STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 15
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

INDICATORS FOR THE EDIT & IMPUTATION PROCESS IN THE AUSTRIAN QUALITY REPORT SYSTEM

Contributed paper

Submitted by Statistics Austria[1]

Abstract: The paper describes the Austrian efforts and plans to describe the quality of edit and imputation within the quality report system which started in the beginning of 2002, with the concepts tested in a pilot phase in the second half of 2001. First the indicators which are built in the system are described for the aspects Editing, Imputation, Coding and Non-Response. Secondly you will find experiences and sources of difficulties which we found in the pilot phase in the second part of the paper. In the last part we will discuss the planned position of the Quality Report in our office and some aspects of the publication strategy which directly involves metadata about the editing and imputation.

I. INTRODUCTION

1. In the beginning of 2000 there was a fundamental change in the organization of Austrian official statistics. In addition, we received a new management which adhered very clearly to the principles of TQM (Total Quality Management). Since product quality is one of the pillars of this model, it was necessary to develop an instrument for reporting and documenting the quality of statistical products.

2. Almost at the same time when the need for such an instrument for quality reporting arose, Eurostat presented a concept for describing the quality of statistical products and requested quality reports for some products (SBS, external trade). So it was decided to base our quality report system on the concepts of Eurostat where quality was broken down into the well-known components of Accuracy, Timeliness, Clarity, Coherence, Completeness and Comparability. The set of quality indicators which were suitable for our purposes was the product of the discussions of a team of well-proven statisticians.

3. Since the set of information concerning the various quality dimensions is a large one, it was inevitable to develop a software tool where the responsible person for a survey can compile the report. After some discussion we decided to develop an MS-Access database which was finished at the beginning of 2001. The final product is called Quality Report Database (QRD). The QRD must be connectable to other in-house databases (for instance controlling). It, therefore, contains some administrative information as well. Along with the QRD, a manual was developed which should serve as a guideline the statisticians in filling out the QRD.

4. First we had no idea how the QRD would work for the persons responsible in the different statistical areas. Therefore we decided not to invent it with a big bang but to start a pilot phase in which we selected a couple of projects to test the concepts. The aim of this pilot phase was to see where the difficulties for the statisticians in the different fields are located. Selected for the pilot phase were the following products: Structural Business Statistics, HCPI, Structural Agricultural Statistics, Microcensus and Foreign

[1] Prepared by Thomas Burg (thomas.burg@statistik.gv.at).

Trade Statistics. As you can see, our purpose was to cover the whole field of statistics in the pilot phase. The pilot phase was complete at the end of 2001. The implementation for all of our products (excluding national accounts) is planned for the first quarter of 2002. It should be noted that we did not include the field of National Accounts in the concept.

5. The indicators contained in the QRD are not only the product of interior discussions. Since the new legal status of Statistics Austria, there exists a statistical council containing representatives of important users such as universities and research institutions. A subgroup of this council – called quality panel - is concerned with the matters of quality. During discussions with this group, the indicators of the QRD were checked to see if they could deliver sufficient information for external users as well.

6. In the following, we will first describe the parts of the QRD that deal with edit, imputation and coding. When we talk of indicators we do not necessarily mean numbers or rates or ratings. Often indicators in the QRD are short textual statements or binary information that indicates the existence of a specific processing step. There is even a form of indicator which represents a link to an external more detailed description of a certain procedure.

II. INDICATORS OF THE E&I PPROCESS IN THE QRD

7. Edit and Imputation (E&I) is embedded in the quality component accuracy and in the large field of non-sampling errors. Edit and Imputation was, until now, not a standardized procedure in Austrian statistics which means that it is done in most cases but there are no guidelines or recommendations for it. Therefore, the information about Edit and Imputation can also be seen as a survey about the existence of editing and imputation in the various statistical production fields.

8. Some parts of the processing are not directly connected to Edit & Imputation but information about it are directly related to it. So in the QRD you can find information on how the data were collected, the legal status of the survey (for instance mandatory or not), sampling information, if it was a sample survey information about the sampling survey, information about coverage problems and the nature of a possible sampling frame.

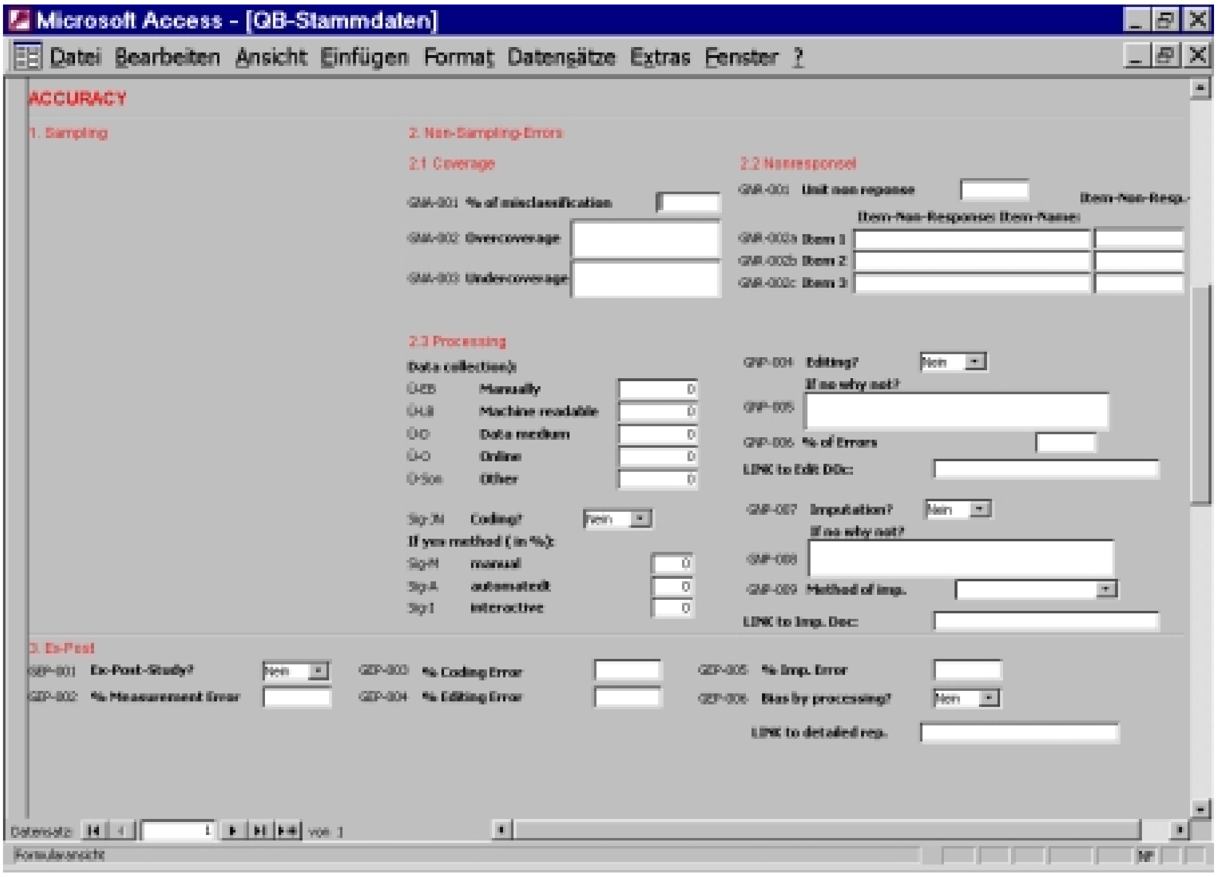
9. Non-response is one of the crucial criteria for every survey. Unit non-response is to report in the QRD as one of the main quality indicators. The definition we give for unit non-response does not include units which are not reachable for some reasons. Only the percentage of units which are in the scope but not willing to cooperate is subject of the report. The case of item non-response is more difficult because usually in an official survey you have many different items. So you have to select 3 key items for which the item non-response must be reported.

10. As mentioned before, editing is not necessarily a standard procedure in Statistics Austria. So on rare occasions, it can happen that there is no editing. However, in most cases it is done. But the problem is that it is performed in a way in which it is often not documented. So the first indicator in this field is the question if there was any editing or not. If the answer to this question is “no” we request a statement why there was none. If there is any editing, we want to know the percentage of erroneous records. The reason for this is that in most of our micro-editing procedures which are performed on mainframe, we have implemented counters for records which violates any editing rule. So you can exploit these counters to evaluate your data. The final indicator in this area is a link to general documentation about the editing procedure. As mentioned before, it is often not standardized and so with this we want to try to force our colleagues to bring down a complete documentation. Another purpose of this is to get a set of existing editing procedures in the different fields.

11. The indicators concerning imputation are relatively similar to the ones in the field of editing. First we ask about the existence of any imputation. If there is none a reason for that is requested. On the other hand, if there is any imputation we are interested to know the basic feature of the applied method. Possible answers are “model based”, “Fixed Values”, “Donor Based Methods”. In the guidelines you can find definitions for these basic categories. However, it can be the case that there are more than one

method used. In this case, the respondent should give the method which is mainly used. Finally, as is the case in the editing part, there should be given a link to detailed documentation of the imputation methods used for the product.

12. Coding can also be seen as a process step on the way to the authentic dataset. So the QRD provides also some indicators for it. So the first filter is if there was any coding. If the answer is yes we are interested in the way the coding was performed. Basically we have three modes of coding in Austria: automated, interactive and manual coding. Of course manual coding is decreasing in the different areas but nevertheless the respondent of the QRD has the possibility to split the coding activities into this 3 modes of coding. The following illustration shows the corresponding part of the QRD as it appears for the person who is expected to fill in.


Microsoft Access - [QB-Stammdaten]
Datei Bearbeiten Ansicht Einfügen Format Datensätze Extras Fenster ?

ACCURACY

1. Sampling

2. Non-Sampling-Errors

2.1 Coverage

GNA-001 % of misclassification

GNA-002 Overcoverage

GNA-003 Undercoverage

2.2 Nonresponse

GNR-001 Unit non reponse

Item-Non-Response Item-Name: Item-Non-Resp.-

GNR-002a Item 1

GNR-002b Item 2

GNR-002c Item 3

2.3 Processing

Data collection(s)

Ü-EB Manually 0

Ü-LB Machine readable 0

Ü-D Data medium 0

Ü-O Online 0

Ü-Son Other 0

Sig-JN Coding? Nein

If yes method (in %):

Sig-M manual 0

Sig-A automatedt 0

Sig-I interactive 0

GNP-004 Editing? Nein

If no why not?

GNP-005

GNP-006 % of Errors

LINK to Edit DOc:

GNP-007 Imputation? Nein

If no why not?

GNP-008

GNP-009 Method of imp.

LINK to Imp. Doc:

3. Ex-Post

GEP-001 Ex-Post-Study? Nein

GEP-002 % Measurement Error

GEP-003 % Coding Error

GEP-004 % Editing Error

GEP-005 % Imp. Error

GEP-006 Bias by processing? Nein

LINK to detailed rep.

The Part for Non-Sampling errors in the QRD

III. THE EX-POST STATEGY

13. The indicators described above are mainly focused on the principal question of the existence of edit & imputation and to a certain extent on data quality. That does not give any information about the quality of the process itself. Nevertheless, we want to evaluate these kernel procedures.

14. For this purpose, we have planned to perform ex-post surveys to check how valuable the procedures are which were performed during processing. This is a completely new approach for our office. Therefore, we excluded the request of filling in the corresponding fields in the pilot phase because we did not have the resources to perform ex-post studies in the several pilot statistics.

15. The indicators of the ex-post area are mostly simple to understand. We see the main purpose of such an ex-post study is to compare in a small sub-sample the real values with the values you generated with a certain processing procedure. So the indicators contain the percentage of wrongly coded, edited and imputed values. To complete this type of indicator you can also find a percentage of wrongly measured values.

16. Of course for some users it is of analytical interest if the applied procedures have yielded to any bias in some estimates. So, in the QRD a binary indicator related to this subject is included and again a link to a more detailed report of the ex-post study.

17. One of our next tasks will be to elaborate guidelines and practices on how to select a suitable sub-sample for the purpose of an ex-post study. In official statistics you have areas that are very different from each other. So it is clear that a layout for an ex-post study will have to differ depending on the subject of interest

18. A problem in performing an ex-post study lies in the fact that by this the total costs of a product will increase. Therefore it seems necessary to bring over the benefits for quality assuring measures to the user and along with that to our financiers. This is important because since Statistics Austria has a new legal status we have the obligation to assure our costs. Of course we can justify ex-post studies for some very important statistical products on the basis of the importance to improve the quality of the processing steps over time for these products, but nevertheless, the cost factor will play a more and more important role in the future of official statistics and not only in Austria as it seems.

IV. EXPERIENCES OF THE PILOT PHASE

19. The first aim of the pilot phase was to determine if there are difficulties for the statisticians in the different areas to fill in the QRD with the delivered guidelines. On the other hand we could receive some substantial information about possible deficiencies in the fields of editing & imputation.

20. In the non-response part, as we expected, the main difficulties concerned the question which key variables should be selected to report the item non-response. Unit non-response represented some definition problems because it was not clear which rate is exactly requested. As a result of the pilot study, it turned out that it would be useful to have a field where it can be stated if there is a non-response adjustment other than imputation (such as weighting with an eventual post-stratification), despite this information being located somewhere in the sampling area.

21. For the information concerning editing, it was found difficult at some pilot projects to determine the percentage of deficient data records. The reason for this lies in the fact that the editing in large-scale statistical projects runs in different steps. So the definition when a record is in error is not homogeneous over all statistical products. If you say in general that a record is wrong, if you fine any edit rule in any procedure you will find that this does not fit the needs of interpreting the quality of your data material. So we learned from the pilot study that this definition must be adjusted to the particular statistical products you have in the scope. Another problem was the question of documentation. It was often the case that some parts of the editing procedure were documented and located on several places. Additionally some main know-how was found in the orders which were given from the expert in a specific statistical field to our EDP department. In order to have complete and standardized documentation, we shall have to combine several components available at various parts of the statistical office. Since large data sets were used in the pilot project, at least few edits occurred in each of the data sets.

22. For imputation, the situation seems a little more structured than in editing, so in this area we made some methods available which are relatively transparent for the colleagues. So it was clear to the statisticians of the various departments in the pilot projects how to categorize the methods. Concerning documentation you can say basically the same for editing. The only difference is that the process of imputation is more centralized so the collection of the necessary information is easier. In some statistics the methods used are not very sophisticated (so it is a common practice for instance in SBS to use values from the last survey). It is imaginable that based on the results from QRD there will be an improvement of the quality of the used methods.

23. The fields for Coding were understood very easily in the compilation of the QRD in the pilot phase. It can be said that the amount of purely manual coding is decreasing but nevertheless in some fields you can find it. The invention of automated or at least interactive supported coding in this area should be a direct result from the results in this field.

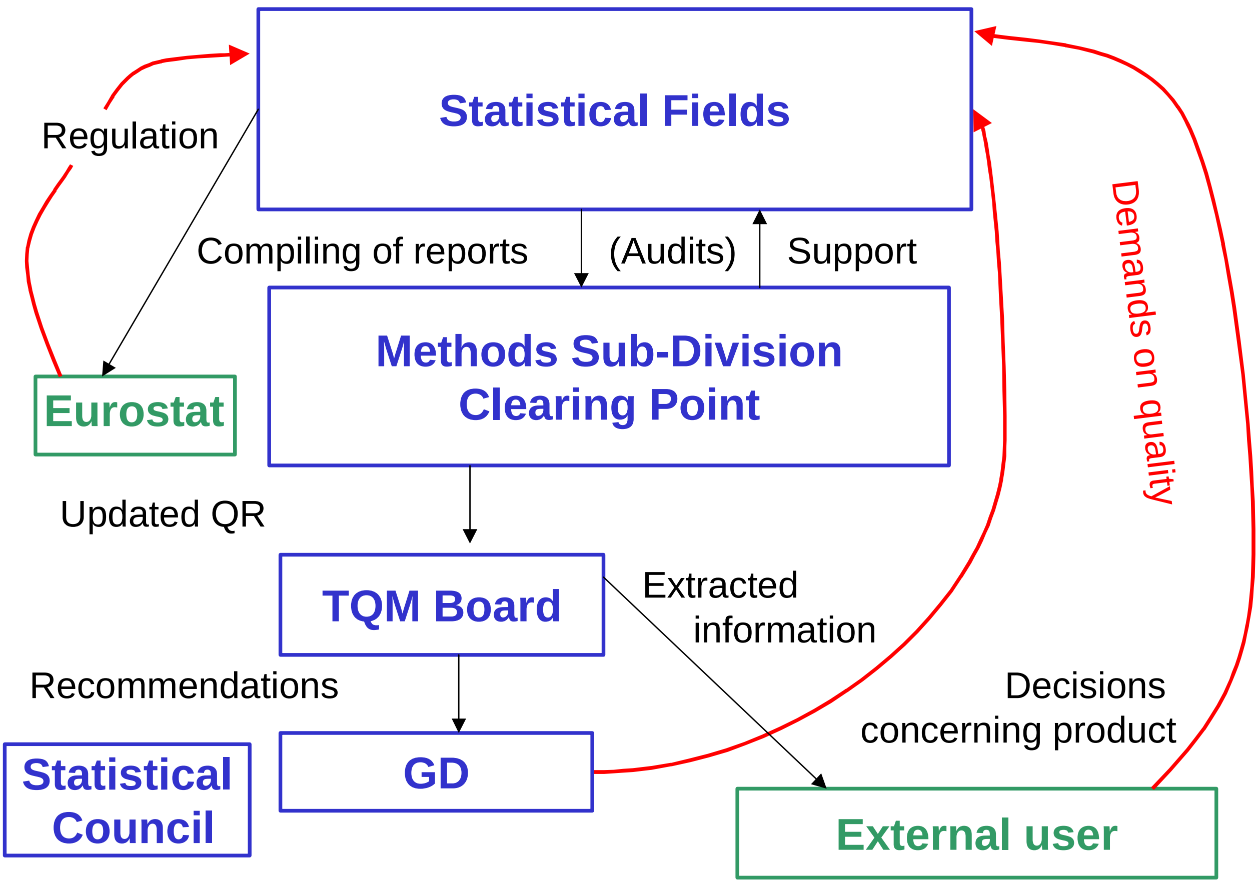
24. In discussions with the colleagues compiling the QRD in the pilot project, it turned out that they wanted to give additional textual statements to the single points in the QRD. This brought us to the decision to put in an additional text field for each quality aspect where such statements can be filled in. So if there is the necessity to add something additional to Edit & Imputation it can be done under the point "Additional statement concerning Accuracy". For technical reasons this statement must be very short. We also do not put it in the subchapters of the quality dimensions. In conclusion, we can say that we hope that the desire to say something more will yield to the achievement of complete and standardized documentations in the field of edit & imputation.

V. POSITION OF THE QRD AND PUBLICATION STRATEGY

25. The QRD for a certain statistical product is compiled in the directorate responsible for it. Then it will be transmitted to the methods department. After discussion with the methods department, the QRD will be updated. The finalized QRD can be made available to certain decision-makers as there are the directors general of our office or the TQM-board: a panel of several people who are responsible for the activities regarding Total Quality Management. This panel can also serve as the point where projects which aim to improve the quality of some products will be launched.

26. Obviously, one of the main goals is that the results of the QRD is to locate deficiencies in the applied processes for statistical products. So the method department will have to evaluate the QRD especially in the fields Edit & Imputation because most of the know-how of best practices is located there.

27. Eurostat requests special detailed quality reports for more and more statistical products. So of course the QRD should also serve as an information provider where all the necessary information for fulfilling the obligation to Eurostat can be extracted. In the above-mentioned quality panel of the statistical council we also discussed very detailed quality reports of some statistical products along with the first completion of the QRD. The following picture gives an idea of the planned position of the QRD in the Austrian Statistical system



28. Along with the QRD there will be a large set of metadata. So of course we want to transmit at least a part of the QRD to our users. This is also one of the reasons why we take care in the completing of documentation of the single process steps. In our statistical law from 2000 we also have the obligation to put results onto the internet. For this reason we will make results available from the QRD on this platform. Of course we must distinguish between information which is only for internal use and indicators and descriptions which are necessary to understand the impact of processing to the data and the resulting estimates. Translating the concept to the special case Edit & imputation you have seen that there is no explicit description of Editing and/or Imputation within the QRD. However we have built in a link to the documentation for each process. So if you imagine the normal procedure of internet you can have basic results of a certain statistical product as portal and from there you can be directly connected to the corresponding QRD. If you are interested in more detailed documentation of Edi & Imputation you can activate the corresponding link in the QRD

V. CONCLUDING REMARK

29. As you can see we are in the heart of this big and ambitious project now. Some of the described concepts are maybe part of revisions as we find out that they or not suitable in practice. But basically, we can say that in the pilot phase, the principles were accepted very well. We now agreed in our TQM-Board and with the quality panel of the statistical council that we want to have the whole structure of QRD together with our new publication strategy in running mode at the end of this year.