

STATISTICS SWEDEN's EDITING PROCESS DATA PROJECT¹

Svein Nordbotten
P.O. Box 309 Paradis
5856 Bergen, Norway
svein@nordbotten.com

ABSTRACT

Statistics Sweden has continuously been working to improve the total quality of official statistics. One of the processes in which considerable work has been invested is in editing and imputation of statistical data. Recently a project for developing, evaluating and introducing methods for preserving process data has been established. The process data will be another dimension in the already formalized and implemented metadata system. One component will be the editing process data. This paper discusses some of the aspects of collecting, saving and using editing process data and outlines some possible approaches to the solution of this task.

Keywords: Statistical editing, Process description, Process data, Metadata

1. Introduction

Statistics Sweden (SCB) has been working to improve its methods for the preparation of official statistics for most of its 250 years existence. In more recent years, an objective for the statistical research and development of SCB has been to identify those methods in different parts of the statistical production which contribute most to the improvement of the total quality of statistics (Granquist 1997a).

Considerable research and development efforts in SCB have been invested in editing and imputation of statistical data. A project headed by Mr. Leopold Granquist has been established to explore and define which process data are needed for deciding how to design the best suited editing process for a new survey. Editing process data represent an obvious component in the metadata system already formalized and established in SCB. One purpose of the Swedish metadata system is to support users with better understanding of the statistics disseminated by the office (Sundgren 1991; Sundgren 1994). Another equally important aim is to provide a database for survey designers in SCB to make the best decisions for new statistical surveys, or improvement of existing surveys. The establishment and maintenance of the current metadata system is based on a report system that should be used by all groups in SCB participating in the statistical production.

This presentation includes an outline of the editing process data project and some reflections on how the results from the project may be used in future work.

2. Research Framework

Editing is a process aimed at improving the quality of the statistical products from statistical surveys (Nordbotten 1963). International research indicates that in a typical statistical survey, editing may consume up to 40% of all costs. It has been questioned if the use of these resources spent on editing is justified, and if they can be used more efficiently in other statistical processes (Granquist 1997b).

The answer to the first question can in principle be imagined retrieved from a table as illustrated in Figure 1 and representing the framework of the editing process data project. When designing a new survey, or redesigning an old one, we want to identify first the row of the table corresponding to which type the survey belongs, and second the column corresponding to the editing process considered. The cell of the row and column intersection should inform us about the expected characteristics of editing results including their costs. The table should also

¹ Preparation of this paper was supported by Statistics Sweden. The author is grateful to Mr. Leopold Granquist for a number of constructive comments. The views expressed are not, however, necessarily those of Statistics Sweden.

inform the designer about the expected relative merits of alternative editing processes for the type of survey considered.

	Type of editing process
Type of survey	Type of editing results

Figure 1: The framework of the editing process data project

The construction of such a table will require systematic collection of data concerning the editing processes and their results. It should be noted that it is not survey data, but metadata about the survey processes we are focusing on. There are three types of data, which are important in this connection, i.e. data characterizing:

- the statistical surveys as editing tasks
- the editing process including architecture, specifications, available resources, and process performance, and
- the survey results and their costs.

We must select the variables of editing process, which we expect are most influential for the final statistical results. The aim is to classify observed editing processes by survey type and process type in a finite number of categories with as small as possible variations among the processes in the same category and as large as possible differences among processes in different categories.

As for the second question quoted above, it will not be discussed in this presentation although it is equally important reflecting the interrelationships among different processes which have to be taken into account by the survey designer.

3. Statistical survey descriptors

Every survey, even if it is a periodic, monthly survey, is special. First, all the facts represented by the survey data are different from the survey taken previous month. Second, it is impossible to replicate all collection and processing procedures exactly because numerous factors outside the observation and control of the survey manager are changing continuously. Each survey could therefore be described as different from all others. To express these differences in a meaningful measure requires extensive investigations (Nordbotten 1993).

We take a pragmatic view and aim at the construction of a finite set of exclusive survey categories based on a few variables which we believe reflect the major differences considered for our purpose. The categories must be defined so that they express characteristics of the editing task quite clearly. The following variables seem particularly useful:

- The domain to which the survey belong
- The periodicity of the survey
- The number of records to be edited
- The number of fields contained in each record.
- The primary source from which the data is collected.
- The data collection method.
- The number of fields in other sources, which can be used as auxiliary data for editing.
- The number of statistical estimates required from the survey.

In the SCB, most of this information is already recorded in its metadata system for the majority of surveys, and can easily be retrieved for creation of survey categories and use as background variables in analysis of the editing process.

4. Editing processes descriptors

As already pointed out the aim of the editing activities is to detect and adjust for errors in the data in order to improve the results of the survey. Assume ideal operational definitions for all survey variables exist. If these were implemented truthfully, no errors would exist in the data collected and processed, or in the statistics prepared. These hypothetical data are referred to as the *target data* of the survey.

In practical applications, the target data cannot usually be achieved because the ideal processes required would be cost and time prohibitive for all units in the survey. Instead more convenient and practical procedures are established and executed. The result of these procedures is the *raw data* records.

Description of an editing process can be considered as divided into three parts, the editing architecture describing principles for the process, the editing specifications describing the numerical details of the architecture and resources available/used, and finally, the editing performance describing the execution of the editing process.

In the next paragraphs we shall explain the three main components of the editing metadata.

4.1 Architecture

If we could compare the target records with the raw records for each unit in a survey, the micro errors would be identified. Similarly, if statistical aggregates based on the raw data could be compared with the corresponding aggregates of the target data, the macro errors would be discovered.

Obviously, the problem is that we do not know the target data records. The first component to be considered in an editing architecture is therefore a method to detect the raw data records that contain the most influential errors. A number of principles have been proposed and methods developed for this purpose. Most of the methods exploit some kind of background knowledge about the units to define classes or ranges of erroneous or suspicious record patterns (UN/ECE 1997).

If a set of raw records has been identified as erroneous or suspicious, the next component of the architecture will be to find or develop a method for adjusting the records. Also for this task a repository of methods exists. The methods range from a complete re-collection and -processing of records for the units which were associated with the original suspicious raw records to using background information about the unit with a suspicious record, or similar units, to impute new more probable values to replace suspicious values.

Combining existing and/or new methods produces a number of possible editing architectures. The architecture data will therefore consist of descriptions of or references to individual methods applied and how they are combined in the editing process.

4.2 Specification

To adapt an architecture to a specific application, numerical specification of a number of parameters, such as lower and upper bounds for ratios of variables, determination of acceptable and suspicious combinations of categories for discrete variables, etc., are required.

The specification of the editing methods also implies available or allocated resources of different kinds, such as human inspectors, computer capacity and time. Data on specifications of the type indicated will play a central role in evaluation of editing processes because they in large will reflect the resources used in the application. A large part of the specifications has never been systematically recorded.

4.3 Performance

The part of the editing process, which has received most attention, is description of the process performance (Engström 1996; Engström 1997; Engström and Granquist 1999). Some typical variables, which can usually be recorded during the process, are shown in *List 1*. These basic variables give us important facts about the editing process.

- N: Total number of observations
- N_C : Number of observations rejected as suspicious
- N_I : Number of imputed observations
- X: Raw value sum for all observations
- X_C : Raw value sum for rejected observations
- Y_I : Imputed value sum of rejected observations
- Y: Edited value sum of all observations
- K_C : Cost of editing controls
- K_I : Cost of imputations

List 1: Typical operational and cost variables

If the number of observations classified as suspicious in a periodic survey increased from one period to another, it can, for example, be interpreted as an indication that the raw data have decreasing quality, or that the edit parameters have become out of date since they were established. On the other hand, if the editing procedure has a satisfactory quality, it can also be regarded as indication of increased quality of the results, because more units are rejected for careful inspection. A correct conclusion may require that several of the variables be studied simultaneously. As a first step toward a better understanding of the editing process, the basic variables can be combined in different ways. *List 2* gives examples of a few composite variables being used for monitoring and evaluating the editing process.

The *reject frequency*, F_C , indicates the relative extent of the control work performed. This variable gives a measure of the workload a certain control method implies, and is used to tune the control criteria according to available resources. In an experimental design stage, the reject frequency is used to compare and choose between alternative methods.

The imputation effects on the rejected set of N_C observations are the second group of variables. The *impute frequency*, F_I , indicate the relative number of observations which have their values changed during the process. F_I should obviously not be larger than F_C . If the difference $F_C - F_I$ is significant, it may be an indication that the

rejection criteria are too narrow, or perhaps that more resources should be allocated to make the inspection and imputation of rejected observations more effective.

Frequencies:

$$F_C = N_C / N \quad (\text{Reject frequency})$$

$$F_I = N_I / N \quad (\text{Impute frequency})$$

Ratios:

$$R_C = X_C / X \quad (\text{Reject ratio})$$

$$R_I = Y_I / X \quad (\text{Impute ratio})$$

Per unit values:

$$\underline{K}_C = K_C / N \quad (\text{Cost per rejected unit})$$

$$\underline{K}_I = K_I / N \quad (\text{Cost per imputed unit})$$

List 2: Some typical operational and cost ratios

The *rejected value ratio*, R_C , measures the impact of the rejected values relative to the total value of all raw values. A small rejected value ratio may indicate that the suspicious values are an insignificant part of the total of values. Alternatively, the total of rejected values could be compared with the edited value sum of all observations. In both cases, the indicator may hide large changes in opposite directions.

If the rejected value ratio is combined with a high F_C , a review of the process may conclude that the resources spent on inspection of rejected values cannot be justified and are in fact better used for some other process. R_C may show that even though the F_C is large, the R_C may be small which may be another indication that the current editing procedure is not well balanced.

The *impute ratio*, R_I , indicates the overall effect of the editing and imputation on the raw observations. If R_I is small, we may suspect that resources may be wasted on editing.

Costs per rejected unit, K_C , and *cost per imputed unit*, K_I , add up to the total editing cost per unit. The costs per product (item) have to be computed indicators based on a cost distribution scheme since only totals will be available from the accounting system.

The process data are computed from both raw and edited micro data. The importance of preserving also the original raw data has now become obvious and it should become usual practice that the files of raw and edited micro data are carefully stored.

The process variables computed are often used independently of each other. The editing process can easily be evaluated differently depending on which variables are used. The purpose of the next section is to investigate how the process can be described by a set of interrelated variables, which may give further knowledge about the nature of the editing process and a basis for improved future designs.

5. Statistical quality descriptors

Editing itself has no meaning in itself. It is the impact it may have on future statistical results, which counts. We will use the quality concept as a summary descriptor for a statistical product, and discuss this descriptor in more detail in the next paragraphs.

5.1 Quality and errors

The quality of a statistical product, is determined by a number of factors including product *relevance* (correspondence between the target concept and the concept required by a typical application), *timeliness* (the period between the time of the observed events and the time at which the product was used), and *accuracy* (the deviation between the target size and the product size) (Depoutot 1998). Wider quality concepts, as used for example by Statistics Canada, include also accessibility, interpretability and coherence (Statistics Canada 1998). It should be noted that the quality as experienced by two or more users might be different because their needs may have different requirements. In this presentation, we consider mainly the accuracy dimension of quality.

The *users* want quality descriptors to decide if the supplied statistics are suitable for their needs, while the *producers* need data on quality to select among alternative production strategies and to allocate resources for improving overall production performance. Quality can never be precisely described. One obvious reason is that the precise quality of a statistical product presumes knowledge of the target size, and if we knew the target size there would be no need for measuring the fact. Another reason is, as mentioned above, that the desired target concept may vary among the users. The descriptor reflects the quality of some kind of average user. While a quality statement expresses uncertainty about a statistical product, uncertainty will also be a part of the quality measurement itself.

5.2 Measuring quality

So far, the statistical product quality has been discussed from an abstract perspective. To be useful, the abstract notion of quality must be replaced by an operational variable, which can be measured and processed.

As already discussed, quality cannot be observed exactly, but it can, subject to a specified risk, be *indicated* by an upper bound for the deviation of the product size from the target size.

Consider the expression:

$$Pr (|Y'-Y|>D)=1-p$$

which implies that the risk that the product size Y' deviates from its target size Y with more than D is $(1-p)$. D is a *quality* metric even though it decreases by increasing quality and in fact is an error metric. Because Y is unknown, we must substitute D by an estimate D' (Nordbotten 1998). Subject to certain assumptions, it can be demonstrated that $D' = k (p) * var (Y')$ where the value of k is determined by the probability distribution of Y' and a specified value of p . Assuming that Y' has a normal distribution, k is easily available in statistical tables.

In order to compute the estimate D' , we need a small sample of individual records with edited as well as raw data to estimate the variance of Y' . If the raw records for these units can be re-submitted for an approximately ideal editing to obtain a third set of records containing individual target data, we can compute Y as well as D' for different confidence levels p . It can be shown that a smaller risk $(1-p)$ is related to a larger D' for the same product and sample.

Because D' is itself subject to errors, the estimator may or may not provide satisfactory credibility. It is therefore important to test the estimator empirically. In experiments with individual data for which both edited and target data versions exist, statistical tests comparing estimate D' , and the target D , can be carried out (Nordbotten 1998; Nordbotten 2000; Weir 1997).

Manzari and Della Rocca distinguish between *output oriented* and *input oriented* approaches to evaluation of editing and imputation procedures (Manzari and Della Rocca 1999). In an output oriented approach they focus

on the effect of the editing on resulting products, while in an input oriented approach they concentrate on the effect of the editing on the individual data items. The quality indicator D' presented in here is a typical example of an output oriented approach while the performance variables discussed in section 4, are examples of an input oriented approach to evaluating the editing procedures.

6. Analyzing editing data

Metadata of the type outlined above offer opportunities for systematic exploration and evaluation of relationships among the statistical product quality and the editing process variables considered. The type of relationship we assume to exist between the data can be depicted by *Figure 2*.

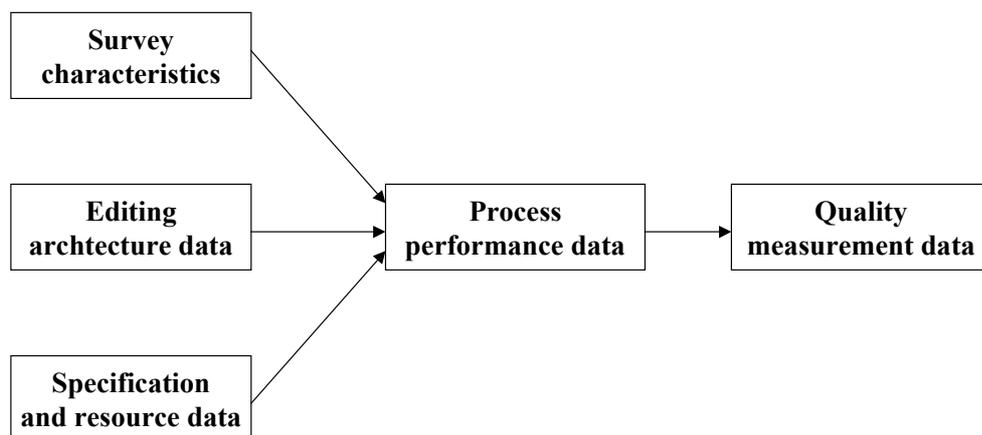


Figure 2: The editing-quality model

The model assumes that data on the left hand reflect the factors determining the performance of the editing process described by the data in the middle of the model. The process performance data in turn characterize the process determining the quality of the statistical results represented by the data at the right side of the figure.

While the first set of data to a large extent already is collected systematically in SCB, the performance data should be made a part of the metadata system. The quality measurements, which will require an additional processing of a small sample, added to the statistical production. This will as we pointed out above be an extra process, which can prove to be both and have a negative effect on the timeliness, another dimension of the quality expensive, and have a negative effect on the timeliness, another dimension of the quality.

It has been assumed that the performance data also indicate valuable information upon the quality of the editing results. A major task for the SCB project might be to investigate correlation between the performance data and the quality measurement data for some selected surveys. If there is evidence for correlation between some of the performance variables and the quality measurements in representative surveys, it is hoped that the less expensive performance variables can be used as quality indicators substituting the expensive quality measurements.

7. References

- Depoutot, R. (1998), "Quality of International Statistics: Comparability and Coherence," Presented at the Conference on Methodological Issues in Official Statistics, Stockholm.
- Engström, P. (1996), "Monitoring the Editing Process," Presented at the UN/ECE Works Session on Statistical Data Editing, Voorburg.
- Engström, P. (1997), "A Small Study on Using Editing Process Data for Evaluation of the European Structure of Earnings Survey," Paper presented at the UN/ECE Work Session on Statistical Data Editing, Prague.
- Engström, P. and Granquist, L. (1999), "Improving Quality by Modern Editing", Presented at the UN/ECE Work Session on Statistical Data Editing, Rome.
- Granquist, L. (1996), "The New View on Editing," Presented at the UN/ECE Work Session on Statistical Data Editing, Voorburg.
- Granquist, L. (1997a), "On the CBM-document: Edit Efficiently", Presented at the UN/ECE Work Session on Statistical Data Editing, Prague.
- Granquist, L. (1997b), "An Overview of Methods of Evaluating Data Editing Procedures," *Statistical Data Editing, Vol. 2, Methods and Techniques. Statistical Standards and Studies No 48. UN/ECE. pp. 112-122.*
- Jong, W.A.M. de (1996), "Designing a Complete Edit Strategy - Combining Techniques," Presented at the UN/ECE Work Session on Statistical Data Editing, Voorburg.
- Nordbotten, S. (1963), "Automatic Editing of Individual Statistical Observations," *Statistical Standards and Studies. Handbook No. 2. United Nations, N.Y.*
- Nordbotten, S. (1993): "Statistical Meta-Knowledge and –Data," Presented at the Workshop on Statistical Meta Data Systems, EEC Eurostat, Luxembourg, and published in *Journal of Statistics, UN/ECE, Vol. 10, No.2, Geneva, pp. 101-112.*
- Nordbotten, S. (1998), "Estimating Population Proportions from Imputed Data," *Computational Statistics & Data Analysis, Vol. 27, pp. 291-309.*
- Nordbotten, S. (2000), "Evaluating Efficiency of Statistical Data Editing: General Framework," UN/ECE, Geneva.
- Sundgren, B. (1991): "What Metainformation should Accompany Statistical Macrodata?" R&D Report, Statistics Sweden, Stockholm.
- Sundgren, B. (1994): "Statistical Metadata and Metainformation Systems," Statistics Sweden, Stockholm.
- UN/ECE (1997), "Methods and Techniques," *Statistical Standards and Studies No. 48. UN/ECE, Geneva.*
- Weir, P. (1997), "Data Editing and Performance Measures," Presented at the UN/ECE Works Session on Statistical Data Editing, Prague.