

CONFERENCE OF EUROPEAN STATISTICIANS

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

Topic (i): Measuring the impact of editing in various phases of statistical survey processing

STRATEGIES FOR IMPROVING STATISTICAL QUALITY

Submitted by Svein Nordbotten¹

Invited paper

I. WHY IS DATA EDITING A FOCAL POINT

1. Data editing is a step in the preparation of statistics, and the goal of editing is to improve the quality of the statistical information. International research indicates that in a typical statistical survey, the editing may consume up to 40% of all costs. It has been questioned if the use of these resources spent on editing is justified, if more effective editing strategies may be applied or if the quality may perhaps be improved more by allocating some of the editing resources to other statistical production processes to prevent errors [Granquist 1996 and 1997].
2. The purpose of this presentation is to inform about research sponsored by Statistics Sweden to gain more knowledge about the statistical quality and ultimately approach answers to the above questions.

II. STATISTICAL QUALITY AND A MARKET PERSPECTIVE

3. Large statistical organisations regard their activities as processes producing statistical *products*. The overall task for a statistical organisation is to specify, tune and run its processes to deliver products with as high quality as possible with the available resources. We assume here that quality can be conceived as a measure of how well the statistical producer succeeds in serving his users.
4. Different users will frequently use the same statistical product as an estimate of different target concepts. In this presentation, we ignore the multiplicity target concepts required by different applications and assume that all users share the same conceptual definition, which measured by a perfect process, would give a value referred to as the *target value* for the product.
5. The quality related to a statistical product, is determined by a number of factors including product relevance (correspondence between the concept measured and the concept required by the application), timeliness (the period between the time of the observations and the time to which the application refers), and accuracy (the deviation between the target value determined by a perfect process and the value determined by the imperfect process) [Depoutot 1998]. Wider quality concepts include also accessibility, interpretability and coherence [Statistics Canada 1998].

¹ The author is Professor Emeritus at the University of Bergen, Norway. He works as a consultant on statistical information systems and can be contacted at: svein@nordbotten.com.

6. Figure 1 symbolises by arrows how the different factors may affect the statistical product value and its deviation from target value. The deviation or error is an inverse indicator of quality.

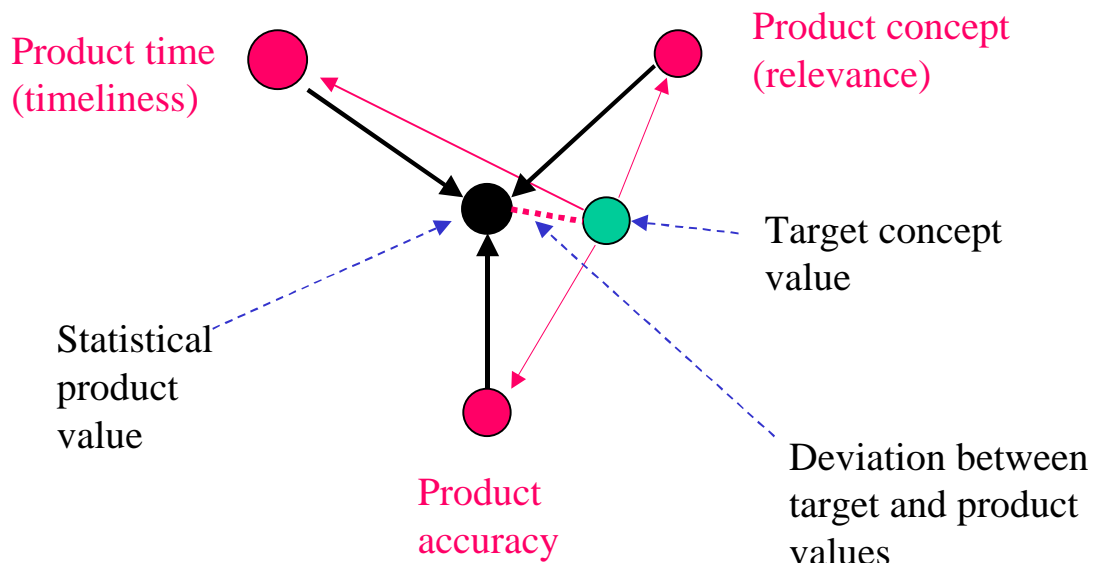


Figure 1: Factors affecting statistical quality

7. To justify the preparation of statistics, the users must benefit from the products. We can imagine a market place in which the statistical producers and users trade. The market mechanism can be described by the sum value of the product for the users as a function of the product quality, and the cost of producing as another function of its quality.

8. Figure 2 presents a simple graphical model of such a market. According to the elementary theory of production, the statistical producer should aim at a quality level, which justifies the costs, i.e. at a quality level for which the value curve is above the cost curve. The market would theoretically be in balance when the marginal value and cost are equal.

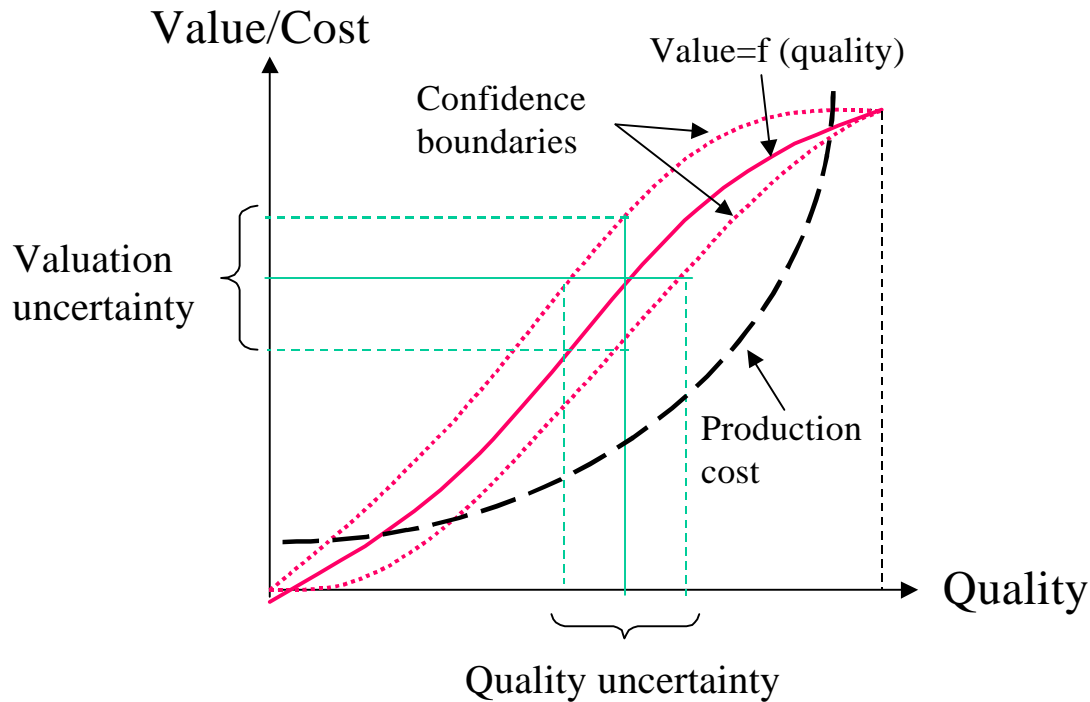


Figure 2: A statistical market mechanism

9. The *users* want data about quality to evaluate if the supplied statistics are suitable for their needs, while the *producers* need data on quality to analyse alternative production strategies and to allocate resources for improving production performance. However, quality can never be precise. One obvious reason is that the precise quality of a product presumes knowledge of the target value, and then there would be no need for measuring the fact. Another reason is, as mentioned above, that the desired target concept may vary by application. Therefore, while a quality statement expresses uncertainty about a statistical product, uncertainty will also be a part of the quality statement itself. This can be illustrated by the stippled curves in Figure 2 indicating a confidence interval for the value-quality curve.

III. STATISTICAL EDITING

10. The preparation of statistics is frequently presented as a number of processes. Figure 3 illustrates how different processes contribute to the quality of the statistical products and require resources, methods and professional competence. Editing is a process aimed at improving the accuracy of the statistical product. It has always consumed a substantial part of most survey budgets, which justifies our interest in evaluating, tuning and improving this process.

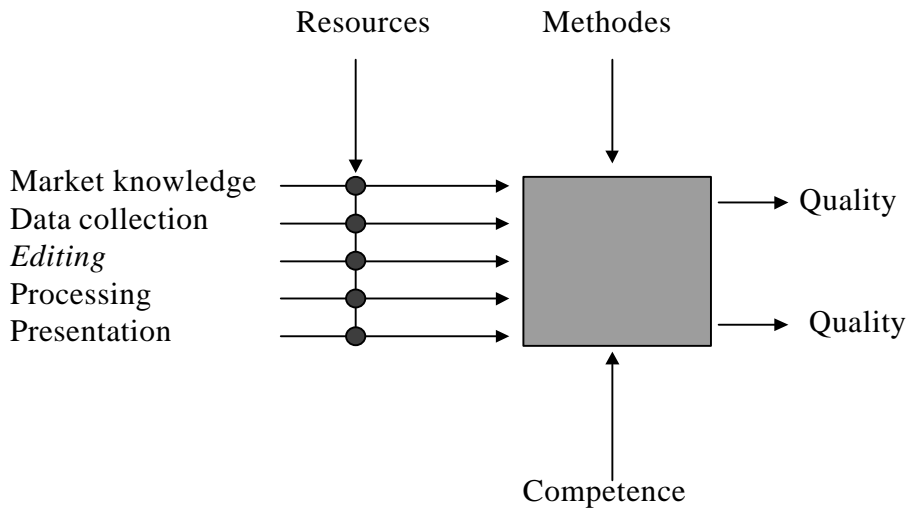


Figure 3: Allocations to statistical processes

11. Assuming that all users have the same conceptual target, it could only be obtained if the statistical process was carried out according to an ideal procedure without any resource restrictions. The deviations of produced individual and population values from their respective targets, are the errors. The aim of the editing process is to catch the errors as effectively as possible within the allocated resources.

12. To describe an editing process, we need four types of data. The first are data about the editing *architecture* describing how the process is designed from different, available methods. These may, for example, be a control algorithm for detecting errors, an algorithm for imputing one category of errors and instructions for manual actions of another category of rejected data. The architecture data inform us how the designer wanted the process to be constructed. The second type of data needed are data about the implemented *structure* of the editing process with operational characteristics, such as specific numerical bounds for edit ratios, imputation functions, etc. These data describe how the process was implemented with all its detailed specifications. The third type of data needed is the *performance* data, which document the operational results of the process applied on a specific set of data. The performance data will include data on the quality of the statistical results and how the quality was operationally obtained. The last type contains *cost* data on which kind of resources were used and how they were spent on different activities.

13. Description of the editing process by means of these four types of data will, in addition to useful information for evaluating and trimming the process, also give us indications about alternative approaches to improve the quality in statistical products.

14. In the following section, the performance data will be discussed in more detail.

IV. MEASURING STATISTICAL QUALITY AND EDITING PERFORMANCE

IV.1 Predicting quality

15. Quality cannot usually be observed by an exact measurement, but can be approximated by a probabilistic approach *predicting*, subject to a specified risk, an upper bound for the product error, i.e. for the deviation of the product value from the target value. This takes the form

$$\Pr (|Y' - Y| > D') = 1 - P \quad 4.1$$

which states that the probability risk that the product value Y' deviates from its target value Y with more than D' is $1 - P$ (small). Because Y is unknown, the metric D' is a uncertain indicator associated with the value Y' [Nordbotten 1998]. We shall denote D' as a *quality predictor* even though it decreases by increasing quality and therefore is an inverse quality indicator.

16. To obtain a prediction of D' , we assume that a small sample of units is available with edited data for individual records. If these units can be resubmitted for an approximately ideal editing to obtain individual target data, we can compute 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.

17. Because D' is itself subject to errors, it may or may not have satisfactory credibility. It is therefore important to test the quality indicator empirically. In experiments with individual data for which both edited and target data versions exist, we can perform statistical tests comparing predicted quality, D' , and factual quality, D , of the edited data [Nordbotten 1998 and Weir 1997].

18. Table 1 illustrates how 504 product values or estimates were classified in an evaluation. The figures based on 1990 Norwegian Population Census data, refer to imputed population totals compared with corresponding target values. Only estimates with a deviation from the target with 5 or less people were acceptable. The quality prediction algorithm classified 430 product values as satisfactory while in fact 414 were within the pre-set requirement. 51 values were predicted as outside the boundary while they in fact were acceptable, a typical Type 1 classification error. On the other hand, 67 values were predicted acceptable while their deviations were greater than 5, misclassifications of Type 2.

19. With a $P=0.75$, we should expect that 25 percent of the values would be subjected to a Type 1 misclassification. As the table shows, the number of Type 1 errors is well within the expected limit. The explanation of this unexpected result is that the distribution of D' does not approximate closely the normal distribution.

		Factual deviation D	
		$D \leq 5$	$D > 5$
Predicted deviation D'	$D' \leq 5$	363	67
	$D' > 5$	51	23

Table 1: Testing 504 product values requiring $|Y'Y| \leq 5$ assuming $P=0.75$.

20. Manzari and Della Rocca distinguish between output and input oriented approaches to evaluation of editing and imputation procedures [Manzari and Della Rocca 1999]. Because they evaluate editing processes by means of data with synthetic errors introduced, they are able to follow an input oriented approach. The quality indicator D' presented in this section is a typical example of an output oriented approach.

IV.2 Measuring process and cost data

21. Two logical steps constitute the editing process: 1. *Classification* of an observation as acceptable or suspicious, and 2. *Correction* of components believed to be wrong.

22. Before the event of automation in statistical production, subject matter experts carried out editing, frequently with few formal editing guidelines. Later, computers were introduced and provided new possibilities for more efficient editing but required also a formalisation of the process [Nordbotten 1963]. Editing principles were developed and implemented in a number of tools for practical application. Today, a wide spectrum of different methods and tools exists. An editing architecture adjusted to the actual survey can be designed by a combination of available tools [UN/ECE 1997].

23. While the quality evaluation focused on the *final effects* of the editing process on the statistical products, the objective of the process evaluation is to describe what is happening with data *during* the editing process [Engström 1996 and 1997].

24. The measurement of the quality effects of editing is based on comparison between edited data and target data. The process measurement on the other hand, is based on comparison between raw (unedited) and edited data. Process data are generated during the process itself and also frequently used for continuous monitoring of the process.

25. Some typical variables which can be recorded during the process are shown in List 1. These basic variables give us important facts about the editing process. They are descriptive facts, and can be useful if we learn how to combine and interpret them.

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

26. As a first step toward a better understanding of the editing process, the basic variables are combined in different ways. List 2 gives examples of a few ratios frequently being used.

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

27. 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.

28. 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 is significant, it may be another indication that the rejection criteria may be too narrow, or perhaps that more resources should be allocated to make the inspection and imputation of rejected observations more effective.

29. 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. If 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 in some other area. In skew distributions, low valued records are frequently rejected. 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.

30. The *impute ratio*, R_I , indicates the overall effect of the editing and imputation on the raw observations.

31. *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.

32. The process data will comprise parameters of the process function, which 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.

V. ANALYSIS

33. *Meta data* of the type outlined in section 4 offer opportunities for systematic research and evaluation of relationships among the variables considered. The objective of an analysis is to arrive at a model of the editing process, which describes the causal relationships among the variables discussed.

34. The set of editing architectures, the users' demands and the resources available, are the environmental conditions within which the implementation structure for a specific statistical product can be selected. The selected structure is next assumed to determine the operational performance and finally the quality and cost of the editing associated with the product.

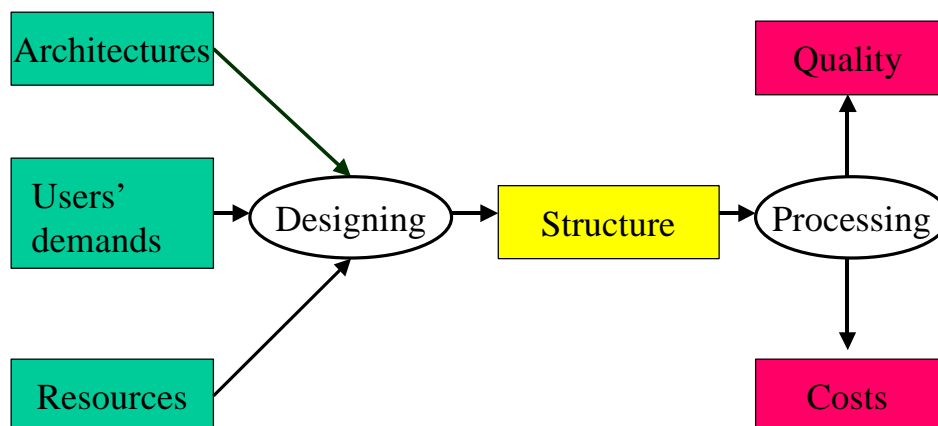


Figure 4: Causal relations

36. Figure 4 outlines the general structure of the model in mind. On the left hand are the different alternative conditions represented. From these sets, a certain structure is selected (designed). Each alternative structure provides to certain quality and cost levels.

37. The causal relations among the different components are symbolised by arrows in the figure. Using the notation already introduced, we can write down the model in a symbolic form by:

$$S=f(A, D, R), \quad 5.1$$

where A, D and R are an architecture, a specific set of users' demands and a level of available resources, respectively, selected from the sets A, D and R. The function f represents the design determining a specific structure S from the set S of feasible structures. The selected structure S in turn determines the quality Q and the costs K from the respective sets Q and K by the process functions q and k:

$$Q=q(S) \quad 5.2$$

and

$$K=k(S). \quad 5.3$$

When the Q and K both are determined, Q needs to be compared with K. A model corresponding to the market Figure 2 is needed, i.e. an equation reflecting the relationship between the quality Q and the market value V:

$$V=v(Q). \quad 5.4$$

The quality can by V be compared with the associated K. Alternative designs, i.e. different structures, can also be evaluated, compared and ranked.

38. Exploring these relations empirically will be an important challenging and long-term project of the future because data from several surveys as well as from the statistical market will be needed.

VI. IMPROVING THE EDITING STRATEGY

39. We want answers to the question which is the 'best' editing strategy for a given market situation subject to available architectures and resources. To obtain a tool for improving the editing strategy, we have to 'turn around' the causal model discussed in the last section into a decision support model. This 'inverted' model is depicted in Figure 5.

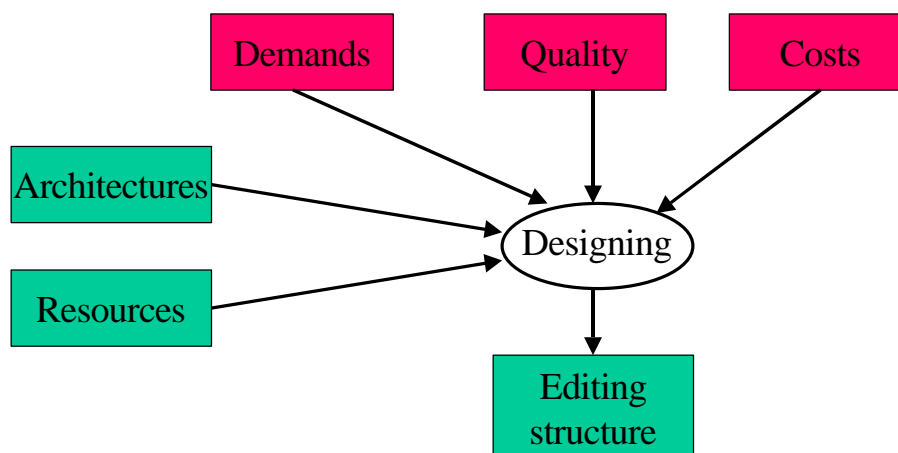


Figure 5: Strategy model

40. If a statistical market demands a statistical product subject to the condition that the market value of its quality must exceed its cost, the strategy model should assist the statistical producer to investigate if a feasible design (editing structure) exists given the architectures (editing methods/techniques) and resources (competence, funds, etc.) he commands. If several feasible structures are identified, the model should advice, which will give the highest quality value/cost ratio.

41. Based on the model outlined in section 5, we can imagine the problem specified as the structure S that gives the highest non-negative solution to the expression:

$$H=V(q(S))-K(S) \qquad 6.1$$

for the demands D and subject to a selection of S within the available sets \mathbf{A} and \mathbf{R} as restricted by 5.1.

42. In the long run, research must be generalised to include also other statistical processes, e.g. data acquisition, estimation and presentation of statistical products [Jong 1996, Nordbotten 1957]. Only such research can hopefully give the necessary tools for tuning resource allocations across all processes in order to obtain the best quality statistics in a wider sense.

VII. FURTHER WORK

43. Considerable research is needed for further evaluation of editing procedures. Among the necessary tasks, the following steps seem obvious:

- i. Development of a conceptual framework for description of editing processes.
- ii. Collecting empirical data sets suitable for experimentation.
- iii. Comparison and evaluation of relative merits of available editing tools.
- iv. Research on causal model description of the editing process.
- v. Discussion of strategy models for improving quality by editing.

44. The essential assumption for comparison and evaluation of editing architectures is access to empirical microdata. Examples of data sets used for comparison have been reported [Kovar and Winkler 1996]. Such sets should be stored in a standard form in a repository and made accessible to researchers working with statistical editing evaluation. Both raw and edited microdata should be stored. When existing, a sample of re-edited ('target') data will be very useful for quality evaluations. If not existing, an alternative is data sets with empirical target and synthetic raw data as discussed in another paper [Manzari and Della Rocca 1999].

45. Research on causal models will require detailed data from the editing process. Data from simulations, may also be a starting base for evaluating the relations outlined in section 5.

Acknowledgement

Helpful comments and suggestions from Mr. Leopold Granquist and Mrs. Joan C. Nordbotten to previous versions of this paper are gratefully acknowledged.

References

- Depoutot, R. (1998): Quality of international statistics: comparability and coherence. Conference on Methodological Issues in Official Statistics. Stockholm.
- Engström, P. (1996): Monitoring the editing process. UN/ECE Work 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. UN/ECE Work Session on Statistical Data Editing, Prague.

Granquist, L. (1996): The New View on Editing. UN/ECE Work Session on Statistical Data Editing, Voorburg. Also published in the International Statistical Review, Vol. 65, No. 3, pp.381-387.

Granquist, L (1997): 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. UN/ECE Work Session on Statistical Data Editing, Voorburg.

Kovar, J. And Winkler, E.W. (1996): Editing economic data. UN/ECE Work Session on Statistical Data Editing, Voorburg.

Manzari, A. and Della Rocca, G. (1999): A generalized system based on simulation approach to test the quality of editing and imputation procedures. UN/ECE Work Session on Statistical Data Editing, Rome.

Nordbotten, S. (1957): On errors and optimal allocation in a census. Skandinavisk Aktuarietidskrift. pp. 1-10.

Nordbotten, S. (1963): Automatic editing of individual statistical observations. Statistical Standards and Studies. Handbook No. 2. United Nations, N.Y.

Nordbotten, S. (1998): Estimating population proportions from imputed data. Computational Statistics & Data Analysis, Vol. 27, pp. 291-309.

Statistics Canada (1998): Quality guidelines. Third Edition, Ottawa.

UN/ECE (1997): Statistical Data Editing Methods and Techniques. Vol. 2. Statistical Standards and Studies No. 48. UN/ECE, N.Y.

Weir, P. (1997): Data editing and performance measures. UN/ECE Work Session on Statistical Data Editing, Prague.