SN, 19 May 2000

Meta-data about Editing and Accuracy for End Users[1]

by

Svein Nordbotten[2]
P.O.Box 309 Paradis
N-5856 Bergen, Norway

1. Introduction

This paper is focused on the needs of the end users of statistics for meta-data about the statistics offered by the statistical producers. It is set within a framework proposed in an earlier paper [Nordbotten 2000a]. Since the discussion is in the context of data editing, the discussion is mainly limited to meta-data about the accuracy of statistics and statistical data editing.

Several questions are addressed in this paper. The first question discussed is what are the end users' needs for meta-data and how will the users of statistics react on meta-data. A second question addressed is how should a statistical producer react to the information needed by the users and provide meta-data, which might serve the needs. At the end of the paper, a short discussion is included about required research and development for implementation of a meta-data service.

2. Information needs of end users of statistics

In the paper referred to above, it was proposed that statistics could be considered as ***products*** delivered from a statistical producer. We shall assume that each product is described by ***4*** production attributes: ***identification***, ***accuracy***, ***process data***, and ***size.*** Identification determines which real world fact the product is assumed to reflect, accuracy indicates the quality of the measurement, process data explains how the product was generated, and size refers to the measured value of the product. The identification, accuracy and process data are ***meta-data***, which tell the user about the measurement while size is the statistical figure, which the end users want to use to solve their respective problems.

Modern meta-data systems are expected to include a number of data useful for the users and producers [Sundgren 1991, Nordbotten 1993]. Examples of meta-data are conceptual and

[1] This paper is an invited paper to the ECE Work Session on Statistical Data Editing in Cardiff, 18-20 October 2000. The paper has been prepared in co-operation with and is funded by Statistics Sweden.
[2] The author can be contacted by email: svein@nordbotten.com and through http://www.nordbotten.com.

operational definitions of facts represented by identification attributes, description of procedures for and performances of the important processes such as the selection of units observed, the collection techniques, editing, estimation, presentation and dissemination of results, which are all components of the statistical production. This paper addresses meta-data originating from the editing process because this process has as its main objective to contribute to the accuracy of statistical products.

The aim of the editing is to detect and adjust errors in the collected microdata. From the process, important process data can be obtained useful for producers to improve the production design and useful for the end users to evaluate the available statistics for their purposes [Nordbotten 2000b].

In early days, the main objective for producers of official national statistics was to provide statistics satisfying the needs of their national governments. The production was typically also funded by government grants and the producers were assumed to prepare the statistics as accurately as the granted funds permitted. In the 20th century, research and private industry became important users of official statistics. In the beginning of the present century, electronic technology, communication and globalisation trends dominate. In the future, we expect statistics, as any kind of information, to become of vital interest to new groups of users and to be valued as commercial commodities. This will require an exchange of information between producers of statistics and the growing variety of users.

It is therefore important to have a conceptual picture of which information producers and users need, and how the actors react to meta-data about accuracy. We illustrate in ***Figure 1*** how the future statistical market mechanism may work. Each quadrant of the figure represents a special aspect of the market.

The ***Southwest quadrant*** explains the relation between production cost, ***C=C (A)*** and the accuracy, ***A***. It assumes that increasing the quality of a product requires resources and increases the cost of producing the product.

The ***Northwest quadrant*** reflects the number of uses of a product, ***N***, as a function of its accuracy, ***A,*** assuming the product is free, ***N= G(A; P=0)***. The more accurate the product is, the more it will be used. This corresponds to the traditional situation when statistical products are available without cost for the users and they have to guess the accuracy of the products.

When a product is provided for a price set by the producer, the number of uses is assumed to be determined by ***3*** factors, the kind of product, its accuracy, and its price. For each accuracy level, there may be a different demand curve. As depicted in the ***Northeast quadrant***, the demand curve, ***N= F(P; A=a)***, determines the number of uses of the product, ***N***, as a function of its price, ***P***. A product with a higher accuracy is assumed to have a demand curve to the right of the curve for the same product with a lower accuracy. Note the relationship between the two functions in this and the previous quadrant.

Finally, the producers gross income, ***I***, from a product with a given accuracy, ***a***, is determined by the income curve ***I=F (P; A=a)*P*** in the ***Southeast quadrant***. The gross income curve can be compared with the cost curve corresponding to the accuracy of the product, and the net income***,*** ***i-c***, deducted.

In a statistical market in which the statistical products are free or charged only with the cost of printing and distribution, the number of uses is determined by the curve in the Northwest quadrant if data on the accuracy are available for the users. If the product is free and no accuracy data are available, the users will evaluate the product on basis of earlier experience, and decide if the product can be used or not for their specific. On the other hand, information on the product accuracy does not imply that the producer must request a price for his product.

For the statistical producer, several productions and marketing strategies are possible:

1) If the goal of the statistical producer is to maximize the number of uses of a product, he may choose to allocate as much as possible of resources to increase accuracy by editing, etc. Maximum of uses will then obviously be obtained when a product is offered free of charge, ***P=0***. The number of uses would in the example be ***N=n₁***.

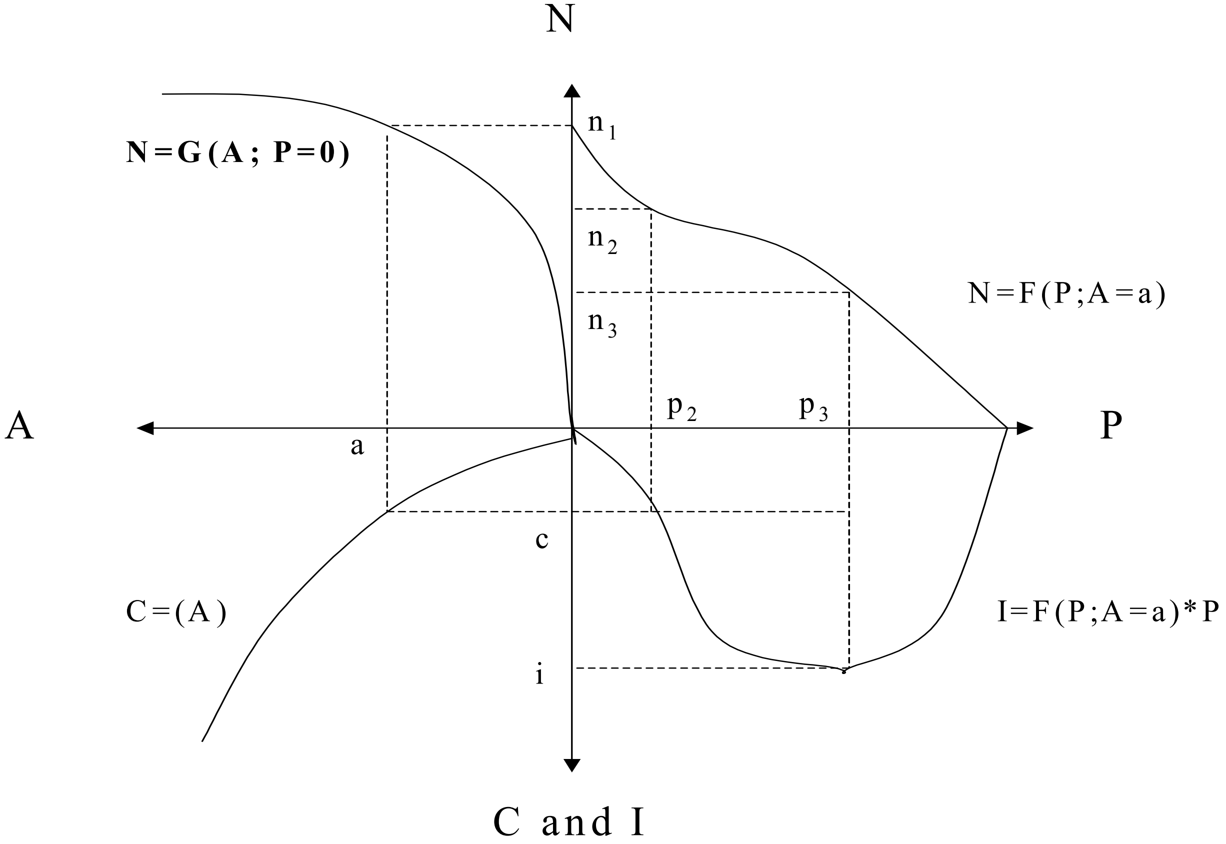


Figure 1: The statistical market

2) An alternative strategy for the producer would be to recover production cost, ***C=c***, for a product with a given accuracy ***a***. The producer must then set the price, ***P=p₂***, such that, ***I=F (P=p₂; A=a)*P=p₂***, equals the costs, ***C=C (A=a)***. The implementation of this strategy requires that the users have information about the product accuracy and that the producer knows the demand curve for this accuracy. The number of uses in this case would be ***n₂.***

3) A third producer's strategy would be to aim at a product price for a given accuracy, which would give a maximum income. This objective would be satisfied when the price is set to maximize the demand function $\boldsymbol{I=F(P; A)*P}$, i.e. $\boldsymbol{P=p_3}$. This also assumes that the producer has acquired knowledge about demand curves for the relevant accuracy level, and that the users are informed about the accuracy and the price set for the product. The number of uses would in this situation be $\boldsymbol{N=n_3}$.

A more challenging problem would be to locate both the accuracy level and the price for the product, which simultaneously generates the maximum net income. All the above-indicated strategies are referred to a single product. In a real world system, there will be multiple products, and the producer must consider the optimum accuracy and price simultaneously for all products.

3. Meta-data concerning editing

In the previous section, we emphasized that providing the end users with information on accuracy/editing must be a two-way communication to be effective. The end users need to be informed about ***3*** meta-aspects of the statistical product as well as the price asked:

- identification of the social/economic fact measured,
- accuracy of the product,
- process data including how accuracy measurement was done ant its reliability
- price requested for the product.

This meta-information should be available in alternative forms depending on the wishes of the end users.

On the other side, statistical producers must also acquire knowledge of the users' reactions to the information about product accuracy and price setting in order to be able to adjust adequately to needs and accuracy.

A ***2-way*** exchange of information about the statistical products between the producer and the users as indicated in ***Figure 2*** is therefore needed. Without this mutual exchange of information, the users' behaviour will be arbitrary and dominated by uncertain knowledge and experience while the producer's choice of strategy for serving the market will be inefficient and dominated by traditions.

The conditions for the 2-way exchange to function effectively are:

1) procedures for providing the users with information about the existing products, and
2) procedures for acquiring feedback responses from the users.

To fulfil condition one, the statistical producers must decide on a marketing policy for providing meta-data about accuracy about their products. Traditional marketing implied that a statistical product was provided with a name permitting the users to identify which real fact the product aimed at describing and a price, if not free. A modern marketing policy must also include a measure of the accuracy, information about how this accuracy was measured, and advice about how it should be interpreted and used.

A national statistical producer disseminates a large number of statistics each year, and it will probably never be possible to provide accuracy measures for each. The aim must be to give accuracy measurements for products which can be considered representative for others with respect to accuracy.

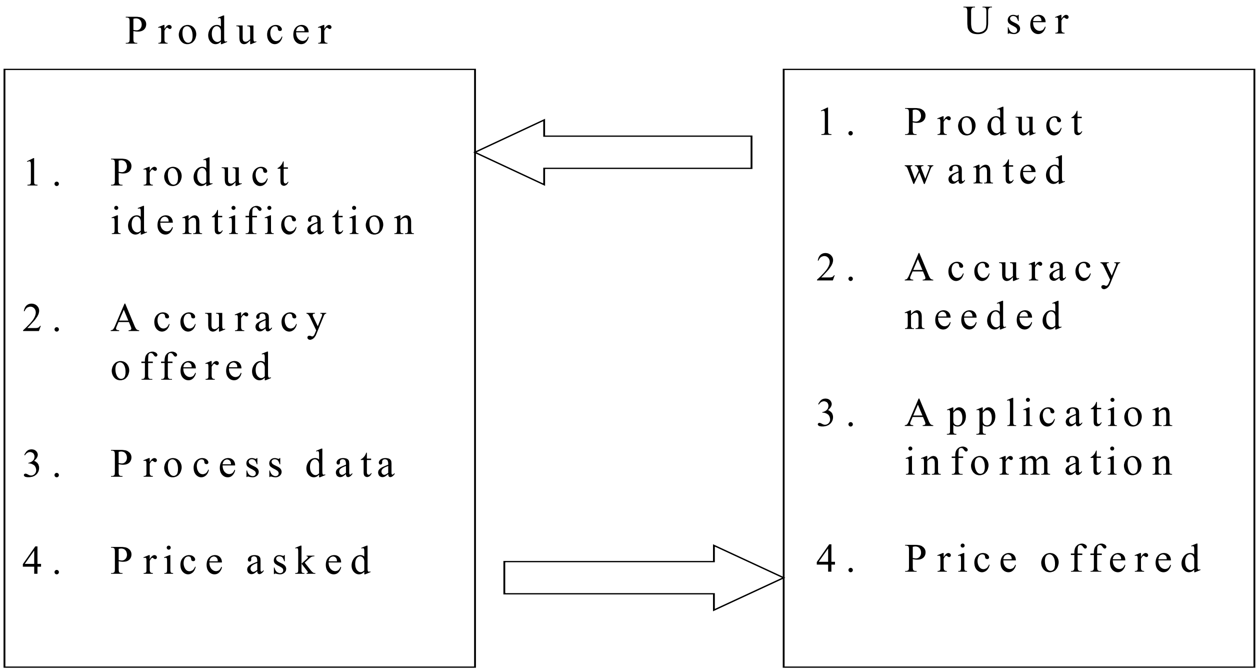


Figure 2: Exchange of meta-data between producer and user of statistics

This policy has been an aim of national statistical offices for many years. Because of the implied costs, it has in most cases been postponed. With exploding needs for statistical information and a trend to consider information products as any other information products, it is now time for preparing a new marketing policy. As a starting point, a few products should be selected for introducing accuracy and editing process meta-data. The best form of meta-data presentation is not obvious and needs research in close cooperation with users.

Feedback from the users is equally important, but a much more difficult problem because it requires active participation of users. Because the statistical products traditionally have been supplied without any or for a small charge, the statistical producer has not had any strong incitement for acquiring data and knowledge about the users needs and their evaluations of products disseminated. However, up to ***40%*** of the typical statistical survey costs, is spent on editing for improving accuracy [Granquist 1996 and 1997]. It should be justified to systematically ask users about their accuracy needs and evaluation of products disseminated. Users should be recruited for representative panels to provide the wanted information.

4. Meta-data on accuracy

So far, we have referred to meta-data on editing in general terms. We need, however, to decide how accuracy from editing should be measured and provided to the end users. We shall refer to the accuracy measurements as ***accuracy indicators***. The general question of measuring product quality was discussed in detail previously [Nordbotten 2000a]. In the present paper, we shall concentrate the discussion on alternative accuracy indicators.

Manzari and Della Rocca distinguished between ***output-oriented approaches*** and ***input oriented approaches*** for evaluation of editing procedures [Manzari and Della Rocca 1999]. Both approaches can provide meta-data reflecting the accuracy of the product. An output-oriented approach focuses on evaluating the impact of editing on the final output comparing edited and true values (in real surveys this can only be possible for small samples), while in an input-oriented approach the evaluation is based on the changes of input data during the editing process. We shall follow a similar distinction for different indicators of accuracy.

4.1 Output-based indicators

The most obvious form for an output-based accuracy indicator is the ***error indicator*** expressing the size order of the error in the product size. Since the error of a statistical measurement can only be determined exactly in controlled experiments (if we in a real survey knew the exact errors, we would of course adjust the estimates to their correct values), we have to be satisfied with an error indicator of accuracy subject to uncertainty. An example of an indicator of this type is an ***upper bound D*** for maximum error specified with a certain degree of confidence. This means that the actual error may be less then the indicator, but that there is also the specified risk that it may be larger.

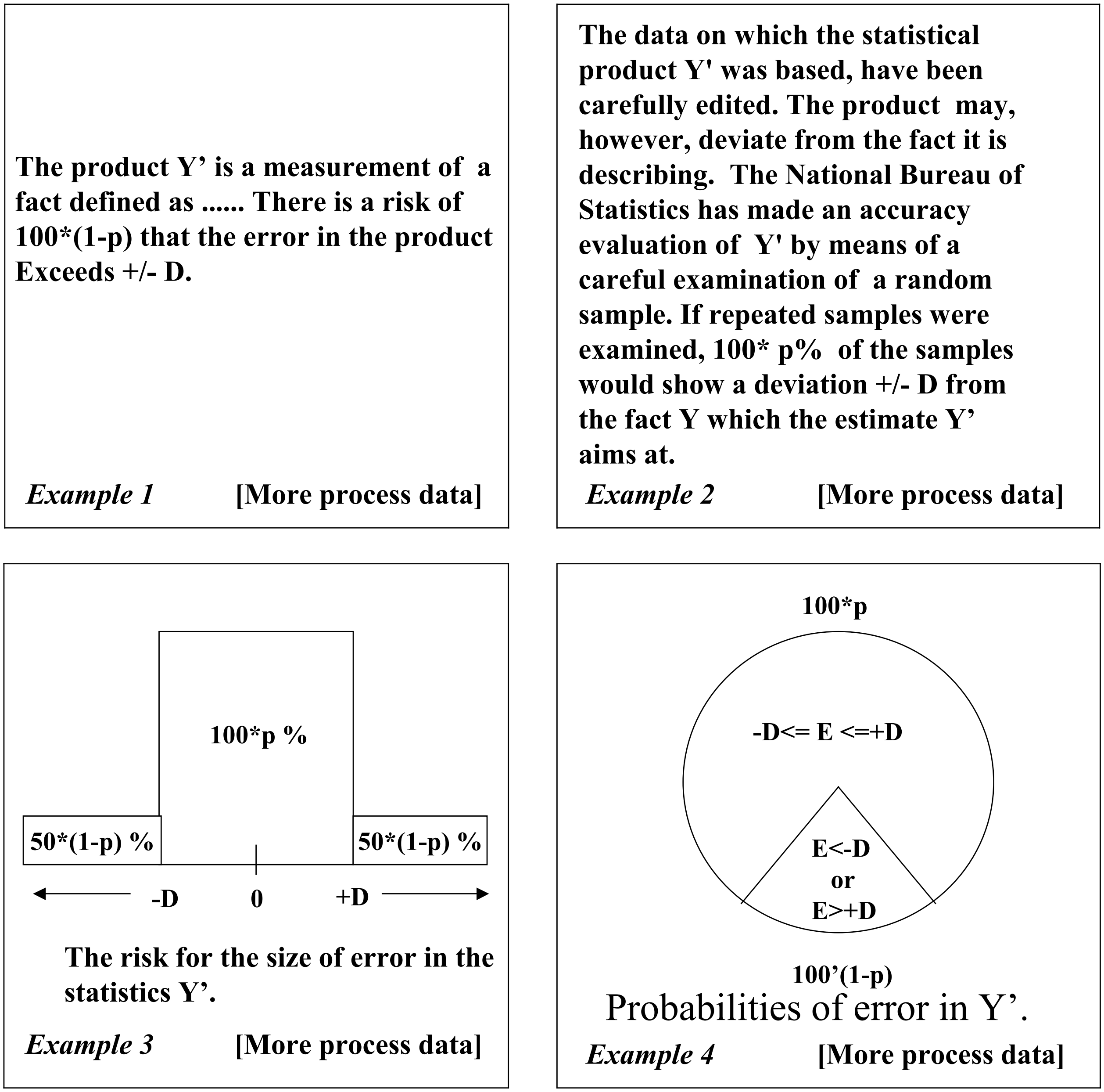
Formally, the indicator ***D*** and a confidence probability p can be presented as:

$$Pr\ (|Y'-Y|<=D)=p$$

where ***Y'*** is the size of the statistics prepared and ***Y*** is the unknown, true value of the fact estimated.

It is not a trivial task to present a probability statement of this type to the end user in such a way that it is understood correctly. It may be formulated as a quality declaration in different ways. In ***Figure 3***, the probability statement is presented in four different examples, two by text, and two by means of graphical tools. Do we know which type will be understood and correctly interpreted by a majority of users? The formulation of the declaration requires careful consideration to be useful for the end users. The four examples in the figure also illustrate the difficulty in conveying the content of probability statements in a simple way.

In any case, there should always be an option for the users to obtain more specific meta-information about the indicators and how they were computed.



The product Y’ is a measurement of a fact defined as ...... There is a risk of 100*(1-p) that the error in the product Exceeds +/- D.

Example 1 [More process data]

The data on which the statistical product Y' was based, have been carefully edited. The product may, however, deviate from the fact it is describing. The National Bureau of Statistics has made an accuracy evaluation of Y' by means of a careful examination of a random sample. If repeated samples were examined, 100* p% of the samples would show a deviation +/- D from the fact Y which the estimate Y’ aims at.

Example 2 [More process data]

The risk for the size of error in the statistics Y’.

Example 3 [More process data]

Probabilities of error in Y’.

Example 4 [More process data]

Figure 3: Four simple examples of accuracy indicators for a statistical product

A serious drawback of output-based accuracy indicators is that they require a second, time-consuming and expensive editing of a sample of already edited data.

4.2 Input-based indicators

In the input-based approach, focus is on raw, input data and edited data. Three types of accuracy indicators, ***frequency, ratio,*** and ***relative*** indicators, can be developed from process data. A typical ***frequency indicator*** based on process date from the editing process is:

Reject frequency = No. of rejected observations/ Total no. of observations.

This indicator tells the users that a certain number of collected values were identified as suspicious, and submitted for further inspection. What does a large frequency indicate? It may indicate that many records have been reinspected and many errors are probably eliminated, i.e. the product has a high quality. Alternatively, it may indicate that the original raw input data had a low quality in general. It will always be difficult to interpret the differences expressed by such an indicator from one survey to another. In most cases, more process information will be useful, for example about possible changes in the collection process.

A number of editing process performance measures can easily be developed as a by-product of the editing process without significant expenses. Assume that a total of imputed values obtained during editing and a corresponding total of the original raw values are computed. We can then define the ***ratio indicator***:

Imputation ratio = Total of imputed values/ Total of raw values

The accuracy of the statistics has been monitored and evaluated during the processing, and it is considered to have improved by x % compared with the accuracy of the statistics for the previous period.

More information on the accuracy can be obtained from the National Statistical Office.

Example 1 **[More process data]**

The data used for preparation of these statistics have been screened. X% of the collected data were rejected as suspicious and re-inspected. As a result Y% of the estimate values is based on data which have been changed.

More information on the accuracy can be obtained from the National Statistical Office.

Example 2 **[More process data]**

Figure 4: Two examples of relative accuracy indicators

A large imputation ratio can indicate high accuracy because the editing process changed many dubious raw values. This conclusion is correct if the adjustment of rejected values always will improve the data. Unfortunately, this cannot always be assumed for example when an automatic imputation has been used. A large imputation ratio may in this case indicate a high degree of uncertainty with respect to the accuracy of the product.

Usually, correct interpretation of process data indicators requires an extensive knowledge of the production process. The metadata reflecting this knowledge should therefore be available to end users. The statistical producer should select the process data indicators, which he believes to be the most reliable indicators of accuracy and present these for the end users in the form of ***relative indicators*** with reference to a base year or the previous period.

4.3 Comparative remarks

Two alternative approaches to inform the end users about the accuracy of statistical products have been outlined. Each has its comparative advantages and drawbacks. Comparing the approaches, we can sum up the results:

1. The output-based indicators permit the end users to evaluate the probably maximum errors he shall have to deal with. He can also compare the errors of different statistics. The main drawback with this kind of error indicators is the extra resources and time needed for computing the indicators.
2. The input-based indicators can be considered as inexpensive by-products from the editing process and are frequently required for improving the process. In many situations, experienced statisticians can deduce accuracy information from these indicators and present it for the end users. Because the preparation of input-based indicators do not include a comparison with any true values, they can not inform about accuracy levels, but about the direction of change in these levels. The user-friendliest form may be to make them available as relative indicators.
3. Meta-data about the editing process should always be available as an option for end users wanting more information about the reliability of the accuracy indicators.

5. Required research and development

To provide meta-data to users about editing and accuracy about statistical products is a challenging objective. Several research and development tasks have been mentioned:

Which data do the end users need and how to collect the data?

The answer requires a 'market analysis'. A first step may be to extend the producer's established contact with established user groups to discuss the needs for meta-data about editing and product accuracy. As a more permanent solution, representative user panels may be needed. These panels should be exposed to experimental market scenarios and express their reactions. Recruiting members for such panels may require that those serving be offered some kind of payment.

In which form should the accuracy meta-data be disseminated?

Accuracy data can be presented in several alternative forms, which may not exclude each other. Most meta-data could be presented as plain text. The challenge will be to find a balance between a shorter presentation, which is read, but frequently misinterpreted, and a longer version needed for correct interpretation, but read only by few. Alternatives can be tabular and graphical presentations.

How to disseminate the accuracy meta-data?

Should meta-data be printed and made easily available as promotional material with product prices, and if so how should this be made known to the general user community? Should links meta-data, etc. be made available at the producer's web site?

7. References

Granquist, L. (1996): 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 Work Session on Statistical Data editing, Voorburg.

Granquist, L. (1997): 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.

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. (1993): "Statistical Meta-Knowledge and –Data," Presented at the Workshop on Statistical Meta Data Systems, EEC Eurostat, Luxembourg, and published in the Journal of Statististics, UN-ECE, Vol. 10, No.2, Geneva, pp. 101-112.

Nordbotten, S. (2000a): Evaluating Efficiency of Statistical Data Editing: General framework. UN/ECE Conference of European Statisticians. Geneva.

NORDBOTTEN, S. (2000b): Statistics Sweden’s Editing Process Data Project. Presented at the ICES II Conference. Buffalo.

Sundgren, B. (1991): “What Metainformation should Accompany Statistical Macrodata?” R&D Report, Statistics Sweden, Stockholm.