WP. 14
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (ii): Implementing editing strategies and links to other parts of processing

INTRODUCING AND IMPLEMENTING A NEW DATA EDITING STRATEGY

Supporting Paper

Submitted by the Federal Statistical Office, Germany[1]

I. INTRODUCTION

1. Statisticians face today a lot of developments:

- Since a few years there has been an ongoing variation of statistical results. Rapid developments induce an increasing demand for actual and reliable enough statistical results. On the other hand, the increasing individualisation of societies leads to individual problems and thus requires specific statistical analysis. Users have nowadays the IT tools to do their own analysis on the basis of high accurate statistical micro data.
- The statistical offices have to satisfy the increasing demands with fewer resources. Thus there is a need for a careful transformation of users' demands in a data editing strategy and the use of efficient data editing methods. The trade off between accuracy and timeliness was expanded by the factor "resources".

2. Destatis, the Federal Statistical Office of Germany (FSO), implemented a new data editing concept in July 2004.[2] It:

- induces the introduction of three new IT tools for the planning and management of data editing with strong effects on the specification of metadata and management of data editing processes,
- implements new data editing methods to be prepared by a practical test of a combined selective and macro-editing method in combination with an automatic error determination and correction,
- introduces a new training concept for data editing.

The concept is planned to be expanded to all statistical offices of the Länder in the first half of 2005 (decided by the head of offices in November 2004).

3. The introduction of modern data editing methods represents a remarkable paradigm shift of the German data editing tradition: manual editing of all erroneous records will be more and more replaced by selective manual and automatic editing – depending on demands on statistical results. The numerous changes and activities induced by the concept will clearly overstrain subject matter units. Thus the introduction of the concept in a decentralised statistical system requires a systematic preparation. The aim of this contribution is to describe how statisticians were made familiar with the new editing strategy and how parts of the new editing strategy were tested by subject matter statisticians.

---

[1] Prepared by Elmar Wein, elmar.wein@destatis.de

[2] The new data editing concept was developed by a national joint working group of the FSO and the statistical offices of the Länder.

II. CHANGES INDUCED BY THE NEW DATA EDITING CONCEPT

4. The changes induced by the new data editing concept affect the planning of data editing, the data correction, and the general methodology. Due to limited resources the optimisation of data editing processes will only be supported after the introduction of new data editing methods at the present time. The concept was realised via IT tools.[i] The changes as regards the planning of data editing are:

- the introduction of an electronic guideline for the collection and judgement of relevant information (new IT tool)
- the provision of predefined, adaptable process chains for data editing as part of the data gathering process and data processing
- the introduction of the PL-Editor with an office wide database for the specification of checks (new IT tool)
- the introduction of two process managers for the planning of data editing sub processes (new preliminary IT tools) and
- the supplementation of the existing data editing methodology, e.g. a collection of methods used for data editing, an actual error scheme, methodological considerations as regards the specification of checks and their assignment to the data gathering process and data processing, or descriptions of data editing processes.

5. The data editing process will be affected by:

- the introduction of new data editing methods, e.g. selective and macro editing methods, automatic error determination and correction as part of the data processing
- the introduction of two process managers for the management of data editing sub processes and
- methodological considerations as regards the management of data editing.

6. The changes documented in the previous two sections affect subject matter units and several IT units (IT production management and development of Internet questionnaires) in all German statistical offices. The fact that there were no recommendations as regards the planning and standardisation of data editing processes makes the real dimensions of the changes obvious: they require a change of the management culture that means the replacement of individual best performance by a systematic processing. In addition to this the new IT tools are very user friendly but in spite of this fact some users need to be trained. Another not negligible aspect is the fact that some statisticians think that they deliver exclusively excellent results if they correct all erroneous records ("bookkeeping culture"). The common opinion that only expert knowledge guarantees high quality statistical results is an additional contra-productive factor which clashes with the planned introduction of automatic editing methods. Thus one task was to convince managers and subject matter statisticians so that they accept modern data editing methods.

III. INTRODUCING THE NEW DATA EDITING CONCEPT

A. Convincing Statisticians of a New Data Editing Strategy

7. The introduction was developed by a steering committee which consists of IT specialists who developed the new IT tools, methodologists who are responsible for the data editing concept, and experienced data editing lecturers who developed the new training courses.[3] The committee proposed an introduction strategy that was approved by superiors.

8. The great majority of German statistics is performed in a decentralized way but with one survey specific and unique data editing strategy. In general the specification of checks and the planning of the data editing strategy belong to tasks of the Federal Statistical Office that is consulted by the statistical offices of the Länder. The statistical offices of the Länder collect the data and perform data editing in accordance with the given data editing strategies and specified checks. A minority of statistics is

[3] The author expresses his thanks to: Andree Hähnel and his team from the company WERUM, Norbert Glaser, Carsten Kuchler, Karl Günter Köhler, Hans Joachim Schwamb, Alfred Steilen, Volker Stutzer, Corina Teichmann, Tatjana Theis, and the numerous colleagues of the FSO (around 20 persons) who contributed to parts of the data editing concept.

performed exclusively by the FSO. Thus the primary interest of the FSO is focussed on methods and IT tools used for the planning of data editing, especially on the specification of checks. Opposed to that the statistical offices of the Länder are interested in effective counter activities to prevent errors and new methods and IT tools for data correction. Consequently a successful implementation strategy should support these main interests. Thus the steering committee chose two focal points: the specification of checks with the PL-Editor (FSO) and the test of new data editing methods (statistical offices of the Länder).

9. Besides these specific interests all statistical offices have to realise budget cuts and reduction in staff since few years. The reduction in staff concerns all sections of the offices inclusive the subject matter units – one focus group of the introduction strategy and concept.

10. Based on this central decision the steering committee concentrated its activities on the respective IT tools and methods that were actively managed. The other tools, e.g. the electronic guideline for the collection of relevant information and the process managers were also provided and trained but the subject matter units are solely responsible for their employment.

11. Subject matter statisticians with a specific (mostly non universal) degree of training specify the checks and use selective and macro editing methods. The IT production manager consults subject matter statisticians as regards the data processing of a survey. In addition to this she uses the specifications of the PL-Editor. Both types of addressees form the most important target groups because subject matter statisticians should be convinced of the benefits provided by the new IT tools and methods and the IT production managers have to integrate these methods in their work flows. Due to the central role of the IT production managers it was decided to train them intensively before the training of the subject matter units starts. Thus the IT production managers are able to coordinate and consult subject matter units as regards the specification with the PL-Editor.

12. The data editing concept was realised by IT tools which (supposed to) support relevant activities as regards data editing that means subject matter statisticians apply automatically the data editing concept by the use of IT tools. The main demand on every new IT tool was the realisation of a significant value added e.g. more powerful tools for priority setting among erroneous data, the multiple use of existing metadata, more reliable information for the management of data editing processes, or the reduction of the coordination effort.

13. The data editing concept, IT tools, materials, training courses, and the concept used for the implementation are very innovative and the reaction of the respective addressees was unknown. Thus members of the steering committee tried to get feed back from potential users or data editing experts before the implementation of the concept started: the PL-Editor was internally tested under real circumstances, a pre-training of the PL-Editor was organised, parts of the concept were presented at international conferences, and the selective editing method was carefully developed and simulated.

14. To facilitate the conviction of the personnel an intensive information campaign before, during and after the successful introduction of the concept had taken place. One year before the implementation (03/2003 – 03/2004) short contributions were published in the internal journal of the FSO: an overview of the concept, presentations of the PL-Editor and Data Editing Intranet, a report on the simulation of a selective editing method and a contribution on automatic editing.

15. Next to the implementation workshops and presentations were performed in a top down approach: first an overview was presented at the monthly meeting of the heads of departments and subsequent workshops were announced. In addition to this the heads of departments were asked for promotion. After this presentation the workshops were announced which were addressed to managers and specialists. Focal points of the two intensive workshops for managers in June/July 2004 were the influence of new data editing methods on the production of statistical results, the underlying principles of a selective and automatic editing and the demands on specialists. The three workshops for specialists focussed the features of the new IT tools and gave more an overview of the modern data editing methods.

16. In spite of the fact that many people were on leave around 180 managers and specialists participated at the workshops – a lot compared to the turnouts of other workshops. The workshops lasted around 150 minutes and the participants received handouts at the beginning. This duration was just acceptable but at the end of some workshops it was obvious that some participants (a minority) didn't want to get additional information. Opposed to them other users comment the workshops as very intensive, interesting, and helpful events. Isolated negative developments occurred when some subject matter statisticians tried to use the events to clarify their survey specific problems ("Does the PL-Editor also support my special checks?") or didn't want to be convinced (In one case a non planned very controversial discussion on automatic editing methods with a stubborn manager in the presence of nearly 30 participants had to be aborted).

17. The current plan foresees follow up workshops in the midst of 2005. The aim is to keep in touch with the "clients" and to give a report on the latest results and further developments. The test of modern data editing methods confirms the need for these workshops because the establishment of the new methods as regular procedures will induce higher skills of subject matter statisticians and may presumably lead to organisational changes. It is planned to perform these post-implementation workshops again in the approved top-down-approach.

18. A very important part of the implementation strategy was the introduction of selective and macro editing methods because it is obvious that they will be demanded and used by subject matter units. Selective, macro, and automatic editing methods were presented and explained at a very simple level so that the participants could comprehend the statements. The selective editing method was introduced by a simple example which shows the influence of an error dependent on its degree and the number of records to be used for the computation of the respective statistical result.

19. Another action was the use of a simple method for the test of a selective editing (distances of values from the mean in relation to the estimated plausible statistical result) and the request to the subject matter unit to set priorities by the choice of the characteristics and subject matter oriented weights.

20. The data editing concept, training courses, and the introduction concept possess a lot of innovative aspects that affect the subject matter units. To get pre-information on available capabilities, possible reactions, and existing knowledge a pilot phase took place after the implementation that ended in December 2004. After this phase some of the IT tools, materials, and concepts were optimised so that they are now ready for the use by all statistical offices. In addition to this the current plans make arrangements for user workshops on the IT tools with the aim to obtain best production versions. These arrangements also mark the end of the implementation.

21. It is planned to realise the concept of the two-phase implementation also for the statistical offices of the Länder in 2005. The successful integration of the data editing concept in the procedures of the FSO was considered as a test in this context.

IV. CONCLUSIONS

22. All in all the implementation strategy was successful; especially the assumptions were very realistic. Minor amendments concern the training courses, the contribution of subject matter statisticians as regards the use of the PL-Editor, and the workshops could a little be shorter. In spite of the fact that the developed macro strategy reflects the specific situation of German statistical offices the following conclusions as regards a successful introduction of modern data editing methods can be drawn:

- The basis of an introduction strategy is the tasks to be modernised. Choose activities that allow satisfying best the needs of subject matter units and implement them in helpful IT tools. Try to find out the addressees who are responsible for the tasks and try to convince them, obtain pre-information on their needs and try to test parts of the objects / IT tools to be implemented.
- Advantageous preconditions facilitate changes: the ongoing reductions of staff – how challenging they are for a statistical office – create an incentive for the introduction of modern data editing

methods. Another positive precondition can be created if target groups obtain information on the expected results. This was done by the simulations of new data editing methods.

- Manager of the upper hierarchy should be informed first. Their support is an important precondition of a successful implementation but it facilitates only the start of the implementation. It has to be confirmed by a positive feedback from the specialists in the middle term to ensure that the new methods will be a part of the standard processes. Inform all concerned groups very intensively on the most important aspects. Keep in touch with the target groups and communicate first positive results until new methods and IT tools are regarded as standing procedures.
- Implement simplified methods first so that subject matter statisticians and managers can verify the changes easily. Sophisticated methods should follow to optimise the methods.
- In the case of complex projects which affect heterogeneous aspects of the survey processing the members of the steering committee should be very communicative and be able to cooperate with specialists (interdisciplinary work). Engaged and interested assistants should be responsible for parts of the concept and be able to present the results of their activities (team work in the case of a complex project).
- Though an adequate portion of fortune is no precondition for the successful implementation of a new data editing strategy, it facilitates the procedures. Special aspects cover the choice of the subject matter unit for the test of new data editing methods, the consultation by colleagues and the partners of the steering committee.

Implementation of a Combined Macro and Selective Editing Method

23. The introduction provided the background for the implementation of new data editing methods. As the FSO started with a "white paper" in this new area (no experience, no IT tools) the implementation of a single selective editing method was decided.

The Annual Survey on Costs of the Producing Industry

24. The annual survey on costs of the producing industry asks a sample of 16 000 enterprises on employees and costs (50 characteristics) via paper and Internet questionnaires. The enterprises are stratified in accordance with the NACE on the fourth level and five classes of employees. The sample is renewed every fifth year but big enterprises with numerous employees are so rare that they are not replaced. The projected statistical results (sums) are first stratified in accordance with the NACE classification on the fourth level and second in five classes of employees. At the end of October that means after two months of data editing 16 characteristics of the plausible records are used for the dissemination of preliminary results which are projected and roughly stratified.

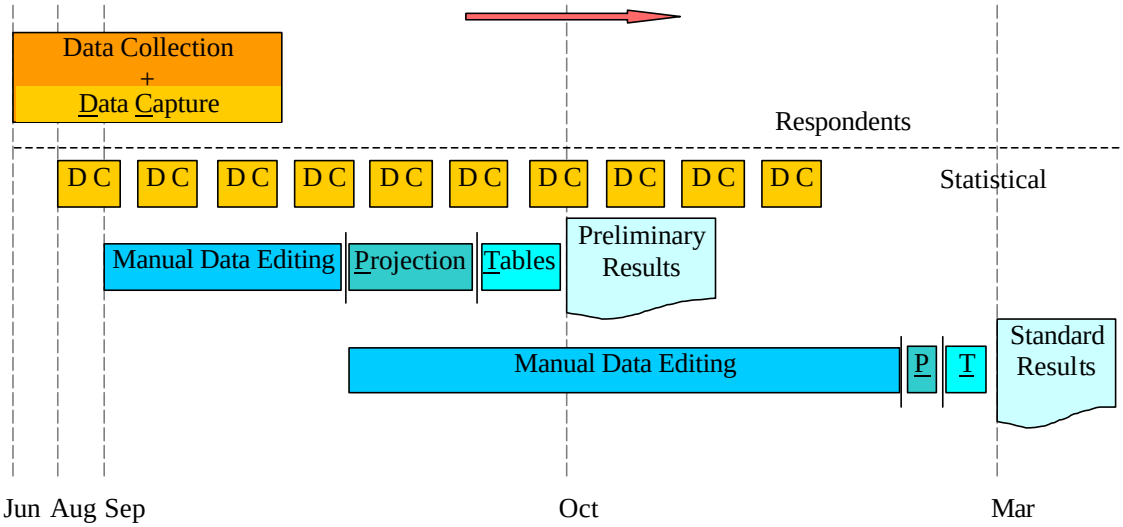
25. The short timeframe for the dissemination of the preliminary results requires efficient procedures. One activity is the use of an Internet questionnaire which was used by 2 000 enterprises in 2004 (1 000 enterprises the year ago). To restrict respondents' burden the questionnaire contains only simple completion checks. The data collection process regularly starts in July and the data editing process in August / September. It terminates in February of the subsequent year. Being aware of the fact that the enterprises have to report over four years the subject matter unit contacts them in the case of serious errors and tries to receive corrected values. The subject matter unit uses an IT application which checks records on the base of specified checks, supports the correction of answers in a dialogue mode and displays the original answers and the ones of the last year. In addition to this the experts also use the original paper questionnaires with possible additional information for the correction of erroneous records.

26. As regards data editing the survey is subdivided into branches because of their specific aspects. Completed paper questionnaires are prepared and sent in packages of 200 up to 500 to the central data capture. The returning raw data are imported in the IT application and the error status per record is determined. As regards the dissemination of the preliminary results under tough time restrictions the clerks use their experience derived from previous surveys and focus big enterprises (but most of them are trained due to their permanent participation) and branches which possessed a lot of erroneous data in the past. Tools for the efficient detection of serious errors were not available. The clerks try to correct as

many erroneous records as possible. Three days before the dissemination of the preliminary statistical results the IT production manager transfer the plausible records (around 10 000) to the mainframe and performs the projection though a simple method is used. After receiving the projection factors the subject matter unit tabulates the data checks roughly the preliminary results and disseminate them to Eurostat. After the dissemination the data editing continued for the standard statistical results. Figure 1 shows the data processing of the survey on costs of the producing industry:

Figure 1: Existing data processing of the annual survey on costs of the producing industry

- No Scale -



The Test of a New Selective Editing Method

27. The selective editing method was the second one that the FSO developed.[4] To achieve an acceptance of this type of method and to facilitate the approach the use of a simple method was decided. Preparing analysis revealed that the majority of the corrections led to the augmentation of the raw data. Most of the corrections were done to the characteristics on costs whereby characteristics on raw materials can be found on the top ranks. As the subject matter statisticians correct erroneous answers dependent on the situation of an enterprise one year ago the following model was chosen:

z	the national category of the NACE for a branch of the producing industry
b	the category "number of employees" for a producing enterprise
m	an index of the characteristics; $m = 1,..., M$
n	an index of the records (enterprises); $n = 1,\ldots,N$
$x_{mn}(t)$	the value for the characteristic X_m of the enterprise n at time t as raw date
$x^*_{mn}(t)$	a model based plausible value for the characteristic X_m of the enterprise n at time t
l_m	a subject matter oriented weight of a characteristic m
w_m	a methodological weight of a characteristic m
v_m	an indicator variable with value 1, if an error score shall be computed for a characteristic m
$h_n(t)$	the preliminary projection factor for the time t for a record n
k	arbitrary index of the modified Minkowski Distance

[4] The author thanks Nicole Lorenz, Technical University of Chemnitz, for her contribution to the development of the method.

28. Then an error score a_n for a record n (enterprise) can be computed with:

$$a_n = h_n(2002) \cdot \left\{ \sum_{m=1}^{M} l_m \cdot w_m \cdot v_m \cdot \left(\frac{\left| x_{mn}(2003) - x^*_{mn}(2002) \right|}{\sum_{n \in N_{zb}} x^*_{mn}(2002)} \right)^k \right\}^{\frac{1}{k}}$$

29. One crucial point of the existing data processing is the early availability of numerous raw data for a data editing. One ad hoc measure was the reduction of the number of questionnaires per data capture package but this was only a provisional solution. Thus it was decided to use a comparison with the plausible values of the previous year as error detection mechanism and to abstain from a feedback on the basis of performed corrections.

30. The selective editing method was tested from September to October 2004 and started with the training of the subject matter statisticians and dissemination of the lists. A first computation was undertaken in the beginning of September (15/9) and a first evaluation of the method in the beginning of October 2004 (8/10). A second evaluation comprised the data which were used for the dissemination of the preliminary results (27/10). The first computation comprised 10 322 records, the second (14 431) with 2 130 corrected records and the third is based on 14 611 records with around 5 135 corrected ones. The first evaluation of the method and the availability of additional 4 000 records led to the dissemination of second lists for the corrections. The evaluation didn't indicate a need for an adaptation of the method. The analysis of the plausible data could not be done when this paper was written but it is planned for February 2005.

31. The overall result of the selective editing method is that 50 percent of the most important corrections reduce the mean plausibility deviation per characteristic of around 80 percent. Opposed to that the last 20 percent of the corrections eliminate only less than 10 percent of the mean plausibility deviation per characteristic. Besides them the most important aspects of the test were:

- Did the method work properly under completely changing conditions (new sample)?
- How did missing records affect the behaviour and power of the method?
- How did the method cope with the existing work flows and internal procedures?
- Did the preliminary IT tools work properly and fulfil the needs of the subject matter unit? Did the lists contain the relevant information for the corrections?
- Did the personnel control the method and new IT tools after a specific training?

Power of the selective editing method:

32. Though the overall results of this first method seem to be satisfactorily further analysis should be done to explore possible aspects of improvements. Thus the ranked error scores of the corrected and still erroneous records were analysed. A rank of 1 indicates a high priority of an erroneous record:

Plausibility of Records	**Max**	**U. Quartile**	**Mean**	**Median**	**L. Quartile**	**Min**
Corrected 27/10	100	73	48	45	22	1
Erroneous 27/10	100	79	55	55	33	1
Erroneous 15/9 + 8/10 + 27/10	100	81	58	58	37	1

33. The table shows that the subject matter unit tended to correct the more important erroneous records but the analysis revealed that at the end of October still important records remained erroneous. This result confirms the first impression of the analysis from September 2004 when the subject matter unit waited for corrections to be done by enterprises. To obtain more information on this development it was found out that 1 836 out of 14 600 records were erroneous over the entire testing phase. The analysis of their ranked error scores (last row of the table above) shows that they tend to represent more the less important ones but not all of them. As the subject matter unit prioritised the most erroneous records the presence of important erroneous records at the end of the test indicates that the corrections made by

enterprises seems to be a bottleneck of the existing data editing strategy that may distort the preliminary results.

Recommendation:

The influence of important missing erroneous records on preliminary results has to be checked. In the case of an important influence of erroneous records on the preliminary results an automatic editing should be tested.

34. A second analysis should provide more detailed information on the error detection and priority setting by the selective editing method. Thus the means of the absolute corrections per characteristic were ranked by the quartiles of the record specific error scores. The following table shows the means of the quartiles in relation to the maximum per row for the respective characteristic (C). The table contains only selected characteristics with specific developments, the column "Cases" informs on similar developments of other characteristics (total of characteristics: 16) not explicitly shown:

Characteristic	Cases	Mean of the absolute corrections ordered by quartiles of the error scores			
		1. Quartile	2. Quartile	3. Quartile	4. Quartile
C40	6	100.0	61.5	14.4	1.5
C52	2	100.0	95.0	57.0	52.4
C63	1	61.8	100.0	6.2	19.6
C64	7	100.0	70.1	18.7	45.5

35. The table shows four different types of plausibility improvements: seven characteristics possessed a decreasing plausibility improvement but with a – partly considerable – improvement in the last quartile of the error scores. Six characteristics possessed a decreasing plausibility improvement, and the plausibility improvements of three characteristics differ completely from the others. In spite of the different developments the tables shows that the method led to non negligible improvements in the last two quartiles. One main reason may be the use of a weak error detection mechanism due to the new sample (80 percent of new enterprises). In addition to this the table seems to confirm the impression derived from the development of the method: the error score is only a "trade-off" of 16 characteristics such that some of them displace others. Consequently the error scores should be adjusted dependent on the plausible improvements of the statistical results combined with information on their plausibility from the previous year.

Recommendation:

The error detection needs further improvements: Comparisons between current raw data and past plausible data should replace the computation of the distance between the mean and raw date. An adjustment of existing error scores should be established which bears in mind achieved plausibility improvements.

36. The test started when the majority of completed questionnaires were available but four weeks before the dissemination of the preliminary results more than 4400 new and unchecked records were available. An analysis was therefore carried out to explore the influence of the new records on the priority setting within the strata. The following table shows the differences of the ranked error scores of the already 1 836 mentioned erroneous records at the beginning of the test (15/9), the first evaluation (8/10), and the second evaluation (27/10) that means after the availability of more than 4400 new unchecked records. The results can only provide some restricted insights because they are heavily determined by corrections performed during the respective periods and the randomly happened sequence of the types of questionnaires.

Difference of the ranked error scores	Max	99. Pctl	95. Pctl	90. Pctl	U. Quartile	L. Quartile	10. Pctl	5. Pctl	1. Pctl	Min
8/10 – 27/10	78	37	14	7	1	-3	-8	-13	-26	-57
15/9 – 27/10	67	34	17	11	4	-4	-10	-15	-33	-99

37. The first row of the table shows the effect of the new unchecked records. 80 percent of the erroneous records changed their priority of around 8 ranks. Two percent changed up to a maximum of 37 ranks. The second row shows the effect of the corrections and new records.

The Test of a Macro Editing Method

38. As regards data editing the sample of the survey was subdivided in branches. To minimise the changes of the existing work flow this subdivision was considered to be an unchangeable determining factor. On the background of the tight time schedule a priority setting among the strata required the implementation of a macro editing method.

Important factors for the priority setting among the strata were:

- The provision of projected preliminary results on a part of the sample. Consequence: the projection factor should be part of the method. (projection factor)
- The need to improve the plausibility of the strata very rapidly. Consequence: the priority setting should also be based on a key figure which indicates rapid plausibility improvement of a stratum. (error indicator)
- The structure of the costs is heavily determined by the size of the enterprises. As the total costs of the producing industry are of primary interest appropriate stable factors which express the contribution of the branches to the total (number of employees, turnover) should contribute to the priority setting. (fraction)
- A last, minor important aspect was to maintain a continuity of the results. Thus actual results had to be compared with ones of the previous year. (comparability factor)

39. The macro editing method had to develop within a few days. Consequently it could only be based of subject matter considerations. As there was no experience on the power of the selective editing method it was decided to use the error indicator and the fraction. Let

i	index of the strata i=1, ..., I
t	period of the survey t=1, ..., T
m	index of the characteristics m=1, ..., M
n	number of sample units n=1, ..., N
G	number of units of the universe G > 0
x^*	plausible value
x	partly plausible / raw value
v	a stratum's fraction of the total, here: turnover and number of employees per stratum
a	error score of the selective editing method per record
py	priority of a stratum

40. Then the priority of a stratum can be computed with:

$$py_i = \frac{G_{i;t}}{n_{i;t}} \cdot \left(\sum_m \left| \frac{\left(\frac{n_{i;t}}{n_{i;t-1}} \cdot \sum_{n_{i;t-1}} x^*_{m;t-1} \right) - \sum_{n_{i;t}} x_{m;t}}{\frac{n_{i;t}}{n_{i;t-1}} \cdot \sum_{n_{t-1}} x^*_{m;t-1}} \right| \right) \cdot \frac{r(a_{n_i;t})}{\tilde{a}_{n_i;t}} \cdot \left(\frac{n_{i;t}}{n_{i;t-1}} \cdot f\left(v_{i;m;t-1}\right) \right)$$

41. As both methods were new it was decided to compute the indicators with additional IT modules. Another reason for this decision was to restrict the effort for the realisation of these methods. As SAS is a standard software of Destatis two standardised SAS-projects were developed which used new SAS-macros for the selective and macro editing method and supported (graphical) comparisons between actual and previous results.[ii] The output of the SAS project which supports the selective and macro editing method is a list with the sorted enterprises.

42. The priority setting by the macro and selective editing method can be shown by the following figure. It contains a list with the erroneous records of a subject matter statistician at the beginning of the first half and a second for the same person at the beginning of the second half. Both lists contain only the first 30 percent of the assumed most erroneous records per stratum (example: 2442/5, 2442/3). The lists of the following figure represent first versions which complicated the work of the clerks who ranked the enterprises on the basis of the identifiers so that they can easily retrieve the original questionnaires. Thus the ranks of the records within a stratum defined by the selective editing method were provided as preliminary solution.

Figure 2: Excerpts of two lists for setting priorities among erroneous records for the same statistician

WZNr	UGrKl	UNr	ErrScore	Med ErrScore 2002	Max ErrScore 2002
2442	5	014028274	587,38	71,44	274,60
2442	5	004021965	403,28	71,44	274,60
2442	5	058834720	379,62	71,44	274,60
2442	5	054212963	97,01	71,44	274,60
2442	5	008219776	90,50	71,44	274,60
2442	3	039081382	633,14	82,86	180,63
2442	3	006006955	303,05	82,86	180,63
2442	3	013048846	252,91	82,86	180,63
2442	3	022521145	227,81	82,86	180,63
2442	3	027039320	151,07	82,86	180,63
2442	3	018159153	114,01	82,86	180,63
2442	3	054286270	101,21	82,86	180,63

WZNr	UGrKl	UNr	ErrScore	Med ErrScore 2002
2442	5	014028274	499,22	71,44
2442	5	058834720	310,59	71,44
2442	5	006914206	98,70	71,44
2442	5	014003252	83,99	71,44
2524	5	027196801	84,91	75,71
2524	5	027082186	70,47	75,71
2524	5	018116700	69,11	75,71
2521	5	056127992	67,57	72,22
2521	5	008350453	56,74	72,22
2521	5	027305772	54,30	72,22
2521	5	018680983	52,16	72,22
2524	3	039094906	72,08	58,46

43. The screenshot on the left shows a priority setting driven by high error scores at the beginning of the test (15/9). Opposed to that priority setting on the right (beginning of the second half, 8/10) was more driven by the fraction of a stratum at the total result. This development was discovered on other lists too. So there was a need to get further information on the influence of the four factors on the priority setting. First the medians and means of the factors were computed for the quartiles of the priority setting. The computations were performed for the first and second half of the test:

Factors of the Makro-Edit.-Meth. / Priority setting	**1. Quartile**		**2. Quartile**		**3. Quartile**		**4. Quartile**	
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Data of 15/9								
Comparability	101.97	74.49	109.53	72.30	128.48	77.11	92.34	53.41
Fraction	0.27	0.07	0.06	0.03	0.02	0.01	0.01	0.01
Error Indicator	3.98	3.12	1.77	1.42	1.15	0.97	0.85	0.69
Projection	3.84	2.33	3.14	2.38	2.56	2.00	2.02	1.50
Priority setting	134.48	45.93	9.43	8.69	2.31	2.12	0.43	0.38
Data of 8/10								
Comparability	108.40	77.40	104.40	66.40	135.50	78.20	99.20	56.50
Fraction	0.27	0.07	0.05	0.03	0.02	0.01	0.01	0.01
Error Indicator	4.90	4.00	2.20	1.90	1.40	1.20	0.90	0.80
Projection	2.50	1.80	2.20	1.70	1.60	1.30	1.40	1.00
Priority setting	138.40	42.30	8.50	7.90	2.10	1.90	0.40	0.30

44. The table shows a significant difference of the priority setting between the first and other quartiles. With the exception of the comparability factor the influence of all other factors decreased. The factor "fraction" didn't change because it is computed on the plausible data of the previous year. The statistics of the projection factors of the second half are smaller than the ones of the first half. This development can be explained by the additional records. Opposed to that it seems that the influence of the error indicators grew. To obtain more detailed information on the influence of the factors the partial correlation coefficients were computed:

Part. Corr.-Coefficients, N = 1 137 (Priority Setting of the macro-editing-method with factors) P > |r| under H_0: Partial Rho=0[5]		
Factor	**Spearman's Rank Correlation Coefficient**	**Pearson's Correlation Coefficient**
Data from 15/9 (beginning of the test)		
Fraction	0.927**	0.892**
Error Indicator	0.855**	0.538**
Projection	0.689**	-0.006
Comparability	0.840**	0.093*
Data from 8/10 (second half of the test)		
Fraction	0.930**	0.932**
Error Indicator	0.859**	0.479**
Projection	0.672**	0.022
Comparability	0.853**	0.096*

**: $\alpha_{emp} < 0.0001$; *: $\alpha_{emp} < 0.01$

45. Spearman's Rank Correlation Coefficient was used to see whether there are similar basic rankings. Compared to that Pearson's Correlation Coefficient shall provide more information on the strength of the influence derived from the absence or presence of linear relationships. Missing or less linear relationships indicate no / less influence of a factor. The table shows that all Spearman's Rank Correlation Coefficient show a high partial correlation between the priority of a stratum and the factors with the factor "fraction" on top and the projection factor at the end. In the second half the projection factor lost a little bit of its influence. The Pearson's Correlation Coefficients show that the factor "fraction" determined the priority setting, followed by the error indicator. Opposed to the first half the factor "fraction" gained more influence and the error indicator less.

Recommendation:

The macro editing method worked rather well. The influence of the error indicator should be strengthened as regards the provision of accurate preliminary results.

Control of the projects:

46. The subject matter unit received two standardised SAS projects for computing the error scores and priorities of the strata and comparing actual results with previous ones. The projects were well described and the personnel were trained over four hours. Last but not least the comparison of the statistical results revealed that the personnel needed more training especially on exploring data analysis and the use of the projects.

Conclusions and Further Developments

47. The test of the selective and macro editing method succeeded: the most important strata were prioritised and the subject matter unit detected the most erroneous records. Another important result is that the head of the open minded subject matter unit regards the methods as standard procedures. The support of two other subject matter units shows that first the subject matter statisticians were surprised of the "new world" besides their traditional checks but the longer the co-operation lasted the better they could be convinced.

48. The test clearly shows that the improvement of the data processing is not only an affair of the data editing methodologists but requires a broad approach. The following synopsis lists the disadvantages of the existing data processing and faces them with possible improvements. The right column functions as an ideal case so that dependencies between improvements become obvious.

No	Current Data Processing	Improved Data Processing
	Data Collection	
1	Paper Questionnaires:	Paper Questionnaires: Redesign on the basis of available characteristics of the

[5] Rho = $\frac{(n-k-2)^{1/2} r}{(1-r^2)^{1/2}}$; Rho ~ t(n-k-2) with n: number of records, here: strata; k: number of partial excluded characteristics, here: factors of the macro editing method; Rho = 0 if partial r = 0.

No	Current Data Processing	Improved Data Processing
		enterprises and error analysis
2	Internet Questionnaires: Relatively slow number of users	Internet Questionnaires: Increasing the usage by the following possible projects: a contest for users, a link between participation and donation, a reuse of existing information
3	Checking: No Checks	Checking: Integration of ordinary checks
4	Transforming paper questionnaires: Data Capture	Transforming paper questionnaires: Scanning combined with completion checks (on record level)
	Data Editing	
5	Serious errors: Selective and macro editing (manually)	Serious errors: Optimised / output oriented selective and macro editing (manually) -> integration of a simple projection and analysis in the editing process to determine the end of the manual editing
6	Marginal errors: manual editing	Marginal errors: automatic editing
7	Coordination between different sections: Simple projection performed by the IT specialist	Coordination between different sections: Reducing the effort for coordination by integrating the simple projection for the production of statistical results in the workflow of the subject matter unit -> performance by subject matter statisticians[iii]

49. The table shows various possible measures of improvements which cover IT specific, methodological and organisational aspects. As it is an annual survey with around 16 000 enterprises a cost-benefit-analysis should determine the realisation of expensive improving activities, e.g. the establishment of an automatic scanning system. Alternatively it should be checked if this system is multiply usable. A new precondition for the establishment of this system is a standardisation of processes and data collection instruments.

50. Other improving measures are more easily to realise but would require an enhanced knowledge of the subject matter statisticians and thus induce additional training. A typical example is the reintegration of the projection in the workflow of the subject matter unit. This proposal may be only applicable for a simple projection method. The alternative will be reasonable if it disburdens the IT specialist who supports more subject units and whose support is a bottleneck as regards further improvements.

51. The redesign of the data processing – as proposed in section 31 – would led to a forward displacement of data editing activities to respondents and a combined data editing activities by subject matter units. In addition to this it is assumed that preliminary results will be disseminated solely on the basis of manually corrected important records and plausible ones – an aspect which has to be verified. An alternative will be to replace the preliminary results by standard ones that were produced on the basis of a selective and automatic editing.

52. Another conclusion is that too much time was needed for the correction of the major errors on the basis of updated information provided by the enterprises. An automatic correction of these errors could improve the timeliness of the results. The implementation of this method seems to be conflicting: the subject matter statisticians mentioned that they have to cooperate with enterprises for four years. If they didn't give a feed back on erroneous information in the first year the plausibility of the last three years would decrease significantly.

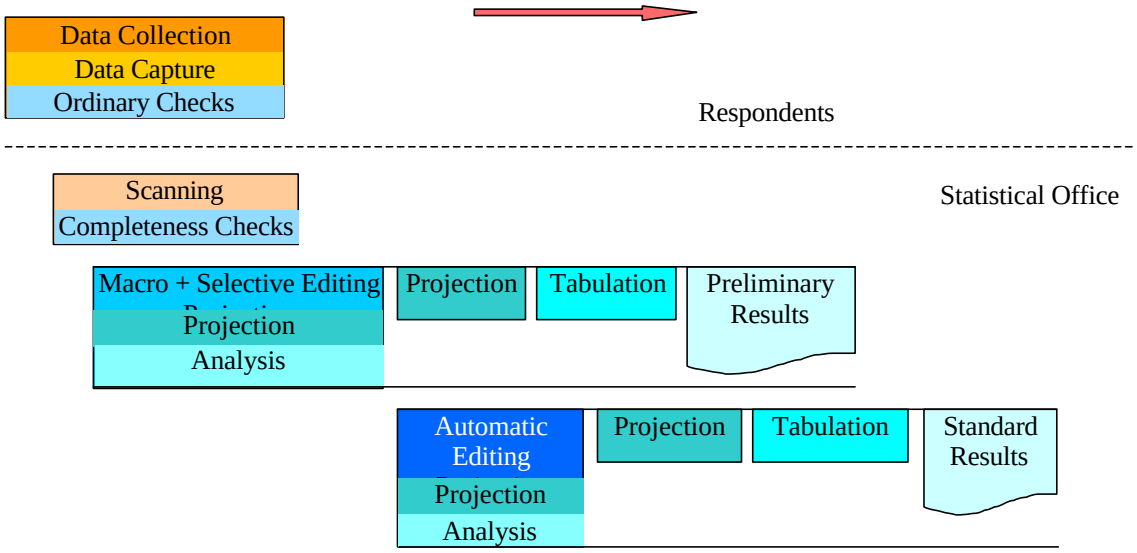
To achieve an efficient employment of the available personnel it will be necessary to improve the analysis tools of the subject matter unit and the respective selective editing methods.

53. As it is planned to replace manual editing by an automatic one it should be checked whether it makes sense to employ less subject matter statisticians within one unit or to establish a central editing

unit in the future. An important counter argument is the needed subject matter knowledge for manual corrections of economic branches. This consideration is a clear sign that modern data editing methods may influence the organisational structure of a statistical office as well as an existing process organisation. as shown by the following figure:

Figure 3: Proposal for a redesigned processing of the annual survey on costs of the producing industry

- No Scale -



54. The realisation of the remaining measures not explicitly mentioned should be decided dependant on the weaknesses of the current data processing, their effort needed for realisation, available similar methods, and software. Modifications of the survey contents may represent – dependent on the degree of changes – a key measure that induces subsequent (tremendous) changes. Another important aspect is that the legal basis of the statistic may reduce possible changes in this area.

55. The aspects discussed in the last chapters clearly show heterogeneous aspects induced by the implementation of modern data editing methods. Thus data editing methodologists have to collaborate and to be able to communicate with IT managers, questionnaire methodologists, subject matter statisticians, organisers and lecturers. Another consequence is that organisational developments are determined by available know how, qualified personnel, financial means, existing strengths and weaknesses, and equipment. The may be an optimal practice but the way to reach it clearly depends on specific organisational circumstances. These considerations induced by the test of one selective editing and macro editing method are confirmed by similar results occurred during the development of two other selective editing methods.

56. Thus we conclude that the comprehensive process and project management approach of the new German data editing concept was the best decision we had made when we started with its development.

---

[i] For further information see Elmar Wein: "Concepts, Materials, and IT Modules for Data Editing of German Statistics", SDE work session, Ottawa 2005; "Improvement of Data Editing Processes, Q2004, Proceedings of the Conference"

[ii] For further details on the SAS-macros see Elmar Wein: "Concepts, Materials, and IT Modules for Data Editing of German Statistics", SDE work session 2005, Ottawa

[iii] Manfred Schulte-Zurhausen (1995). "Organisation" (in German only), München; Günter Schmidt. "Methods of the Process Management" (in German only), WiSt, 5