

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 5 of the provisional agenda

ON THE CURRENT BEST METHODS DOCUMENT: EDIT EFFICIENTLY

Submitted by Statistics Sweden¹

¹ Prepared by Leopold Granquist.

I. INTRODUCTION

1. In 1994, Statistics Sweden adopted the principle of continuous quality improvement to be applied in all processes of the agency. The decision was based on information from books, papers, courses and conferences on Total Quality Management (TQM), and studies of experiences from statistical agencies, that had incorporated or made attempts to introduce TQM principles into their work. The U.S. private statistical agency WESTAT was finally chosen to educate and train, first the top managers and then employees in their TQM concept. Since then, around 70 *pilots* have been trained to supervise TQM teams in projects. By summer 1997, more than 100 projects have been launched. About 70 have finished their work resulting in reductions of costs, improved timeliness, and a deeper insight and knowledge of the production process. In spring 1997, the General Director stated that the approach of systematic continuous quality improvement is incorporated into many of the operations. In an inquiry to the staff, 60 percent of the employees declared that they consider TQM very important or important to Statistics Sweden.

2. A key element of the WESTAT view of continuous quality improvement is to look at the statistical agency "as a system of processes that produce a final product, the survey, data registries, or other products" (Morganstein and Marker, 1997 p. 475). This approach contrasts with the traditional quality control view that focuses on "measuring the quality with little recognition that better control of the processes will improve the result."

3. One important method in their TQM concept is developing and using current best methods (CBM) for vital processes. The aim of CBMs is to help minimise variation and increase the likelihood that best practises are followed. Thus CBMs should contain checklists that for repetitive processes must be followed in exact sequence of the operation, while for creative and unique complex processes as designing an editing process, they serve as reminders of factors that have to be addressed. In both cases the CBMs have to be updated as soon as better practices will become known. See Morganstein and Marker (1997) for details.

4. This paper concerns the CBM, developed to serve as guidelines for designing efficient editing processes especially for all business surveys for which Statistics Sweden is responsible. Besides the discussion of the CBM concept, some highlights of the CBM document: "Edit Efficiently" are presented. Note that the document "Edit Efficiently" is not a manual. In principle it defines factors that should be addressed, not how to address them, although examples of successful methods are provided. The paper should be considered an attempt to realise the new view on editing as expressed in Granquist (1996) and Granquist and Kovar (1997). During the course of the project the WESTAT consultants have regularly provided valuable advice on how to shape the CBM document and how to work in the CBM team.

II. ON THE CONCEPTUAL ASPECTS OF THE CBM EDITING DOCUMENT

5. As creative CBMs should be developed only for a few vital complex procedures requiring a lot of resources, it was logical for the TQM co-ordinator (the Head of the Research & Development Department) and the TQM guidance team (the top managers of the agency) to select editing to be the first CBM project. The second project concerned best practices to

decrease non-response in our surveys. The CBM documents of the two projects appeared in spring 1997. In connection with the presentation of the CBMs to all personnel, a four page booklet of each document was distributed to each employee. The booklet on editing was written by a team of potential users of the CBM. They had previously consulted a final version of the CBM manuscript, and succeeded extremely well in highlighting the essential points of the CBM.

6. The objective of CBMs is that they should be accepted and used throughout the whole organization. A CBM has to reflect the organization's view of current best practices. Therefore, the selection of the team members to elaborate the CBM is a crucial issue for the project to be successful. In our case, the CBM team consisted of one methodologist and/or one survey manager or a subject matter statistician responsible for the editing process of a survey (viz. clients of the CBM) from each of the four departments producing business surveys. The editing expert of the Research & Development Department was appointed project leader.

7. The project team dedicated more than one year (about 20 meetings) to collect information on different survey processes, penetrate and discuss every single part of editing processes, obtain knowledge of editing practices and needs within the agency and, most important of all, to arrive at a common view on editing of business surveys. All project members were committed to communicate the progress of the work to their department. This provided feedback and helped to assure that the ideas and suggestions of the project team should be accepted by key persons within the department. The progress was orally reported in every meeting with the TQM-pilots. A number of seminars were held, and the project itself and its key chapters were thoroughly discussed at the international research conferences organized by Statistics Sweden in 1995 and 1996. Furthermore, the CBM-projects were on the agenda when the top managers discussed the TQM projects. In this way, the results of the work became widely discussed and disseminated. An evidence of this was the demand for copies when the document was printed in March 1997. It was distributed to all methodologists and heads of subject matter divisions (around 200 persons), together with a letter that additional copies could be demanded. By summer 1997, nearly 400 persons have requested a copy. This means that almost half of the employees (1300) have one copy of "Edit Efficiently." So far we have received only positive reactions.

8. However, it is not sufficient that the document is widely disseminated and accepted. The main objective of a CBM is that it will be widely used. This part of the project will start in August 1997. Discussions on how to proceed with this issue have of course already started.

9. The document in its present form should be useful as guidelines for designers of editing processes and for survey staff in TQM teams to re-engineer the editing operations of their survey. It does not yet contain checklists of factors that should be considered when designing, re-engineering or evaluating a process. For this task WESTAT proposed a new team to be selected. Persons who were not involved in the development of the document should form the team together with the core of the "old" CBM team. The checklist document should be tested first internally, and then in selected surveys before it is disseminated.

10. Another objection from WESTAT against the CBM is its length, 50 pages. Well aware of the requirement that a CBM should be short and easy to read, our aim was a 25 page document. One of the reasons that this goal could not be accomplished were claims from

team members to include examples how to address at least key factors. However, this and other problems will be solved by creating an electronic version as a home page in our Intranet. Then the latest version of the CBM will be available to all employees just when they need it. Thus any change in best practices can be communicated immediately to the users. This is even more important as there is a substantial international activity in this field of survey methodology. New methods are developed, and already known methods are continuously refined and improved. Examples, methods and the best available reference of any method will be accessed just by clicking on the name of the example or method. In particular examples of good editing processes are lacking, because the CBM is advocating a new approach which has not yet been fully implemented in any Statistics Sweden survey. Therefore, it is necessary to insert descriptions of surveys applying the new approach into the home page text data base. A comprehensive bibliography on editing literature is also planned to be included with the notion that the Research and Development Department can provide hard copies on request.

11. The main reason why a CBM should be short and essentially consist of checklists is that it should be easy to comment and update. The very meaning of *current* best methods is that they should contain the best methods of today, and as soon as better practices are known, the CBM has to be changed. An electronic version available to all potential users will facilitate the updating of the document at any time, irrespective of how important the change might be. In fact, a proposal was discussed to produce a CBM on editing regularly and date each issue, for example *Edit Efficiently 1998*, but it was disregarded when the home page idea was launched.

III. THE CBM ON EDITING

12. Editing of business surveys was selected to be the first CBM project, because editing accounts for about 20-40 per cent of the total cost of a business survey, and many studies show that the cost benefit of this operation is rather low in many cases. Implicitly, editing of surveys on individuals and households is treated as a simplified special case of business surveys, especially as we have only a few surveys with many records.

13. The CBM is focused on the process: the computer identifies errors and suspicious values in individual records. In the next step, the survey staff or respondents (in CASC modes of data collection) verify the erroneous or suspicious item values. It means that the CBM does not specifically provide methods or guidelines for automatic imputation of erroneous or missing values.

14. For the forthcoming Client/Server platform, Statistics Sweden does not have generalized editing and imputation software that could be recommended for general use within the agency. Would that have been the case, the CBM might have become a manual on best practices on using such a software. The major part of the chapter on EDP techniques and technical environment deals with editing in a Client/Server environment and the principles of the software that is under development.

15. The intention of the CBM is to change radically the view on editing from a mere error detecting operation to "identifying and collecting data on errors, problem areas and error causes to provide a basis for a continuous improvement of the whole survey vehicle" (Granquist 1996) as outlined in Granquist and Kovar (1997). The document deals with the

following fundamental issues for designing efficient editing processes: the role of editing in a well designed data collection and production process; critical (fatal) and suspicious (query) errors; where in the survey process should editing be done; how to design efficient edits; what should be done in the manual verifying process; how to promote continuous improvement; how to evaluate editing processes and edits.

IV. HIGHLIGHTS OF “EDIT EFFICIENTLY”

Introduction

16. It is stated that editing is a necessary operation in statistical data production, as errors always occur in data collection and processing. Editing in this sense is defined as the procedure for identifying and handling errors and outliers in data that are used for producing statistics.

17. In all computerized editing processes in Statistics Sweden, the computer identifies erroneous or suspicious data by means of edit rules, and the subject matter staff review the flagged records to adjust the data. That manual review is called verifying and includes all the operations that have to be undertaken with flagged data. In Statistics Sweden as in all other official statistical agencies all over the world this process has been substantially streamlined with the advent of personal computers. The gains have been invested in attempts to raise the quality by applying more checks and to selecting more forms for follow-up with the respondents. Thus, editing is as expensive as before, and it is unclear whether the efforts have yielded higher quality. However, the response burden has been raised clearly as a consequence of the increased number of recontacts. The CBM emphasizes that this common approach may impose an undue confidence in the quality of the survey among the survey managers, in particular if editing is hiding problem areas instead of highlighting them.

18. It is also stressed that editing may not always be the most efficient method of improving the data quality. For example, resources might have a better pay-off if a considerable part of the resources spent on editing is used on a more intensive follow-up of nonrespondents. To obtain a basis for a good allocation of resources, it is necessary to acquire information about the different error sources of the particular survey. Here editing has its most important role: to furnish the survey manager with data on error sources and problem areas of the survey. This information can then be used for a continuous improvement of incoming data quality.

19. The purpose of Edit Efficiently is twofold: to change the view on editing and hence broaden the scope of the verifying work; to provide efficient methods for identifying errors and outliers in collected data.

20. The corner stone of good editing is:

- high hit-rates of the edits
- good quality of incoming data
- continuous follow-up and improvement of the data collection and processing.

High hit-rates

21. The CBM distinguishes between two types of errors: *critical* and *suspicious* errors. Critical errors are those that can be identified with certainty, accessing only data on the individual unit. Examples are: partial nonresponse, validity errors, consistency errors, and typing errors. For the other type of errors, the computer (that is the checks) can only tell that indicated data are suspicious. The questionable data have to be investigated further to find out whether they are faulty, outliers, or correct. The corresponding checks are termed *critical* and *suspicious* checks, respectively. (For linguistic reasons in the Swedish language, we deviate from the terms, *fatal* and *query* edits, which are used in, for example, Granquist and Kovar 1997). It is clearly said that the given "definitions" should not be considered strict definitions. The terms are introduced to get editing processes focused on the difficult problem: to identify suspicious errors efficiently. It is the suspicious checks that are responsible for the high costs of editing, while traditional editing is well suited for identifying critical errors. (Most of the critical errors can be removed by automatic imputation as done in many editing processes).

22. The term *hit-rate*, that is the share of the number of flags that result in changes to the original data, concerns suspicious checks only, and is introduced because a high hit-rate is of vital importance not only for the efficiency of the editing, but also for the resulting data quality. In fact, necessities for the editing to have a positive effect on quality are:

- High over all hit-rate of the set of suspicious edits to identify important errors;
- The verifying work is fulfilled by trained personnel following established procedures.

23. A common view is that many and tight suspicious checks will guarantee a good quality. Checks can do no harm to the quality. But unfortunately, it is not quite that simple. On the contrary, the design of individual checks to a co-ordinated and well tuned set of checks is a complex, but fundamental task to get an efficient editing system (Granquist 1995). There are too many examples, where big errors have drowned in the huge lot of error messages produced by the computer, errors that were visible even for persons who are not familiar with the survey. Badly designed checks and too many flagged data often lead to *overediting*, among other things resulting in insertion of new errors. In particular, new errors are easily introduced when the verifiers focus too much on getting data through the editing system, and manipulate data to pass the edits, so-called *creative editing* (Granquist 1995). This also may give a wrong impression about the reporting capacity of the respondents (survey problems are hidden instead of highlighted). Every editing process that is not carefully designed and executed can become counter-productive.

24. The CBM emphasises the following three principles for identifying suspicious data (see Granquist 1996):

- limit, co-ordinate and target the checks on the specific error types of the survey;
- provide each individual check with bounds based on statistics from the data to be edited so that the verifying work will be prioritized to data that have most influence on the estimates;
- focus the verifying work to the most influential records for example by using score functions.

25. The principles are discussed in general terms and illustrated with examples from well known surveys and editing cases. Short overviews of methods are presented together with data

from successful implementations or carefully designed and well performed studies. Much effort has been devoted to selecting references to detailed descriptions of methods, references that will be directly available in the home page version and changed as soon as we find still better references.

26. The use of score functions, selective, and graphical editing is recommended. Graphical editing is illustrated by the method developed by Engström and Ängsved (1997) and has been in use now for a couple of years. The document contains descriptions of the Hidirouglou-Berthelot method according to Höglund (1994) and of the DIFF score function according to Latouche and Berthelot (1992). It is stated that these methods should be considered as representative of a number of similar methods suggested or developed by renowned statisticians all over the world.

27. We emphasise that acceptance bounds should not be set so that a certain in advance fixed percentage of the data will be flagged, because the hit-rate then might be too low. Except for graphical methods, there are parameters that have to be determined at least as start values in most check methods. Therefore, there is an obvious user need for good thumb rules for accomplishing acceptable hit-rates and for deciding which of the possible methods should be used in an application. Actually, we have tried to analyse our experiences, but concluded that we need to collect more data. This will be done in the near future and will provide the CBM with that kind of guidance.

Good response quality

28. It is fundamental for the data quality that respondent data be of high quality. Prerequisites to get high quality incoming data are that the respondents:

- understand exactly the question and underlying definitions;
- have data available in their information system;
- understand differences between definitions used in the survey and in their information system; in case of large differences be able to give an acceptable estimate.

It means that:

- the data collection has to be adapted to the respondent's possibilities and conditions;
- the production process generates data on response quality;
- the editing process also contributes to raising the incoming quality.

29. The key to fulfil the conditions is to extend the verifying process to collecting systematically data on respondent problems and causes of errors. It is underlined that this is more important than to ascertain whether a suspicious value is wrong and finding a more accurate value.

30. But editing should improve the response quality, too. That can be done by urging the respondent to perform certain checks when answering error prone questions. However, in mail surveys usually the space for such checks is limited to a small number of self-checks. In the document we strongly recommend using the re-contacts to educate the respondents in the survey definitions, concepts, and questions, thus helping them to give acceptable answers in the future as well. The editor should not limit this help to the flagged data items, but should do his or her best to find out possible other problems the respondent may have in responding to the questions. In new surveys, the requirement of an over-all high hit-rate (see paragraph 22) should be considered

subordinate to the need to get information about how respondents understand the survey variables and their possibilities to deliver acceptable answers. It is also accentuated that the questionnaire and the instructions should repeatedly encourage the respondent to contact the agency, whenever he or she encounters problems of any kind to fill out the questionnaire. For this task, the survey manager has to establish a respondentservice desk, equipped with the survey editors, who should be well trained and knowledgeable.

31. All data and information from these activities should be analysed and used to advantage in sharpen survey concepts and definitions, and to improve the survey vehicle design (Granquist and Kovar 1997).

Continuous follow-up and improvement of the data collection and processing

32. Efficient editing means that the quality improvements are justified by their cost. Checks are always constructed based on assessments that they are needed. But with regards to the cost, it is necessary to follow up the efficiency of each suspicious check, and whether the assessments behind the check are valid. What is the hit-rate? Which errors are found? What is the impact on the estimates? How much is the timeliness affected by the editing? In repetitive surveys the survey managers have to know the answers of these and similar questions. Suspicious checks that induce verifying cost or losses in timeliness simply have to detect errors. But it is not sufficient to do evaluations occasionally. Checks have an expiry date, although unknown, due to for example natural changes in the population. Therefore the properties of the checks have to be followed up continuously, in order to detect essential changes in the population, and whether the assumptions underlying the checks still are valid. The CBM suggests a set of indicators to be automatically computed and graphically displayed each time the editing system is run. The proposal follows the ideas outlined in Engström (1996), and is described and illustrated in a similar way as in his paper. To get used by the survey staff it is an absolute requirement that data be presented in an attractive way that can be easily analysed. Therefore graphics, top-down and Pareto principles are used to draw the attention to those items and edits, which cause most problems and work.

33. The tool that is under development for generating editing applications in a Client/Server environment will generate and store process data automatically in a uniform way, according to a requirement from the CBM team. Thus a general indicator tool can be developed and be easily incorporated in all surveys using the editing software.

Where should editing be performed

34. Editing can be done in various phases of the collection and processing of data, for example:

- when data are collected, *respondent editing*;
- manually before data entry, *manual pre-editing*;
- interactively during data entry, *data entry editing*;
- after data entry on batches of registered forms, *batch-editing*;
- when all data are collected and captured, *output-editing*.

An editing process may consist of various sub-processes. Mail surveys very often contain all 5 processes mentioned above. However, many processes dramatically increase the risk of rework in the sense that data are edited more than once, see Linacre and Trewin (1989). Therefore a main

issue is how to allocate and co-ordinate the possible sub-processes efficiently. In fact, processes and this resource allocation problem have been the subject of many discussions in the CBM project. Much work remains to be done here.

35. So far, the CBM advocates that there is a great potential for cost reductions in taking measures to prevent rework, when editing is carried out in different stages of the survey vehicle. The best or the easiest way to do so is to rely on only one heavy editing process. This main editing should be performed as near as possible to the data collection. It is also recommended always to include output editing to ensure that there are no big errors left in data.

References

Engström, Per (1996), "Monitoring the Editing Process," Work Session on Statistical Data Editing, Working Paper No. 9, Vorburg 1996.

Engström, P. and Ängsved Christer (1997), "A Graphical Macro-Editing Application," Statistical Data Editing: Methods and Techniques, Volume No. 2, United Nations New York and Geneva, 1997 pp. 92-95.

Granquist, L. (1995), "Improving the Traditional Editing Process," in B.G.Cox, D.A.Binder, N.Chinnappa, A.Christianson, M.J.Colledge and P.S.Kott (eds.) *Business Survey Methods*, New York: Wiley, pp. 385-401.

Granquist, L. (1996): "The New View on Editing", Paper presented at the Data Editing Workshop and Exposition, Washington D.C. 22 March 1996, U-promemoria 1996-03 1996-05-24.

Granquist, L. and J.G Kovar (1997), "Editing of Survey Data: How much is enough?" in Survey Measurement and Process Quality L. Lyberg et al. (eds), New York: Wiley, 1997.

Höglund Davila, Eiwor (1994), "Macroediting - The Hidiroglou-Bertelot Method," Statistical Data Editing: Methods and Techniques, Volume No. 1, United Nations New York and Geneva pp 127-137.

Latouche, M., and J.-M. Berthelot (1992), "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys," *Journal of Official Statistics*, 8, pp. 389-440.

Linacre, S. J., and D. J. Trewin (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.

Morganstein D. and D. A. Marker (1997): "Continuous Quality Improvement in Statistical Agencies" in Survey Measurement and Process Quality, L Lyberg et al. (eds), New York: Wiley 1997.