

# MEASUREMENT OF DATA QUALITY IN THE CASE OF SHORT TERM STATISTICS

Rudi Seljak<sup>1</sup>  
Metka Zaletel<sup>2</sup>

Statistical Office of the Republic of Slovenia

*Key words: quality indicators, quality of short term statistics, system for quality control*

## ABSTRACT

There is a constantly growing demand for the fast data delivery especially in the case of business surveys. Monthly or quarterly surveys which should provide these data are usually based on samples and are focused just on few variables (i.e. turnover, number of employees). The main goal is to get results of sufficient quality as quickly as possible. Since the deadlines for the publishing of results are becoming shorter and shorter there is a temptation of publishing data of poor quality just for the sake of timeliness. Therefore there is of great importance to constantly measure different aspects of quality of these data in order to avoid publishing results which would not satisfy agreed quality standards.

In the paper we present the system for quality control of statistical results in the case of the short term surveys of enterprises from 3 different areas: retail trade, hotels and restaurants, services. The published results of the surveys are different time based indices. Quality is controlled by the set of indicators defined on the base of the Standard Quality Report and other methodological documents prepared by Eurostat. Since most of the statistical process is automated the aim was to incorporate into this process also the calculation of the majority of defined quality indicators so that consequently we would have the indicators at disposal at the same time as the estimated indices. We describe which indicators are calculated automatically, which were the problems that we faced and what are our plans for the future work in this area.

## 1 Introduction

Quality of the surveys and in fact knowing more about the quality of the statistical products is getting more and more important not only for data users but also to survey managers and to management of the national statistical institutes. They all need quick results which are easy to read, understand and compare with the results of other statistical surveys (similar in design, place or time).

Users of quality measurement results are similar to users of results of statistical surveys: some of them demand thick reports with precise descriptions of all possible views, the others on the other hand want to have very brief results with only few tables, lots of graphical presentations and lots of comparisons with applicable data. Understanding this fact, the development of quality control measurement in statistical surveys had been firstly performed into the direction of Standard Quality Report; afterwards the set of quality indicators has been developed. The Statistical Office of the Republic of Slovenia has followed this strategy. We are trying to incorporate the calculation of quality indicators into data process to be fully automated. The target is to get the set of monthly quality indicators as fast as the first results of the survey and to produce the Standard Quality Report as fully automated as it is possible.

---

<sup>1</sup> Rudi Seljak, Statistical Office of the Republic of Slovenia, Vožarski pot 12, 1000 Ljubljana, Slovenia, [rudi.seljak@gov.si](mailto:rudi.seljak@gov.si), phone: +386 1 2415 110

<sup>2</sup> Metka Zaletel, Statistical Office of the Republic of Slovenia, Vožarski pot 12, 1000 Ljubljana, Slovenia, [metka.zaletel@gov.si](mailto:metka.zaletel@gov.si), phone: +386 1 2415 306

The first test of such a process has been done on the system of three monthly surveys and it is described in the following sections. Later on, we will work on the expansion of such a system to other surveys, especially monthly or quarterly surveys.

## 2 General information of the system

The process of quality control which will be the main focus of the article has been firstly developed for the purposes of some short-term business surveys, especially three monthly surveys: Retail Trade Survey, Survey in Hotels and Restaurants, and Survey in Service Sector. The distinctive characteristics of these surveys are:

- They are monthly surveys with short list of variables in the questionnaires.
- Published results are usually indices of different types.
- There has been growing demand for quick results of the survey.

In year 2003 all the main parts of the methodology of these three surveys has been standardized. The main aspects of the standardized methodology are:

- The questionnaire contains only two variables: turnover and number of employees.
- Surveys are based on the samples where all large enterprises are selected with certainty, while random sample is drawn within small and medium enterprises.
- The sample design is rotating panel with 75% of overlapping units, meaning that same units are kept in the sample for twelve months and afterwards one quarter of the sample is replaced with new units.
- Missing values (unit and item nonresponse) are estimated by imputation method usually called Historical Trend Imputation.
- Special weighting system, which takes into account every year selection probability and estimated ratio between newborn and dead units was developed.
- First results of the surveys are published approximately 55 days after the end of the reference period.<sup>3</sup>

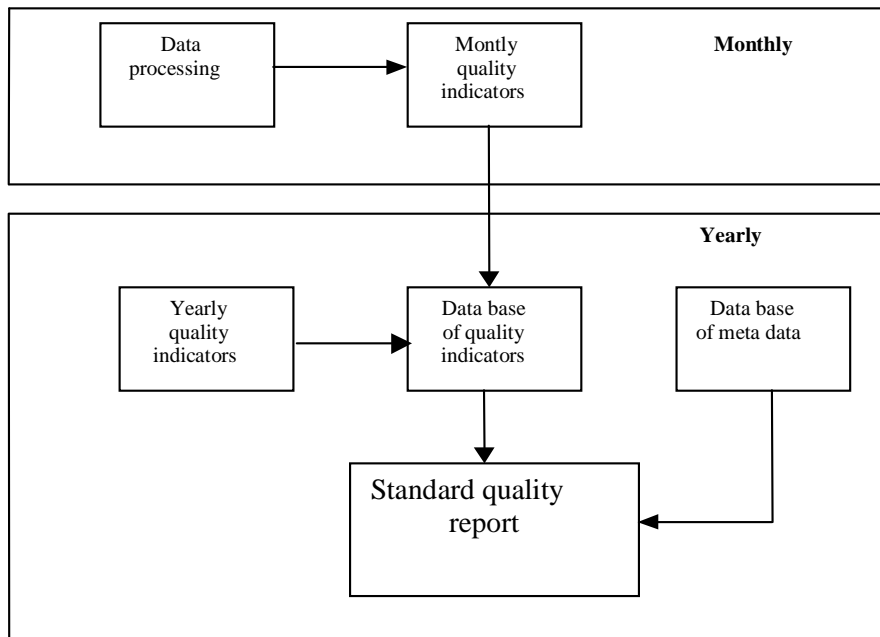
The data of the surveys are stored in two databases. The first database contains information on the units which are defined at the time when sample was drawn (i.e. size class, NACE code, last year's turnover, address, ...). The second database contains information that we get monthly as the result of our survey (raw data). The data in this database are of course keyed every month. The corrections of reported data are allowed for current and previous year.

The data processing including imputation, weighting and tabulation is fully automatized through the set of SAS macro's which are run by using tailor made MS-Access graphical interface. We minimized the time used for this part of the process to approximately 15 minutes and consequently gained pretty large amount of time that we can use for other parts (i.e. data cleaning, data analyzing,...) of the process.

Because of the mentioned demand for quick results there is a risk of publishing results of poor quality. To avoid such publications and to assure quick and effective control of quality of produced results we set up the system of quality control which will be presented in the next sections. The main goal was to incorporate this system as much as possible in the automated process shortly explained above. The basic concept is presented by following picture:

---

<sup>3</sup> The exception are the results for the January when (because of the sample rotation) we publish results approximately 70 days after the end of the reference period.



The system could be shortly summarized as follows:

- Monthly control of produced results is based on the set of quality indicators which are subset of the complete list of quality indicators defined for the purposes of the Standard Quality Report (SQR). We will refer to these indicators as monthly indicators. The calculation of monthly indicators is incorporated into the data processing system and is done automatically every month.
- At the end of the year we calculate »yearly quality indicators« by prescribed methodology. Some of these indicators are calculated by using monthly indicators and some of them are calculated independently. The whole set of quality indicators is stored in the database that contains indicators for different surveys and for different reference years.
- Some additional textual data information (i.e. information on sampling frame, sampling design, media used for publishing,...) that should be included in SQR are also stored in special database which contains these information for different surveys and for different reference years.
- Information from both databases are then merged together into the prescribed and standardized form of the SQR. The standardized form of SQR is defined by Word's template which is directly linked with both databases and enables quick and user friendly procedure of producing final version of SQR.

Each of the above items will be in greater detail explained in following sections.

### **3 Complete list of indicators**

The complete list of quality indicators which should be included in the SQR has been determined on the basis of the list proposed by Eurostat Task Force on Quality Indicators. The proposed list was studied and discussed within the special working group of the SORS. The result of these discussions was the list of 18 indicators defined for the need of the SQR produced in the SORS. The methodology of calculation of these indicators is based on following rules:

- The values of the indicators should be on the interval  $[0,1]$ . The additional values  $\{-2,-1\}$  are also allowed where  $-1$  stands for the case when the value of the indicator is not available and  $-2$  for the case when indicator is not applicable.
- The indicators should be defined in the way that smaller value of indicator means higher degree of quality. Thus value 0 should refer to the ideal degree of quality.

In some cases limitation of values to the interval  $[0,1]$  follows directly from the definition of the indicator, while sometimes some additional calculation to assure the prescribed range is needed. For the later cases it is recommended to publish both, original value (we shall call it nominal value) as well as recalculated value in the interval  $[0,1]$  (we shall call it standardized value).

Firstly, we present the complete list of indicators. For the sake of comparability with the list proposed by Eurostat, we used the same signs for the common indicators.

| Component of quality       | Notation | Title   |
|----------------------------|----------|---|
| Relevance                  | R1       | User satisfaction index   |
|                            | R3       | Rate of unavailable statistics  |
| Accuracy                   | A1       | Coefficient of variation  |
|                            | A2       | Unit nonresponse rate   |
|                            | A3       | Item nonresponse rate   |
|                            | A4       | Editing rate  |
|                            | A5       | Imputation rate   |
|                            | A6       | Overcoverage rate   |
|                            | A7       | Average size of revisions   |
|                            | A8       | Missclassification rate   |
| Timeliness and punctuality | T1       | Punctuality of the first release  |
|                            | T2       | Average time lag between end of the reference period and date of the first release            |
|                            | T3       | Average time lag between end of the reference period and date of the release of final results |
| Accessibility and clarity  | AC1      | Rate of media used for dissemination  |
|                            | AC2      | Rate of means used for dissemination  |
| Comparability              | C1       | Length of comparable time series  |
| Coherence                  | CH2      | Coherence between first and final results   |
|                            | CH3      | Coherence with comparable data from other sources   |

For some of the listed indicators it is recommended to publish also weighted values of indicators. Indicators where (if it is feasible) it should be done are: A2, A3, A4, A5. Weight that should be used for calculation should be sampling weight only.

### Relevance

User Satisfaction Index is based on the User Satisfaction Survey. In the SORS such a survey is planned for this year. Since the methodology of the survey and the methodology of calculation of the results still hasn't been completely defined, we haven't exactly defined methodology of the calculation of indicator R1.

Indicator R2 shows the rate of statistics that are not available according to the number of statistics, which should be available by regulations.

**Accuracy**

Most of the indicators are well known and described in detail in different sources of literature. In the cases where indicator is referring to specific variable, only the values for the key variables should be calculated. As mentioned before, we have limited number of the key variables for described surveys (up to maximum 10 variables).

The calculation of the indicator A7 (average size of revisions) is done according to the formulae:

$$A7 = \frac{1}{k-1} \sum_{i=1}^{k-1} |r(i)|$$

$X_1, \dots, X_{k-1}$  ..... values of provisional results  
 $X_k$  .....value of final result

$$r(i) = \frac{X_i - X_k}{X_k}$$

.....relative revision

**Timeliness and punctuality**

All three indicators T1, T2 and T3 are measuring time lag between two dates. We decided to use day as the time unit in the case of all different surveys of different periodicity. We also had to consider how to recalculate the values obtained directly from the definition into interval [0,1] to fulfill one of our basic rules. We decided to use the following procedure:

For each of the different periodicity we defined the upper tolerance bound where the value of the indicator should be equal to 1. Proposed upper tolerance bounds are:

- T1 (punctuality of the first release):  
 Monthly surveys: 10 days  
 Quarterly surveys: 20 days  
 Yearly (and several year) surveys: 30 days

- T2 (average time lag between the end of the reference period and the date of the first release):  
 Monthly surveys: 120 days  
 Quarterly surveys: 210 days  
 Yearly (and several year) surveys: 730 days

- T3 (average time lag between the end of the reference period and the date of the first release):  
 Monthly surveys: 485 days  
 Quarterly surveys: 575 days  
 Yearly (and several year) surveys: 1095 days

The standardized value of the indicator should then be calculated by formulae:

$$T = \frac{|T_m - T_{nom}|}{T_m}$$

T .....standardized value of the indicator

$T_{nom}$ ...original value of the indicator

$T_m$  ... upper tolerance bound

### **Accessibility and clarity**

The calculation of the proposed indicators (AC1, AC2) is based on two lists: list of all possible media that could be used for the dissemination, list of all possible means that could be used for the dissemination. The first list currently contains 7 items and the second 22 of them. The lists will probably be revised and completed in the future.

### **Comparability**

The nominal value of the indicator C2 is defined as the number of different time points in the time series for the particular variable. The standardized value of the indicator should be calculated by formulae:

$$C2 = 1 - \left[ \frac{Y}{4} \right] \cdot 0.2$$

Y...number of years in the time series

[.]...truncated value of the number

If the value obtained by the above formulae is less than 0, the value of the indicator is 0.

### **Coherence**

The indicator CH2 compares the results of the survey with the results of some reference source which should be defined in advance. The reference source could be data from similar survey, national accounts data or administrative data.

## 4 Monthly indicators

As we mentioned earlier in the case of short term surveys some of the quality indicators are calculated automatically in the process of data tabulation. These indicators are: A1, A2, A3, A5, A6, A8, T1, T2.

Some remarks:

The results of the short-term surveys are exclusively indices of different types. It is clear that indices are not linear estimates and therefore calculation of coefficient of variation (indicator A1) is not straightforward. We should also take into account that indices are generally estimated from the data of two different samples when even the same units could have different weights. Since classical software for variance estimation could not be used to deal with such a situation, we developed our own SAS macro's (the procedure uses SAS procedure SURVEYMENS) for estimation of the coefficient of variation of indices. The CV's are estimated for two different types of indices for all different domains for which the results are published.

Since all the missing (nonresponse) data are imputed the value of the indicator A5 equals to the value of the indicator A3.

The values of the indicators A6 and A8 should ideally reflect the rate of overcoverage and misclassification for the whole sampling frame. Since we only have information for the units included in the sample these two indicators could only be estimated.

Indicators T1 and T2 are not fully automatically calculated as the rest of the indicators listed above. The person who is in charge for the survey should only enter the date of the first release into the pre-defined cell in the EXCEL worksheet and the values of the indicators are automatically calculated.

The monthly indicators are automatically calculated by SAS macro's procedure and then exported into several EXCEL files but the methodologist can see them in just one EXCEL file which is linked with all the produced workbooks. Calculation of indicators is always done for the results of every month of current year. This enables methodologist not only to see the indicators for the results that should be published but also compare them to the indicators for the results of previous months. Indicators could be seen in the form of EXCEL table and could be also graphically presented by EXCEL chart.

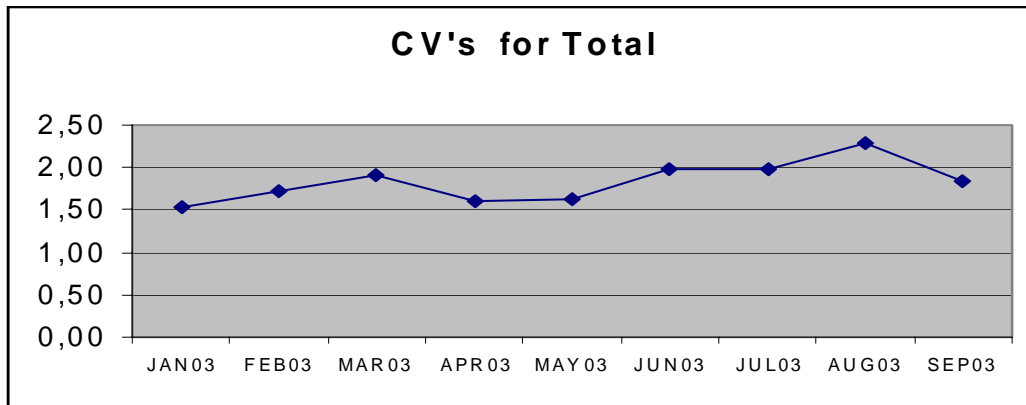
As an example we present results for the coefficient of variation for one of the indices that has been calculated in Retail Trade Survey in September 2003. The following part of the table shows indicators for three of the domains. As we already described, along with the indicators for current month, also indicators for all previous months of the current year are always presented in the table.

Coefficient of variation (percentage)

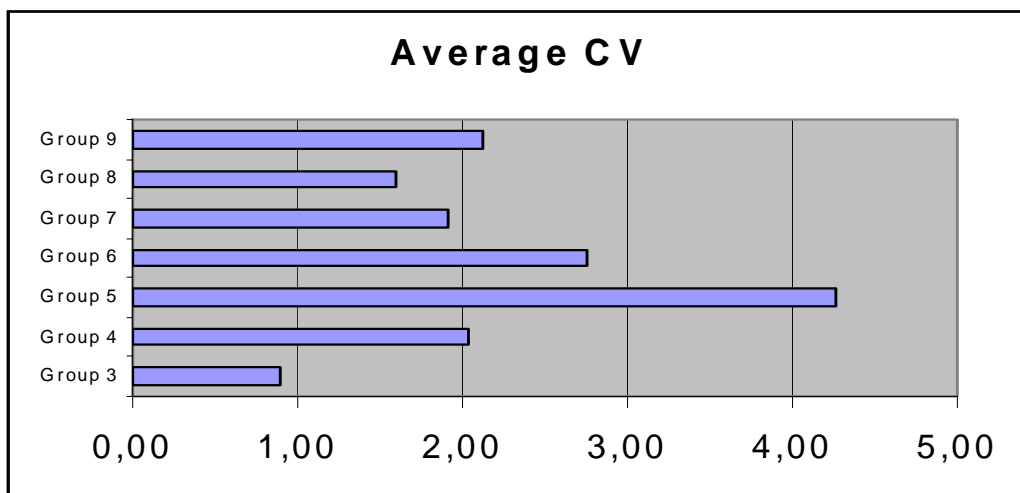
|                    | JAN03 | FEB03 | MAR03 | APR03 | MAY03 | JUN03 | JUL03 | AUG03 | SEP03 |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| <b>Total</b>       | 1,53  | 1,73  | 1,91  | 1,60  | 1,62  | 1,98  | 1,98  | 2,28  | 1,84  |
| <b>Non-food</b>    | 4,81  | 5,63  | 3,98  | 4,01  | 4,23  | 4,47  | 4,30  | 3,81  | 4,67  |
| <b>Motor Trade</b> | 2,22  | 2,91  | 2,84  | 3,58  | 2,93  | 3,43  | 3,22  | 2,92  | 2,72  |

If the methodologist wants to see results graphically presented (s)he just has to click on the command button and gets on the screen the user form where (s)he can choose between following options:

- The line charts where the results for the chosen domain for all months of current year are presented. In our example for the chosen domain Total the chart is:



- The bar chart of average CV's for all activity groups in current years. This chart enables methodologist to detect groups of activity where the CV's of the calculated indices are high and can be considered as »problematic«. In the case of Retail Trade Survey we have 15 groups of activity. Here we present chart just for 6 of these groups.



## 5 Standard Quality Report

Described list of monthly indicators is calculated each month but the Standard Quality Report should be prepared just ones a year. Therefore when all results for particular year were published as final all the monthly indicators should be summarized (usually the average of monthly indicators is calculated). Together with the rest of indicators which could be calculated only after the results for complete 12 months, they should be inserted into the database of quality indicators. As we mentioned earlier the database contains indicators for several surveys. The data in the database are organized in the way that all the indicators for particular survey and for particular year are merged together in one record. The identification of the record is the code of the survey together with the reference year. Some of the indicators are written in different forms: unweighted, weighted, nominal, standardized. In the case where the indicator refers to a particular variable, indicators for all key variables are stored. For these reasons one record in the database contains 136 variables.

Indicators are inserted into database through MS – Access interface which also contains some basic logical controls (i.e. the entered value could only be values from interval  $[0,1]$ ) The interface offers the user possibility of looking over, editing and inserting new indicators. The following picture shows part of the interface:

**ISKANJE ZAPISA**

šifra raziskovanja:

kratica:

leto:

1. USTREZNOST

R3 - delež manjkajočih statistik:

**OSNOVNI PODATKI O RAZISKOVANJU**

šifra raziskovanja:

kratica:

leto:

Zapri

2. NATANČNOST

število ključnih spremenljivk:

A1 - koeficient variacije

|        |                                    |
|--------|------------------------------------|
| A1_V1  | <input type="text" value="0.027"/> |
| A1_V2  | <input type="text" value="0.026"/> |
| A1_V3  | <input type="text" value="0.024"/> |
| A1_V4  | <input type="text" value="0.013"/> |
| A1_V5  | <input type="text" value="0.166"/> |
| A1_V6  | <input type="text" value="0.169"/> |
| A1_V7  | <input type="text" value="0.077"/> |
| A1_V8  | <input type="text"/>               |
| A1_V9  | <input type="text"/>               |
| A1_V10 | <input type="text"/>               |

A2 - stopnja neodgovora enote

|            |                                    |
|------------|------------------------------------|
| A2         | <input type="text" value="0.143"/> |
| A2_utezeno | <input type="text" value="0.152"/> |

A3 - stopnja neodgovora spremenljivke

|        |                                    |                |                                    |
|--------|------------------------------------|----------------|------------------------------------|
| A3_V1  | <input type="text" value="0.143"/> | A3_V1_utezeno  | <input type="text" value="0.152"/> |
| A3_V2  | <input type="text" value="0.143"/> | A3_V2_utezeno  | <input type="text" value="0.152"/> |
| A3_V3  | <input type="text" value="0.143"/> | A3_V3_utezeno  | <input type="text" value="0.152"/> |
| A3_V4  | <input type="text" value="0.143"/> | A3_V4_utezeno  | <input type="text" value="0.152"/> |
| A3_V5  | <input type="text" value="0.143"/> | A3_V5_utezeno  | <input type="text" value="0.152"/> |
| A3_V6  | <input type="text" value="0.143"/> | A3_V6_utezeno  | <input type="text" value="0.152"/> |
| A3_V7  | <input type="text" value="0.143"/> | A3_V7_utezeno  | <input type="text" value="0.152"/> |
| A3_V8  | <input type="text"/>               | A3_V8_utezeno  | <input type="text"/>               |
| A3_V9  | <input type="text"/>               | A3_V9_utezeno  | <input type="text"/>               |
| A3_V10 | <input type="text"/>               | A3_V10_utezeno | <input type="text"/>               |

A4 - delež urejanja podatkov

A5 - delež ocenjenih podatkov

The second database that should be filled after the completion of the particular reference year is the database of textual information on the survey. The database is organized in the similar way as the database of quality indicators meaning that all the information for particular year is stored in one record of the database and identification of the record is also the code of the survey together with reference year. Also similar interface as in the case of the quality indicators database is on disposal.

There is quite large amount of textual information that should be provided for each survey. To illustrate the nature of these information we give here just small subset of complete list of required information:

- Description of the structure of survey users. Structure should be based on standardized segmentation of users.
- Description of the target population and procedure for construction of sampling frame.
- Description of the sampling design.
- Description of the weighting system and imputation method used.
- Reasons for possible delays of the first release.
- Detailed list of all types of means used for dissemination.
- Description of procedure used for the data disclosure control.
- Where and under what conditions can user access the published data.
- Reasons for larger deviations of final results from the first results.

The final step of the procedure of preparing SQR for particular survey and for particular reference year is merging the data from both described databases into prescribed, standardized document form. The standardized part of the document is prepared in the form of MS-Word template. In the template are defined fields directly linked to the variables in the databases. When the person in charge for the survey wants to prepare the final SQR he just has to select the right record in the database (perhaps do some custom cosmetics) and save the document as the SQR.

## 6 Conclusions

The pilot testing of fully automated system of quality indicators production has been described. At the end, it has to be said that the data process system for the described surveys had been developed under special conditions which caused that this system has differed from the "normal" systems within the same office. We estimate that it is quite possible that the establishing similar system for quality control for other surveys could be much more challenging and demanding. On the other hand, very positive experience with the results that many survey managers dream about could be very intense push to develop also the other system of important surveys into the same direction.

At the end of the article let's try to summarize all the main (as we see them) advantages and deficiencies of the system. Advantages would be:

- The system enables methodologist to have quick access to the information on data quality.
- As we have established through the testing period of the system, it doesn't just enable quality control but many times enables to detect some errors in the processing system or even in reporting data.
- Standardized form of the SQR should make the whole amount of information in the SQR more readable and comparable.
- Keeping quality indicators in the same database should enable easy and effective control of the attained degree of quality for particular survey over time.

The main deficiencies of the system which should also be our main challenges for the future:

- Some of the described indicators are by definition still sensitive on revisions and completions.

- There is a danger of false interpretation of set of indicators if one analyzes them without taking into account also textual information. To avoid it, constant and quality education for persons working with these data is needed.
- Since we expect further development in the area of quality indicators inside of the European Statistical System we will probably have to adjust our system constantly. So the goal is to have the system which will be flexible enough to enable adoptions according to future development.

## References

- [1] L. Lyberg et al. – Survey measurement and Survey Quality, Wiley, 1997
- [2] Methodological documents, Handbook “How to make quality report: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003
- [3] Methodological documents, Standard Report: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003
- [4] Methodological documents, Definition of Quality in Statistics: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003
- [5] Standard Quality Indicators - Producer oriented: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003
- [6] Framework – International coordination: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003
- [7] Quality Assessment of Administrative Data for Statistical Purposes: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003
- [8] Quality Reports on Labor Force Statistics: Working Group “Assessment of quality in statistics”, Sixth meeting, Luxembourg 2-3 October 2003