frequently. Most of these methods may be characterized as outlier programmed edits, that is data points outside acceptance regions are flagged as suspicious. Here we will only discuss graphical editing.

III.1. Graphical editing

26. Graphical editing is established as a powerful and efficient method for accurately identifying outliers. It is a parameter free method that allows for efficient examination of large amounts of data at once. The data points are visualised in graphs that are linked to each other which means that marking data points with the mouse will highlight the data points in all the graphs. The editor clicks on the potentially most interesting cases, reviews the up-popping reported values of the selected record, changes the faulty values, and notes the effect on graphs, distributions and so on.

27. Graphical editing systems put in practice are intended for output editing. They use top-down approaches in a macro-editing setting to search for probable contributors to suspect estimates (e.g., Esposito et al. 1994, Houston and Bruce 1992, Weir 1997, Engström and Ängsved, 1997). This means that

the search is driven by programmed edits. Hence the quality improvement depends on the edits like other selective outlier editing methods.

28. Explorative Data Analysis (EDA) methods can now be successfully used to edit data on-line, thanks to the advent of powerful PCs and EDA software. EDA can be described as “a set of tools for finding what we might have otherwise missed” in a set of data (see Tukey 1977). Hogan (1995) characterises EDA techniques by the “4 R’s,” resistance, re-expression, residuals and graphical revelation. Biennias et al. (1997) illustrate with practical examples how they used these “4 R’s” to improve the outlier detection in two establishment surveys. The paper can be considered an introduction to using EDA methods in editing. It points out the following advantages. The methods of fitting data to be relatively resistant to the presence of outliers in the data are useful. Removing linearity in the scatter plots to examine the residuals from the linear fit is valuable. Various methods exist to transform data (re-expressing) so that patterns can be more easily discerned. Establishment survey data are often skewed. Therefore, it is essential to have methods for transforming data to be rather symmetric which make it easier to find data points that are particularly unusual.

29. DesJardins (1997) is an excellent introduction to graphical EDA techniques using SAS/INSIGHT or JMP and the implementation strategy used at the U.S. Bureau of the Census. He describes a number of EDA-techniques and presents a number of new EDA graph techniques developed by him to fulfil needs at the Census Bureau. Furthermore the paper stresses the power and the usefulness of the point and click SAS/INSIGHT software making it possible to produce a number of linked EDA graphs of millions of data in a few seconds. Novices in SAS can do it. However, the editors have to be trained in EDA to be able to understand what graphs are telling. It means that analysts should do the editing.

III.2. Inlier edits

30. Outlier checks cannot identify for example data that are affected by small but systematic errors reported consistently by a number of respondents in repeated surveys. Such errors are termed inliers and are defined as faulty data which lie within any reasonable bounds of ordinary used edits. Inlier methods are probably of greater importance for the quality than outlier methods, irrespective how efficient they are in detecting erroneous outliers. Inliers occur whenever there are discrepancies between the survey definitions and the definitions used in the firms’ accounting systems. Werking et al. (1988) present an illustrative example. In an ongoing Response Analysis Survey (RAS) designed to focus on differences in definitions and how to get firms to apply the survey’s definitions, they found that the estimate of the main item “production worker earnings” for the RAS units became 10.7 (standard error 3.2) in contrast with 1.6 for the control group. A method to cope with that type of inliers is to add some questions into the questionnaire, asking the respondent whether he or she included/excluded certain item components in the answer.

31. Mazur (1990) reports from the Livestock Slaughter Data Survey at the U.S. Department of Agriculture, where many respondents reported the same values every week. Therefore they created an inlier edit by using historical data and Tukey’s biweight (see also Hoaglin et al. 1983).

32. Research on inlier methods is fairly new. Winkler (1997) presents a description of the problem and suggests a number of methods of converting inliers to outliers using additional information that may be available in the files being edited.

33. However, Exploratory Data Analysis (EDA) methods using SAS/INSIGHT or JMP is probably the best method for identifying the presence of inlier problems as they are focused on discerning patterns in data.

References

- DesJardins, D. (1997): Experiences With Introducing New Graphical Techniques for the Analysis of Census Data, Work Session on Statistical Data Editing, Prague, Working Paper No. 19.
- Economic Commission for Europe, (1994): Statistical Data Editing: Methods and Techniques, Volume No. 1, Statistical Standards and Studies - No. 44, United Nations New York and Geneva, 1994.
- Economic Commission for Europe, (1997): Statistical Data Editing: Methods and Techniques, Volume No. 2, Statistical Standards and Studies - No. 48, United Nations New York and Geneva, 1997.
- Esposito, R., Fox, J. K., Lin, D. Y., and Tidemann, K. (1994), "ARIES -- A Visual Patch in the Investigation of Statistical Data," *Journal of Computational and Graphical Statistics*, **3**, pp. 113-125.
- Engström, P. (1995): A study on using selective editing in the Swedish survey on wages and employment in industry, ECE, Work Session on Statistical Data Editing, Athens, Room Paper No. 11.
- Engström, P. (1996): Monitoring the Editing Process, Work Session on Statistical Data Editing, Working Paper No. 9, Voorburg 1996.
- Engström, P. (1997): A Small Study on Using Editing Process Data for Evaluation of the European Structure of Earnings Survey, ECE, Work Session on Statistical Data Editing, Prague, Working Paper No. 19.
- Engström, P and Ängsved, C. (1997): A Description of a Graphical Macro-Editing Application, in ECE, Statistical Standards and Studies, No. 48, Geneva pp. 92-95.
- Granquist, L. and Kovar, J.G. (1997): Editing of Survey Data: How much is enough? in *Survey Measurement and Process Quality*, New York: Wiley, pp. 415-435.
- Granquist, L. (1997): On the CBM-document: Edit Efficiently, ECE, Work Session on Statistical Data Editing, Prague, Working Paper No. 30.
- Hidiroglou, M. A., and Berthelot J.-M. (1986): Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, **12**, pp. 73-84.
- Hoaglin, D. C., Mosteller, F., and Tukey, J. F. (1983), *Understanding Robust and Exploratory Data Analysis*, New York: Wiley.
- Hogan, H. (1995): How Exploratory Data Analysis is improving the way we collect business statistics, Proceedings of the American Statistical Association, August 1995 pp. 102 – 107.
- Houston, G. and Bruce, A. G. (1993): gred: Interactive Graphical Editing for Business Surveys, *Journal of Official Statistics*, Vol. 9, No. 1, 1993, pp. 81-90.
- de Jong, W. (1996): Designing a Complete Edit Strategy, Combining Techniques, Statistics Netherlands, Research paper No. 9639.
- Lepp, H., and S. Linacre (1993), "Improving the Efficiency and Effectiveness of Editing in a Statistical Agency," *Bulletin of the International Statistical Institute: Proceedings of the 49th Session*, Florence, Italy, Contributed Papers Book 2, pp. 111-112.
- Linacre, S. J. (1991), "Approaches to Quality Assurance in the Australian Bureau of Statistics Business Surveys," *Bulletin of the International Statistical Institute: Proceedings of the 48th Session*, Cairo, Egypt, Book 2, pp. 297-321.
- Linacre, S. J., and Trewin, D. J. (1989), "Evaluation of Errors and Appropriate Resource Allocation in Economic Collections," *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 197-209.
- Lyberg, L., Biemer, P., and Japac, L. (1998): Quality Improvement in Surveys – A Process Perspective, Proceedings of American Statistical Association, Dallas 1998, to appear.
- Mazur, C. (1990): Statistical Edit System for Livestock Slaughter Data, Staff Research Report No. SRB-90-01, Washington, DC: U.S. Department of Agriculture.
- Morganstein, D. and Marker, D. A. (1997): Continuous Quality Improvement in Statistical Agencies in *Survey Measurement and Process Quality*, New York: Wiley, pp. 475-500.
- Nordbotten, S. (1998): Improving Editing Strategies, Proceedings of the third International Conference on Methodological Issues in Official Statistics in Stockholm, October 1998 to appear.
- van de Pol, F. and Bethlehem, J. (1997): Data editing perspectives, Statistical Journal of the United Nations ECE 14 (1997) pp. 153-171.

- van de Pol, F., Buijs, A., van der Horst, G., and de Wal, T. (1997): Integrating Automatic Editing, Computer-Assisted Editing and Graphical Macro-Editing, Statistics Netherlands, Research paper No. 9739.
- Tukey, J.L., (1977) *Exploratory Data Analysis*, Addison-Wesley Publishing Company, 1977.
- Weir, P. (1997): Data Editing Performance Measures, ECE, Work Session on Statistical Data Editing, Prague, Working Paper No. 38.
- Weir, P., Emery, R., and Walker, J. (1997): The Graphical Editing Analysis Query System in ECE, Statistical Standards and Studies, No. 48, Geneva pp. 96-104.
- Werking, G., Tupek, A. and Clayton, R. (1988): CATI and Touchtone Self-Response Applications for Establishment Surveys, *Journal of Official Statistics*, **4**, pp. 349-362.
- Winkler, W. (1997): Problems With Inliers, ECE, Work Session on Statistical Data Editing, Prague, Working Paper No. 22.