WP. 2
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (i): Development and use of data editing quality indicators

PROCEDURES TO IMPROVE THE DATA-CLEANING PROCESS BASED ON QUALITY INFORMATION

Invited paper

Submitted by Statistics Austria [1]

Abstract
First the experiences of collecting information on the data cleaning process in the Austrian Quality Report system are described shortly.
Due to the experiences and information received from our quality database (QRD) some improvement projects were launched. The second part of the paper shows based on the example of the Austrian Labour Force Survey how these improvements were established. Different problems during these projects due to the connections and trade offs between various quality dimensions were accompanying this process.
The 3rd part of the paper gives a prospect to useful organisational models that are in discussion in STAT now to use quality information efficiently for optimising data cleaning.
Finally a possible method of measuring the quality of the data-cleaning process as a whole to make it comparable over time is described.

I. INTRODUCTION

1. Quality of statistical production processes became a topic during the year 2000 when TQM was selected as management background for quality issues at Statistics Austria in general. Since Product Quality is one of the piles of TQM one of the tasks was to establish a well proofed quality report system. This was done in the years 2001 to 2002. Now we are in a position to give first feed back on the conclusions we could make after collecting the information.

2. The increasing demands for metainformation especially related to Editing and Imputation is not only a desire originated from the producer but also on the side of our customers.
Since Data Cleaning for sure is understood as one of the cores in the production of official statistics questions concerning the effects of the data cleaning process are raised more often than in the past.

3. You can distinguish between two general groups of indicators concerning the data cleaning process. Such ones who are describing the quality of your data like basically the amount or percentage of erroneous records. On the other hand and often more important for the producer you should be able to say something about the quality of the data cleaning process. It is mainly this question which will be treated in this paper in a sense how this quality is related to the aspects of organisation and management.

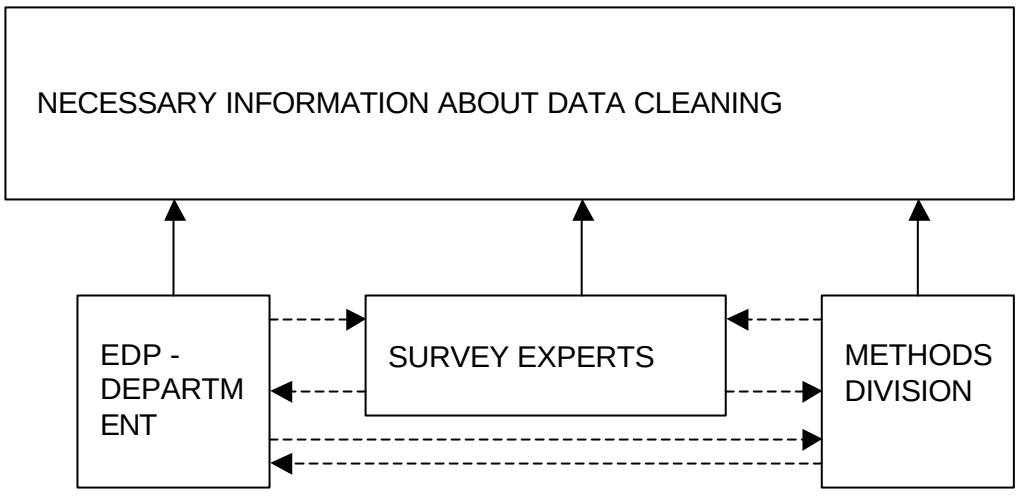
[1] Prepared by Thomas Burg (thomas.burg@statistik.gv.at).

4. One part of our Quality report system consists of a database called QRD (See also [2],[3]) in which the different indicators related to the well known quality dimensions are collected. During the primary collection for this database we soon realized that especially in the non sampling area there is some potential of improvement. So we decided to investigate the effects of data cleaning on the authentic data files. On the other hand and along with this we also wanted to analyse and decompose the various steps of this process.

5. The QRD is not the only part of our quality report system. Secondly we use for important products more detailed quality reports in a textual form. Of course these reports contain a detailed description of editing and imputation. A more or less critical statement of the way the data cleaning is organized should also be within this description. However the problem is that self-critical statements produced by an expert and published in a document are not so favoured.
The detailed quality reports are also used in a series of feedback audits which took place on the demand of a subgroup of our statistical council. Of course the data cleaning process of the eyed product was one of the key questions.

II. COLLECTION OF INFORMATION ABOUT DATA CLEANING

6. When we started to fill our database I often heard the statement:" I don't know how to say something about it" or "Yes we have some possibilities to find out how many records are erroneous but it is relatively difficult to do that". These are only extreme examples of answers you receive when you try to collect more or less standardized information. Generally spoken you can say that the information you receive is not the one you want to have. So one of the first tasks you have is to transform the information and/or compose it into a understandable and usable form.



Model of Information flow when collecting information about Data Cleaning

7. There also often was the problem that you do not get the whole information from one person alone (most of the times the person responsible for the survey). It is often the case that the

knowledge of the data cleaning process is spread between different persons who can even be in different organisational units. So sometimes the collecting of the whole description of editing and imputation procedures is a little bit of a puzzle. Not seldom it was the case that the main information source about the edit and/or imputation was not the expert or the survey manager but rather the EDP department. Along with this phenomenon it can happen that an expert from the EDP is leaving the office (for instance retiring) but his programs are still running and are used by the colleagues in the subject matter division. So we had the situation that we had to analyse programs (often very old, even assembler) to get a clearer view about the things going on.

8. So the collection process for our dataset was sometimes time consuming because of this hidden and sometimes even vanishing information. As a consequence from these problems we decided to put some energy in persuading the experts in the subject matter divisions to analyse their data cleaning process steps more exactly and to oblige them to give more detailed documentation. This is very important and it paid off that we decided to have links in our database to more explicit information. If the links had to be filled out as void we could argue that this was a loss of quality for the whole product because the increasing demand of data users after metadata could not be satisfied sufficiently.

9. So we can say that one of the first consequences of our quality report system was to launch an inhouse campaign to have the whole production process fully documented and along with this naturally as one of the main parts the data cleaning steps.

10. During the collection of the information and after analyzing the indicators in some fields it became quite clear that something has to be done. As I told before some departments did not even exactly know the methods which were exploited. You can imagine that these "methods" sometimes were not very sophisticated. So there was a large potential for improvement in some areas in building in procedures which are nearer to common standards.

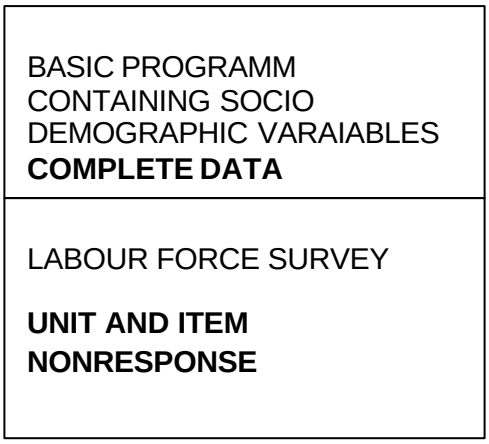
11. The way of this information collection process and some results of the quality indicators you can find in the QRD started in some areas a complete think over on the questions of data cleaning. This can also be seen in the background of a new strategy concerning human resources since the organisational and legal change of our statistical system in the year 2000.There was an increase of our academic staff. This was for sure one of the most important prerequisites to launch useful improvement projects in the field of Editing and Imputation.

12. To have academic persons in your office is one thing but the other side is that you have to train and make them familiar with the methods used in official statistics. Our academic staff is often educated in the field of economics and/or social sciences. Of course there they had some statistical education. But the methods used in data cleaning are most of the time not included in that.

13. Concluding this section we can say that in some fields we were able to invest energy to improve the quality of our data cleaning steps. For instance this happened in the Foreign Trade statistics, the Short Term Business Statistics and some products form population and social statistics. In the following section it will be reported how this process took place at the redesign of the imputation procedure of the Austrian labour force survey.

III. THE EXAMPLE OF THE AUSTRIAN LABOUR FORCE SURVEY

14. The Austrian Labour Force Survey in its current form is performed since 1995. It is embedded in the microcensus a quarterly sample survey containing 1% of the Austrian population. This survey contains a basic program which is mandatory for the respondent. The second part is a special program on which every respondent can decide if he takes part or not and in march of every year this special program serves as Labour Force Survey.



Structure of the LFS embedded in the Austrian Microcensus

15. The structure of this survey gives us a situation where you have the basic program containing many socio-demographic variables like sex, age, occupation, education etc. filled out with high quality. On the other hand for the part containing the labour force variables non-response is a topic. Since the filling out of this part is completely up to respondent there is a certain amount of unit-non response. Usually the rate of unit-nonresponse is form 9-11%.and due to the fact that the LFS is relatively complex survey takes a relatively large time and there and consists of some sensitive questions we have also to deal with item-nonresponse which amounts sometimes up to 20%.

16. In other special programs unit- and item- nonresponse. were not treated as that critical and no imputation procedure was performed. But for the LFS the situation was different because EUROSTAT demanded the delivery of completed records in the same amount as the co-financed sample size was. So there was the necessity to perform imputation.

17. Due to the situation described above it was obvious to use information from the basic part of the microcensus. After some considerations it was decided to use a donor method exploiting a distance function based on a subset of the socio demographic variables contained in the basic program.

18. Of course we did not use all of the variables available from the basic program. So we selected a subset of only few items and built a weighted distance function. This method was applied to the whole LFS-records and used not only at item- but also for the unit non response. I do not want to describe the method fully here. If you want to know more details you can look at [1].

19. So the situation was that the method division received an order by the survey experts to develop an imputation method to come along with the necessity of delivering completed records. After that the method division developed a program which fulfilled the task. The only contribution of the survey division was which variables should be built in in the distance function and what weight should be assigned to such a variable. The method itself was a complete black box to the experts in the subject matter division. Even the running of the program were not done by them.

20. The imputation was executed year by year when the LFS was processed in our office After he results were available there was only a check of one dimensional distribution of several items. There was no check of multidimensional tables or any detailed analysis between imputed or non-imputed values.

21. That was going on until the LFS 2002 was performed. At this time due to the reasons I described in Section II there took place a deep analysis of the results of the LFS 2001 before the processing of the LFS 2002 started. There it was found out that some results where peculiar and the method division was asked to give the new hired good qualified staff a deeper insight into the methodological aspects of imputation.

22. Now the planning process was quite different then 1995. There took place a detailed analysis of the different parts of the LFS. According to the results of this analysis and after consulting the methods division they decided to use this time a more simple imputation method based on hot deck algorithm. The sort used in this hot deck was different for each variable group.

23 The imputation procedure was performed stepwise. After each variable group which was imputed a check of the results took place. This time not only one dimensional but also by cross tabulating several items. The methods division was only involved when there were algorithmic or methodological questions concerning hot deck.

24. The Donor Method applied in former years was only used to generate records in place for the unit nonresponse. This was done in multiple runs testing various distance functions.

25. Retrospectively spoken this procedure of very detailed work on the imputation procedure has for sure increased the quality of the authentic dataset. But to be honest there are also some problematic aspects if you do not focus on the dimension of accuracy only.

26. **Timeliness**: The procedure described above was extremely time consuming, So if you launch an improvement project in the field of data cleaning it is obvious that for the period you do this you have to calculate with delays . A possible solution to this can be that you work simultaneously with the old method while you are developing the new. But this is most of the time as you may expect a problem of resources.
The Time delay in our example was about 4 month and users got a little impatience. However it is to expect that you only have this delay once when you are working on the redesign of your data cleaning.

27. **Accessibility and Clarity**: Our experiences have shown that the Clarity was for sure increased by improving the imputation procedure. Since there is now a group of people located near the considered subject who is well informed about the methods and procedures which were applied during imputation they can give better answers to questions from users related to this topic.

28. **Coherence**: During their investigations the colleagues found out that there was incoherence to other sources (for instance Labour Market Data). This differences were at some items significant stronger in the imputed values. One of the goals during the new imputation was to minimize these effects.

29. **Comparability**: This is always a problem, because if you change your method you wont be able most of the times to avoid breaks in your time series. We also had in the LFS some problems with that and had to deliver arguments to the users why there are some changes for a certain amount of estimators.

30. Finally spoken it can be said that it was worth to invest some energy in overdoing the imputation process for the LFS because as you can see there are more benefits than losses when you go systematically through the different quality dimensions. The graphics below tries to illustrate this.

.

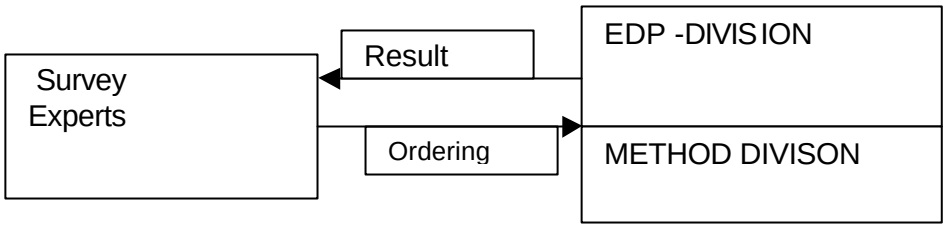
POSITIVE EFFECTS	NEGATIVE EFFECTS
ACCURACY strong COHERENCE partial CLARITY partial ACCESSIBILIT: A little	COMPARABILLITY Sometimes strong TIMELINESS once

Influence on the quality dimensions of the LFS improvement project

IV. THE MANAGEMENT ASPECT

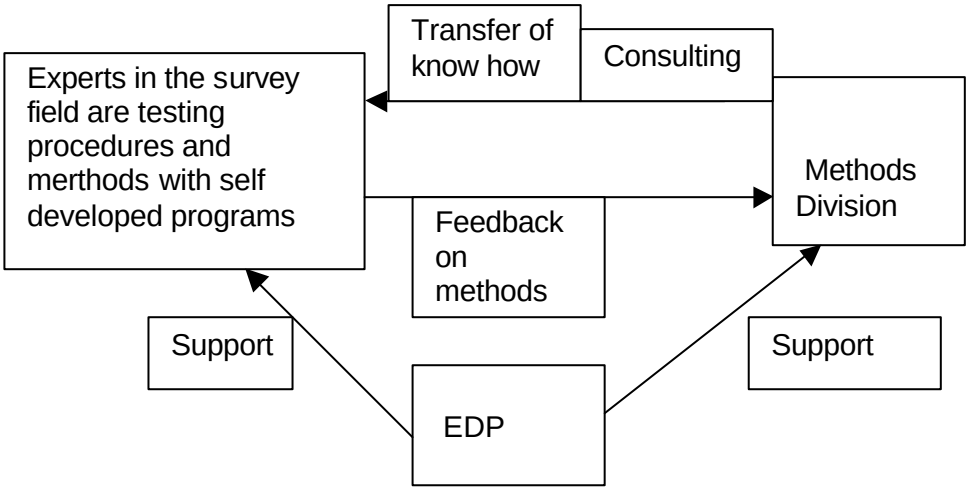
31. After the experiences made at the improvement project we can now analyse what are the necessary ingredients and the motivation for launching an improvement project. We also want to raise the question what is an optimal or at least feasible environment for a successful outcome of such an undertaking.

32, First off all there must be useful organisational structures which allow it easily to change things and to find appropriate ways for getting insights in the work of service departments.
A situation which was often the place probably due to the form electronic data processing developed was the following:



Management of data cleaning development before 2000 at Statistics Austria

As you can see the arrows are unidirectional what means that there was no automatic knowledge transfer from the methods and/or EDP to the survey management. So as a first step you must have a change in this structure to come to a more decentralised model in which the knowledge of the methods can be transferred easily to the experts dealing with the survey. So the situation we desire should rather be like this:



New decentralised management model for data cleaning development

This simplified graphics shows that the main part of the specified know-how should lie in the subject matter division. If they need methodological support there should take place a transfer of knowledge to the survey experts. What means that the methods division is only a consulting unit. The development of tailor made programs for their special problem lies in the responsibility of the survey management. And the role of the EDP-department is limited to supporting the colleagues in special questions concerning technical software problems.

33. To achieve such a situation it is of course necessary to improve the qualification of our staff what is for sure not a process which can be completed in a short time period As it was mentioned earlier it is not done with the increasing of the academic staff. It is also necessary to train the people and make them familiar with methods of data-cleaning.

34. The impetus of changing the organisation must also come from the desire to make more detailed analysis of the data. So the knowledge of the methods of the data cleaning process and the self development of tailor made software solution by an expert in a certain statistical field enables him to enlarge his own expert knowledge and gives him the possibility of answering sophisticated questions by experienced users. In the past it was often the case that questions which concerns methodological issues about the data cleaning process were transmitted to the method division. There again we had the problem that the staff there was in fact familiar with the methods applied but special knowledge about the considered survey or the specific statistical product was missing. The new model ensures that there is sufficient combined knowledge concentrated in one organisation unit.

35. It is necessary that the changes in the working process must be supported not only by the staff involved but also by the management up to the highest level. That will also ensure that staff members who are very long on board and used to the older form can be motivated to change their working philosophy. But this cannot be done in an authoritarian way. Rather is it necessary to show those people that the new way means a lot of job enrichment to them.

36. Having now changed the structure of work we can go to the question how an improvement project in the field of data cleaning can be managed. The LFS experience showed us that we were missing a plan for this improvement project. As mentioned earlier it was a very time consuming process and in some moments a detailed project plan was missing.

37. So retrospectively spoken you can say a project of analysing reorganising and improving your data cleaning process shall include certain steps which must be manifested in a written project plan and be bound to determined timetable. These process steps listed below should be documented in guidelines.

1 .Nomination of Project Team
-Distribution of tasks

2. Analysing of the actual situation in the data cleaning process

3. Discussion of new methods
- What is state of the art, Study of methods used elsewhere
- Consulting by methods division
- Selection of suitable methods
- Education of staff

4. Implementation of new method
- Decision about software
- Tests of results

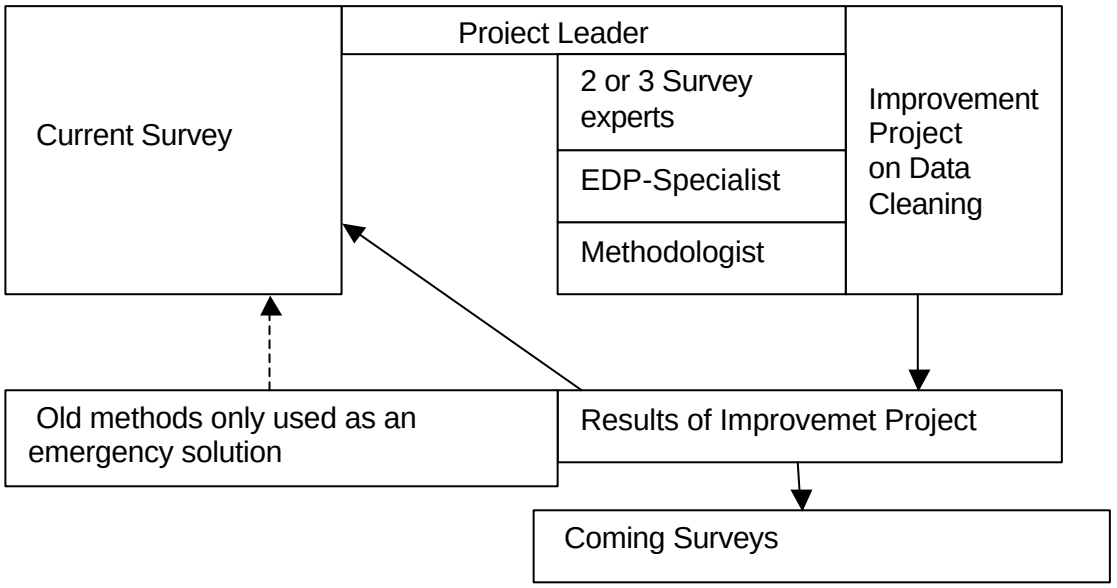
5. Documentation of new methodology
-Decision of publication strategy

Steps for an improvement project for the data cleaning process.

38. Another question is who shall be involved in such a project. For sure there must be a project leader. She/He should be a person relatively high in the hierarchy and should be responsible for all management tasks, i,e, .timetable, resources and a like. The team should consist of two or three persons form the subject matter department one methodologist and one person very sophisticated with programming and algorithmic problems from the EDP department with sound knowledge of the computer environment in the office. You should keep in mind that you do not involve too many people. From the 2 or 3 experts from the subject matter department one should be dealing with the market study. During the implementation phase the main work should be done only by the experts in the subject matter division. They should have been enabled for that in the discussion process before after education and consulting by the methodologist and the EDP-specialist. Of course this need not to be that strict because as you will expect there will be problems or situations where the help of a methodologist or a professional programmer is useful and necessary. This will for sure also happen when test results are discussed and for instance interpretation of curious results must be done.

39. In regular project meetings the project leader should control the time table and should provide additional resources if necessary. The role of the project leader does not only involve the improvement project alone. She/He has also to keep an eye on the fact that there must go on the

regular production process. The implementation phase is for sure the most critical point for this problem. If it is recognised that there are problems which threaten the timetable to much it must be ensured that you can despite of this fact provide results for the current survey.



Structure of an improvement project

40 Now at Statistics Austria we are trying to implement this structure and the project steps described before when establishing an improvement project.

V. EVALUATION OF THE DATA CLEANING PROCESS

41. To ensure that things around the data cleaning process are well planned it seems necessary to have a tool with which you can evaluate the situation. Evaluating a process like data cleaning means like in many other things that you should try to decompose the quality you like to look on into several dimensions. In the case of data cleaning it seems to us that there are 3 aspects which contribute to the overall quality. From the things in the above section it is clear that there are aspects concerning the organisation and the management of the process. Secondly you should evaluate the technical aspects around the data cleaning and as a third dimension there should be of course evaluations concerning the effects on the results of your statistical products.

42. We are now in a planning phase to create a useful checklist which we want to distribute among the survey managers. Despite such a data cleaning checklist should give a good feeling about the quality of your process it is also necessary that it is not too long because you don't want to burden the survey manager too much and on the other hand if it is too long it is not suitable to analyse it

43. Below there is a list of possible questions such a checklist should contain.

- MANAGEMENT AND ORGANISATION
  - Are the Methods of Data Cleaning well known in your division?

  - How many people have sound knowledge about the Data Cleaning in your division?
  - Are your methods approved by the methods division?
  - Do you have contact with other offices/organisations and compare your methods with theirs?
  - When did you perform your last improvement project?
  - Is your Data Cleaning Process fully documented?

- TECHNICAL ASPECTS
  - Is your Data Cleaning process fully automated?
  - Who developed the programs which run the data cleaning process?
  - How much support did you need from the EDP?

- DATA AND RESULTS
  - When did you perform your last ex-post study to evaluate the accuracy of the cleaned values?
  - Do you know on the effect your data cleaning has on the variance of your estimators?
  - Did you test your methods with a simulation study?

44. Maybe this list is not complete but the amount of questions should not increase to much.
As next steps we will form a checklist out of this questions (maybe a few more) and we will try to survey every statistical product in our office.

REFERENCES

[1] Burg "IMPUTATION FEHLENDER WERTE IM LABOUR FORCE SURVEY"
Österr. Zeitschrift für Statistik Jahrgeng 1996 Heft 2 (GERMAN only)

[2] Burg "THE IMPLEMENTATION OF IN INHOUSE QUALITY REPORT SYSTEM"
Paper presented at the Working Group "Assessment of Quality" 2002 I in Luxemburg

{3[Burg "INDICATORS FOR THE EDIT & IMPUTATION PROCESS IN THE AUSTRIAN QUALITY REPORT SYSTEM"
Supporting paper at the UN/ECE Work Session on Data Editing 2002 in Helsinki

[4] EUROSTAT "STANDARD QUALITY REPORT"
Paper presented at the Working Group "Assessment of Quality" 2002 I in Luxemburg

[5] Wein " THE PLANING OF DATA EDITING PROCESSES"
Invited paper at the UN/ECE Work Session on Data Editing 2002 in Helsinki