

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing

(27 – 29 May 2002, Helsinki, Finland)

Topic (i): Planning and management of statistical data editing

THE PLANNING OF DATA EDITING PROCESSES

Invited Paper

Submitted by Destatis, Germany¹

I. INTRODUCTION

1. A powerful contribution to data quality represents an important aim for data editing activities. Modern management techniques promote the organisation and management of (data editing) activities with processes, which leads under these general conditions to the conclusion that optimal data editing processes contribute to a high data quality and a high level of efficiency. Processes have to be planned, so a careful planning of data editing processes seems to be a compelling precondition for a good quality of data editing processes.
2. Users' demands for statistical results, personnel employment, available methods and techniques, and equipment are important parameters of data editing processes, which should be brought into a balance. The respective planning activities seem to be complicated because of the dependencies between the preconditions.
3. Data editing plans are considered to be an essential basis for the management of data editing activities because they are needed for comparisons with performance data¹ on data editing processes. Comparisons will only be helpful if they provide clues to the reasons for deviations between the plans and the performance data. This requirement raises the question what information can be automatically gained from data editing processes.
4. Many national statistical institutes (NSIs) have to compete with other statistical institutes for national surveys and have to offer their services inclusive of data quality and cost agreements. Due to the increasing implementation of business methods in the public sectors of many countries, even ministries have to pay for statistical results. So NSIs need a cost estimate which takes into consideration the effort involved in the statistical production. From this point of view, data editing plans and performance data should deliver information for calculating the prices of statistical results.
5. Against this background the aims of this contribution are:
 - the presentation of considerations on the planning of data editing processes
 - the presentation of some considerations on how to gain performance data from data editing processes, and
 - some conclusions and considerations on further research.

¹Prepared by Elmar Wein (elmar.wein@destatis.de).

II. ASSUMPTIONS CONCERNING THE PLANNING PROCESS

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

7. It is assumed that survey managers plan their data editing activities as described in the contribution "The planning of data editing" presented at the last Work Session on Statistical Data Editing.^{iv} This means that survey managers collect necessary information for the planning of data editing, develop a data editing strategy during a coarse planning, perform a detailed planning, describe and evaluate potential risks and finally check their plans.^v The following considerations are based on the assumptions that a planned data editing strategy is applied and that specifications of the checks are available.

III. DESIGNING OF DATA EDITING PROCESSES

8. The designing of data editing processes then begins with an estimation of the time effort for specified basic data editing activities. Together with the given deadlines this estimation forms the basis for planning the personnel employment and finally the time scheduling. If the need arises the respective steps will be performed several times to take given restrictions into consideration. After the time scheduling and the personnel employment the cost and investment planning can start. If it becomes clear that a planned data editing strategy cannot be financed the plans have to be adapted. The description of risks and final checks of the plans complete the planning. Finally data editing plans are documented by expressive indicators which facilitate a benchmarking within groups of similar surveys.

III.1 Estimation of the time effort for data editing activities

9. The time effort for data editing processes can be calculated on the basis of the time effort for basic activities. An important reason for this approach is that it is easier to measure or estimate the time effort of simple activities instead of complex ones. Another reason for the estimation of the time effort for basic activities is that the deviations between the plans and the performance data become obvious.

10. It is assumed that the time effort involved in computations can be determined more easily than the time effort for human work. Furthermore, detailed information on computer runtimes of performed surveys will facilitate the estimation of time efforts for computations. The computations heavily depend on the software systems applied. So the focus of the further considerations will be on the estimation of the time effort of personnel activities.

11. Some useful advice about the estimation of the time effort for data editing processes is provided by the existing methods concerning the calculation of needed personnel^{vi}, which is generally calculated by the following steps:

- Determination of the basic time which is needed for a basic activity.
- Addition of a contingency allowance for individual and specific activities in relation with a basic activity. The result is the mean time effort.
- The mean time effort is multiplied by the amount of work, i.e. the number of products or orders, so that the total time effort can be computed.

12. A manual check of a questionnaire, an optical character recognition (OCR), the data capture of an answer, its automatic and manual coding, a correction of an error, and a computer runtime generally represent typical basic activities of data editing. They may be considered as quantifiable repetitive work which cannot be subdivided. In principle, the time effort of these activities could be objectively and exactly measured but this measuring requires an extreme effort because the activities have to be simulated under the specific conditions of a survey. As many NSIs don't have the resources for such studies, the estimations to be made should be supplemented by recommendations derived from performance data in the long run.

13. The time effort for a basic activity depends on the individual efficiency of the staff members and their experience especially in the case of coding activities. With respect to the estimation of the time efforts, personnel with average experience and individual efficiency should be assumed. The quality of the estimations can be improved by the following guidelines:

- Estimations should be done for easily understandable activities. In accordance with this recommendation complex data editing processes should be subdivided into basic activities.
- The quality of estimations can be improved if a person simulates the performance of the respective activities or has a clear impression of their performance.
- Experienced personnel and different persons should give their estimates, and the mean estimated time effort should be calculated in case of different estimations.
- Deadlines should not be considered during the estimation because this will lead to the adaptation of the time effort for basic activities.

14. Data editing processes like the cleaning of records generally combine basic activities like data capture, coding and error correction. The time effort for one record is determined by the length of the routes in a questionnaire and the probability of the occurrence of errors. Other activities like OCR are determined by the number of records in data editing processes. One difficulty for the estimation of the time effort for data editing processes arises from the fact that the number of records and errors has to be estimated as well.

15. It is obvious that personnel normally do not spend their whole working hours on correcting errors. Sometimes meetings take place and from time to time personal affairs are taken care of. As these activities can't be measured exactly they should be taken into consideration through a global extra charge in accordance with generally accepted recommendations, i.e. of 15 percent^{vii}.

16. Against this background the time effort of data editing processes like OCR can be estimated as follows: Operators of OCR machines really know the capacities of their equipment to be expressed by $(\hat{t}_S + \hat{t}_{OCR}/\hat{s}_P)$ which means $\hat{t}_S + \hat{t}_{OCR}$ minutes for \hat{s}_P pages with \hat{t}_S as estimated time for scanning and \hat{t}_{OCR} for OCR if it is necessary to separate the time efforts for both sub processes. With \hat{n}_R as the estimated number of questionnaires to be scanned, \hat{s}_Q the number of pages per questionnaire, \hat{r} the expected failure rate concerning the character recognition, and \hat{t}_C the estimated mean time effort for corrections of non-recognised characters per page, the time effort for OCR \hat{t}_R can be estimated as follows:

$$\hat{t}_R = \hat{n}_R \cdot \hat{s}_Q \cdot \left(\frac{\hat{t}_S + \hat{t}_{OCR}}{\hat{s}_P} + \hat{r} \cdot 1,15 \cdot \hat{t}_C \right) \quad ; \quad \hat{s}_P > 0$$

17. 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{t}_{Di} . It is determined by the length of an attribute, the number of attributes and the speed of data capture. Multiple attributes per characteristic will be necessary to be weighted 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 probability \hat{p}_c shall be considered. Based on an

estimation of the average time needed for checking the captured data per questionnaire \hat{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, like personal absence. 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:

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

18. With respect to the estimation of the time effort for the computer-assisted coding two different situations may occur: an information can be coded or not. In the first case the additional time effort for coding can be neglected. In the second case clerks need a certain amount of time for coding which could be estimated by a similar formula. With \hat{n}_O as expected number of records which require a higher time effort for coding, \hat{p}_{Oc} as probability of a content which needs a time effort for coding, and \hat{t}_{Oc} as expected mean time effort for coding a characteristic c the additional time effort for coding \hat{t}_O can be estimated in the following way:

$$\hat{t}_O = 1,15 \cdot \hat{n}_O \cdot \sum_{c=1}^{C_O} \hat{p}_c \cdot \hat{p}_{Oc} \cdot \hat{t}_{Oc} \quad ; \quad C_O \leq C$$

19. The estimation of the time effort for computer-assisted error correction assumes at first the correction of all errors regardless of their influence on statistical results. If not all errors are corrected manually a rate of the total time effort can be estimated and the time for automatic corrections should be estimated by another formula. With \hat{n}_E as expected number of records which form the basis of an error correction, \hat{p}_i as estimated probability for multiple answers, \hat{p}_{Ej} as estimated mean probability of the activation of a check j , and \hat{t}_{Ej} as expected mean time effort for an error correction, the total time effort for corrections \hat{t}_E can be estimated in the following way:

$$\hat{t}_E = 1,15 \cdot \hat{n}_E \cdot \sum_{c=1}^C \hat{p}_c \cdot \sum_{i=1}^{I_c} \hat{p}_i \cdot \sum_{j=1}^{J_i} \hat{p}_{Ej} \cdot \hat{t}_{Ej}$$

The formula may help to estimate the time effort for the correction of all errors. As in practice data cleaning may be finished in case of marginal influence on statistical results a part of the total time effort may be used for the planning activities.

20. The described estimation of the time effort may be suitable for typical data editing processes. In the case of data editing processes with different types of basic activities, the estimation should be performed separately for each type and the specific time efforts should be summed up.

III.2 Personnel employment and time scheduling

21. The next step towards the estimation of process duration is the assignment of personnel to data editing processes. Relevant information in this context is the availability of a person expressed by a starting and a finishing date and the person's working capacity per week - i.e. full time or part time. On the basis of this additional information and the calculated time effort, an IT tool should calculate the process duration.

22. The time scheduling can start after the assignment of personnel capacities to processes. Especially in the case of complex data editing strategies with overlapping processes, commonly used methods like the calculation of earliest starting points for processes from the start to the final point (forward scheduling) and latest starting dates (backward scheduling) should be applied to find out time buffers of processes. Critical processes do not comprise any time buffers; that means they have to be performed just in time.

Finally the time scheduling should discover critical paths, which refers to sequences of data editing processes without any time buffers. Based on this information, an IT tool should document the results by lists and Gantt-Charts, which seem to be suitable means of visualising the sequence of data editing processes even in large surveys.

23. The time scheduling should also encompass the setting of milestones. They should be used to gain an overview of performed data editing processes: this refers to comparisons between planned and performance data with respect to error correction, duration, and costs. Milestones are set in general at the end of critical processes. The first one may be set at the end of the error detection to clarify the influence of deviating record and error numbers on all following activities. Other suitable time points for milestones seem to be deliveries of clean data for the computation of statistical results.

III.3 Completion of the planning

24. The estimated time effort, the personnel involved, and the computer runtimes represent essential data for the cost planning for data editing processes. This information should be supplemented by individual wage rates and flat rates for computer runtimes. Determining all the costs of data editing processes also requires the calculation of the costs of needed materials and equipment.

25. Besides these specific data editing process costs, common costs may be taken into consideration for the development of data editing specific methods and software. A simplifying mechanism may be the use of general offsetting items in this context.

26. The description of risks represents the last but one planning activity. It should sharpen the view on critical data editing processes and prepare survey managers for the treatment of exceptional situations - to some extent. Furthermore, personnel not involved in the planning can easily see if someone is aware of the specific risks of planned data editing processes.

27. Unplanned personnel absence, delays induced by late incoming questionnaires, and unexpected amounts of errors (with greater influence on statistical data than expected) or additional time for corrections represent risks in general. With the exception of unplanned personnel absence all other reasons mentioned can be detected to some extent in the beginning of data editing and enable survey managers to initiate more or less effective countermeasures like an enhanced use of automatic corrections/imputations or a flexible personnel employment if enough resources are available.

28. The last planning step comprises the final check of the data editing plans which should be performed by personnel not involved. The most important aspects may be the consistency of the separate plans, and the attention to output, restrictions and guidelines. A checklist could be a helpful instrument, which should cover aspects like output orientation of a planned data editing strategy, the consistency and robustness of data editing processes and the respective plans, an adequate use of available techniques and efficient use of resources, and respondent- and employee-friendly tools and guidelines.

IV. THE PROVISION OF DATA ABOUT PERFORMED DATA EDITING PROCESSES

29. The activities mentioned above may lead to a solid planning but they are worthless if the plans cannot be compared with performance data. A successful management of data editing processes requires an early identification of the reasons for deviations between the plans and the performance data and a kind of forecasting. So this fact raises the question how adequate performance data can be received.

30. 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/processes – 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 through 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, an overview of the errors and correction, and the improvement of data plausibility. In addition to that, the report should furnish information about the causes of time and cost deviations.

31. The information required for the reports should be gathered from the planning of data editing and measured during the performance of data editing processes. IT systems play an important role with respect to the provision of information required for the management and optimisation of data editing. In accordance with the reflections outlined above they should supply the following information:

Table: Data about performed data editing processes

Field	Example	Remarks
Identification number (ID) of a survey	...	This information is needed for the identification of a survey.
Name of a file	Stat32	This information is needed for the reconstruction of erroneous data.
ID of a record	112	
Date	01222002	It is assumed that there may be general decisions concerning error correction activities. The date is therefore necessary to determine their influence.
Time	1503	The beginning of a plausibility check is needed in combination with the beginning of a subsequent check to gain information on the real work effort.
ID of a plausibility check	A025	The ID of a plausibility check is necessary to calculate the process effort.
Corrected characteristic	SocPos[1]	Implausible data can be reconstructed by means of this information. It is used for determining the most effective checks.
Original value of the corrected characteristic	2	
Personal ID	...	A personnel identifier is needed for the calculation of process costs.

32. The proposed information needs to be elucidated by some additional remarks:

- The information should be provided by a data editing or monitoring system. The further data processing should be performed by a management system which should have access to the data editing plans and formulas as described in the previous sections. This functionality is necessary for comparisons.
- Activities that are not specific for data editing and general 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 use of robust indicators for the calculation of real time efforts and costs.

33. The data in the table enable the calculation of real time efforts for basic data editing activities and supply detailed information about the course of error correction. Together with the numbers of records and an overview of all errors, a kind of forecast can be made by applying the formulas mentioned in section III.1.

V. CONCLUSIONS

34. The way to estimate process duration is focussed on computer-assisted human work and should be enhanced to automatic data editing activities.

35. The rising importance of timeliness as a quality aspect, the scarce resources of NSIs and the fact that personnel costs represent a cost runner of data editing - in spite of an increasing use of automatic methods - demonstrate very clearly the importance of a detailed time scheduling and personnel employment.

36. The activities mentioned above lead to a systematic planning and often require a change of management practices in the meaning of the Capability-Maturity-Model^{viii}. Especially experienced personnel may regard these activities as a new form of bureaucracy but the advantage of the approach is that it may be performed even by less experienced personnel. Furthermore, against the background of continuous budget cuts, NSIs will have no or fewer alternatives concerning their planning processes in the long run.
37. Large surveys and periodic ones may justify the greater planning effort. It would be of benefit to store the respective metadata and to adopt them for the specific runs. In the case of single and smaller surveys, simplified calculation and estimation procedures should be used which take the number of records, variables and checks into consideration. Recommendations based on experience and derived from the more detailed activities described above seem to be required.
38. The above considerations on the designing of data editing processes demonstrate very clearly the complexity of an adequate coordination between deadlines, time effort, personnel, tools and budget restrictions and require a powerful project or process management software. An IT tool should therefore facilitate estimations for shortest and longest questionnaire routes by user-friendly interfaces and should break overall estimations down to the level of checks and characteristics. The software should also enable global settings for e.g. the occurrence of variables as well as their additional adaptation. In addition, an IT tool should display control variables, like the average duration of different types of checks or characteristics, which permit to assess the estimated basic time efforts.
39. The considerations on the planning steps indicate a multiple use of data editing specific metadata like record and check descriptions^{ix}, and information about process factors during the planning activities as well as the use of this information for comparisons with performance data during data editing activities. Furthermore, existing plans are used - sometimes with a few modifications - for repetitive surveys. These facts promote the central storage of data editing specific metadata in office-wide databases so that they are also available for the calculation of central management indicators in the areas of time scheduling, and cost and investment accounting.^x
40. Due to the complexity of the planning activities to be supported, preliminary work like the development of prototypes should be carried out to identify the subject-matter demands on IT-support systems and to explore the degree of support which can be delivered by such supporting systems. Internal specifications of Destatis concerning the IT-support for the planning of data editing indicate that most of the tasks can be supported by commercial project management software. It should be investigated how standard software could be adapted to meet the needs of the planning of data editing processes.

ⁱ UN/ECE, Svein Nordbotten (2000). "Evaluating Efficiency of Statistical Data Editing: General Framework". Geneva, p. 4

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

ⁱⁱⁱ Manfred Schulte-Zurhausen (1995). "Organisation". München, 41pp. Günter Schmidt (1999). "Methoden des Prozess-Managements". WiSt, 9, pp. 241-245. "Total Quality Management Prozesse". VDI/DGQ 5505 (Entwurf), Düsseldorf, pp. 2-17

^{iv} Elmar Wein, UN/ECE (2000). "The planning of data editing". Work Session on Statistical Data Editing, Cardiff

^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} Peter Röthig (1995). "Handbuch für die Personalbedarfsermittlung in der Bundesverwaltung". Bonn, pp. A-9

^{vii} Peter Röthig (1995). "Handbuch für die Personalbedarfsermittlung in der Bundesverwaltung". Bonn, pp. A-15

^{viii} Mark C. Paulk (1995). "The Capability Maturity Model", Reading; Mel J. Turner: "Changing Management Practices", ISIS 2000, <http://www.unece.org/stats/documents/2000.05.isis.htm>

^{ix} Jelke Bethlehem, Lon Hofman (2000). "On the use of XML in the Blaise Environment". International Blaise User Conference. Kinsale

^x Bo Sundgren, UN/ECE (1995). "Guidelines for the modeling of statistical data and meta data". Geneva, pp. 10. Bo Sundgren, UN/ECE (1998). "An information systems architecture for national and international statistical organizations". Meeting on the Management of Statistical Information Technology, Geneva, pp. 9