The Role of Error Detection and Correction Systems in Survey Management and Optimisation

Elmar Wein, Federal Statistical Office Germany

Abstract

This contribution contains reflections about the planning, management and optimisation of data editing processes. The focus is set on information needed for the management and optimisation of data editing. The management of data editing requires information about the state of error correction, real time effort and costs. With the help of the information provided, the causes for deviations between the data editing activities planned and those actually performed should be discovered.

In contrast to that the optimisation of data editing sets the focus on the discovery of poor data editing processes and the most effective corrections. To fulfil these tasks the optimisation of data editing particularly needs raw and plausible data and information about the data editing processing.

This information should be delivered by error detection and correction systems – like Blaise.

Keywords

Process management, data editing, management information, error detection and correction systems, management of data editing, optimisation of data editing

1 Introduction

Increasing user demand for statistical results and (continuous) budget cuts sharpen the view for the necessity of survey processes and consumption of resources. Data editing is a primary survey process which consumes up to 40 % of the resources needed for the production of a statistics.[i] This fact induces a need for powerful data editing methods, edp techniques and management methods. While great progress has been achieved in data editing methods and respective software, the development of specific management methods used for data editing has been neglected. Essential for a management of data editing is a solid planning, the ability to "measure" the performance, to compare real (management) data with the planned ones and to manage data editing activities on the basis of the conclusions drawn.

Rapid progress in the area of information technology (IT) and the fact that not all problems can be solved during the performance of data editing require a continuous optimisation of data editing activities and questionnaires. Theses activities need data which should be provided by error detection and correction systems.

For that reason the aims of this contribution are to describe the specific aspects of the management and optimisation of data editing, data required for the management and optimisation of data editing, and the

requirements set for error detection and correction systems concerning the provision of appropriate real data.

The contents of the following paragraphs have not been tested in surveys, however they have been influenced by the discussion of a German task force which develops guidelines for data editing.

2 Reflections about a management of data editing

2.1 Methodological framework

Essential for the management of data editing are:

- the definition of data quality which sets the aims for data editing activities,
- the way in which data editing activities are organised, and
- criteria and data used for the judgement of the efficiency of data editing activities.

User demands for data quality influence data editing. There is no uniform definition of data quality, but with regard to data editing the focus may be put on the commonly used criteria "timeliness", "accuracy" and "clarity, accessibility".[ii] Clarity is an important criterion for data editing because it has to provide in many cases information for user-friendly quality indicators.

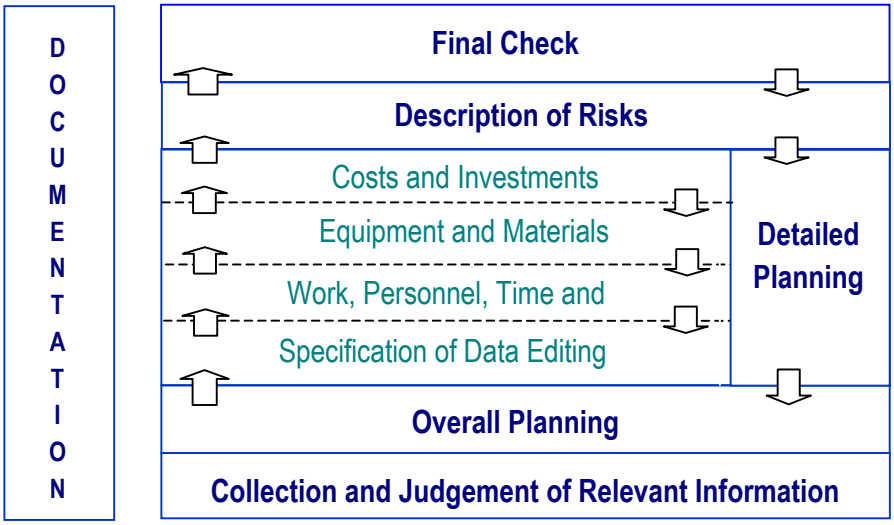
It's a fact that the reorganisation of activities in processes improves the efficiency.[iii] For that reason data editing activities are combined in data editing processes.[iv] A data editing process receives raw statistical data and delivers plausible ones as a result of logically connected activities. The design of a data editing process reflects the individual view of an organiser. Data editing processes are designed in such a way as to contribute to the dissemination of statistical results by short **runtimes**, low consumption of resources measured by **process costs** and a user-friendly documentation which is needed for data analysis. They possess only the absolutely necessary interfaces with other survey processes, and complex data editing processes can be divided into logically separated sub-processes. A process owner is defined for every data editing process who needs information, methodological and subject matter knowledge and adequate (IT) equipment.

The achieved "accuracy" of statistical data, needed runtimes and process costs form essential criteria for the management of data editing processes, measured by *real* data of data editing processes and compared with *planned* data coming from the planning of data editing.

2.2 Planning of data editing processes

It is assumed that there are arrangements between users and producers of statistical results concerning their accuracy. They are transformed into operative criteria for the management of data editing processes. Runtimes and process costs are estimated and calculated during the planning of data editing which may be performed in the following steps:[v]

Figure: The Flow of the Planning of Data Editing



The figure above shows that the planning of data editing consists of the overall planning with the preceding collection and judgement of relevant information. The main aim of this step is the development of a consistent data editing strategy consisting of data editing sub-processes. They are completely specified during the several steps of the following detailed planning. Specifications of data editing activities e.g. OCR, data capture, (manual) coding and checks represent work to be done. For that reason they form the basis for the estimation of process duration - the basis of process costs. With the help of a powerful meta-data system and relevant project management software, process duration can be estimated for typical data editing sub-processes as demonstrated:[vi]

The estimation of the time effort for data capture $\hat{t}_D$ *is based on the average time needed for capturing an attribute* i *of a characteristic* $\hat{\bar{t}}_{Di}$. *It is determined by the length of an attribute, their number and the speed of data capture. If more than one attribute is captured, it will be necessary to weight them with their probabilities of occurrence* $\hat{p}_{Di}$ *and to add the weighted values. If the occurrence of a characteristic* c *depends on routings the respective estimated probability* $\hat{p}_c$ *shall be considered. Based on an estimation of the average time needed for checking the captured data per questionnaire* $\hat{\bar{t}}_Q$ *and the expected number of questionnaires for data capture* $\hat{n}_D$ *the basic time effort can be estimated. It should be generally expanded by time needed for non-specific activities e.g. meetings, training. This will be done by a factor of 15 percent so that the time effort for data capture* $\hat{t}_D$ *can be finally estimated as follows:*

$$\hat{t}_D = 1{,}15 \cdot \hat{n}_D \cdot ((\sum_{c=1}^{C} \hat{p}_c \cdot \sum_{i=1}^{I_c} \hat{p}_{Di} \cdot \hat{\bar{t}}_{Di}) + \hat{\bar{t}}_Q)$$

The main advantage of this procedure is the efficient use of existing metadata of the survey contents and assumed sample size and the use of such a formula with real data. Data editing processes which consist of different activities and the time effort needed for computations can't be estimated in the mentioned way; they should be indicated as total numbers. The disadvantage of such an overall

estimation is that there is no opportunity of finding out the causes of deviations during the management of data editing processes.

The calculation of process duration requires the division of estimated time efforts through the general worktime per day/week/month and the available time of personnel or available hours of computation respectively. The result of this calculation should be multiplied with the set of labor costs of the personnel or computations.

2.3 Management of data editing processes

It is assumed that management activities consist of a periodic observation of the error detection and correction and a kind of stock-taking at critical steps / procedures – preferably at milestones. The periodic observation requires an error report which generally contains accounts and graphics on erroneous data and the influence of corrected errors on data. Furthermore a management report should document the performance of data editing by a comparison of the estimated and needed time, the influence on the beginning and termination of all following processes, a comparison between calculated and real costs and an overview of the errors and correction. In addition to that the report should deliver information about causes of time deviations, e.g. differences in numbers of errors and more or less time effort needed for their correction.

This information required for the reports should be delivered from the planning of data editing and measured during the performance of data editing processes. The proposed formula for the estimation of the time effort opens the opportunity to calculate real time efforts. The basis of this calculation is formed by real numbers, which should be provided by error detection and correction systems. They should also be used for the determination of the causes of deviations and should give ideas about how very effective corrections could be performed.

3 Optimisation of survey processes

Changes in the demand for statistical results, problems during the performance of a survey, methodological and technical progress and changing demands for statistical results are the most important reasons for an optimisation of a survey. In this context the focus will be set on the improvement of data editing processes and questionnaire design. Furthermore recent reflections see in the optimisation of survey processes the chance and necessity to improve the knowledge of the personnel who is involved in the survey operations. The aims of an optimisation may be:

1. the improvement of data quality and efficiency of survey processes.[vii]
   Relevant activities are focussed on performed corrections, the introduction of new data editing methods and process redesign. These activities require information about raw and plausible data and documentations of data processes which should contain information about duration and costs.
2. the reduction of burdens on respondents and the prevention of errors.
   This aim can be achieved by an optimisation of questionnaires which may lead to a reduction of burdens on respondents and of errors as well. Error statistics may provide useful information for the optimisation of questionnaires.

3. the improvement of the knowledge at a national statistical office.
   Knowledge which was acquired during the performance of survey processes, e.g. about the application of data editing methods, may have great influence on methodological guidelines and research which are valuable for colleagues with similar tasks. Error statistics contain important information for methodological research on data editing methods.

The reflections in chapter 2 and 3 demonstrate that the optimisation of data editing sets higher demands on the provision of information.

4 Demands on an error detection and correction system

Error detection and correction systems like Blaise play an important role with respect to the provision of information required for the management and optimisation of data editing. In accordance with the preceding reflections an error detection and correction system should deliver the following information:

Table: Information delivered by an error and detection system

Field	Example	Remarks
Identification number (ID) of a survey	*...*	This information is needed for comparisons between different surveys.
Name of a file	*Stat32*	This information is needed for the reconstruction of implausible data during the optimisation of data editing.
ID of a record	*112*	
Date	*04222001*	It is assumed that there are general management decisions concerning the work effort for error correction activities. The date is therefore necessary to determine the influence of these decisions.
Time	*1503*	The beginning of a plausibility check is needed in combination with the beginning of a following check to determine the work effort.
ID of a plausibility check	*A025*	It is assumed that plausibility checks are combined and integrated in data editing processes. The ID is necessary to calculate the process effort.
Type of correction	*manual / automated*	This information should be handled like a flag which influences the manner in which time efforts should be calculated.
Corrected characteristic	*SocPos[1]*	Implausible data can be reconstructed with this information. It is used for the determination of most effective checks.
Original value of the corrected characteristic	*2*	
Personal ID	*...*	The personal ID is used for the calculation of process costs.

The proposed information needs some additional remarks:

- The information should be provided only on demand and in separate log files. The further data processing should be performed by a separate management system which should have access to the planned data and formulas as described in the previous sections.
- Data editing non specific activities and general management decisions concerning error correction may heavily influence the distance between two points in time recorded. It is assumed that their influence can be eliminated by the calculation of real time efforts for plausibility checks.

5 Conclusions

Higher user demand for statistical results and data quality and continuous budget cuts lead to new requests for data editing methods and management techniques. The implementation of a process management approach for data editing requires more information about the performance of data editing.

The information should be delivered by error detection and correction systems on demand in a structural form and in separate log files. Especially data about points in time, IDs of plausibility checks, personnel, files, records and original values are needed. On their basis time effort for plausibility checks, process duration and costs may be calculated. A management system should generate management reports on the basis of the information provided.

---

[i] Data Editing Subcommittee of the Federal Committee on Statistical Methodology: "Data Editing in Federal Statistical Agencies", Statistical Policy Working Paper No. 18, Statistical Policy Office, Office of Information and Regulation Affairs, Office of Management and Budget, 1990

[ii] Yves Franchet (1998). "Verbesserung der Qualität des ESS". DGINS-Konferenz, Stockholm, 18. Bernard Grais (1998). "The Future of European Social Statistics". Mondorf Seminar, pp. 28-31.

[iii] Bernd W. Wirtz (1996). "Business Process Reengineering – Erfolgsdeterminanten, Probleme und Auswirkungen eines neuen Reorganisationsansatzes". Zeitschrift für betriebswirtschaftliche Forschung, 48, pp. 1023 - 1037

[iv] Manfred Schulte-Zurhausen (1995). "Organisation". München, 41pp. Günter Schmidt (1999). "Methoden des Prozess-Managements". WiSt, 9, pp. 241-245. Verein Deutscher Ingenieure, Deutsche Gesellschaft für Qualität (1998). "Total Quality Management Prozesse". VDI/DGQ 5505 (Entwurf), Düsseldorf, pp. 2-17

[v] Georg A. Winkelhofer (1997). "Methoden für Projektmanagement und Projekte". Berlin, pp. 121-215. Heinrich Keßler, Georg Winkelhofer (1999). "Projektmanagement". Berlin, pp. 162-180.

[vi] Bundesministerium des Innern (1995). "Handbuch für die Personalbedarfsermittlung". Bonn, B-29 pp. Helmut Wittlage (1995). "Personalbedarfsermittlung". München, pp. 24.

[vii] Marco Fortini, Mauro Scanu, Marina Signore (2000). " Use of Indicators from data editing for monitoring the quality of the survey process: the Italian information system for survey documentation (SIDI)". Statistical Journal of the United Nations ECE 17 (2000), pp. 25-35