WP. 16
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21–23 April 2008)

Topic (iii): Improvement of quality through data editing

THE USABILITY OF CORRECTIONS FOR IMPROVING AND PRICING DATA QUALITY

Invited Paper

Submitted by the Federal Statistical Office, Germany [1]

I. INTRODUCTION

1. The annual German statistics on wholesale trade are performed inline with European legislative acts. They deliver structural information on personnel, turnover, and costs of wholesale trade enterprises. With the annual survey on the wholesale trade of Germany up to 12,000 enterprises are questioned. Information is collected by a mix of automatic data transmission (eSTATISTIK.core[2]), internet questionnaires, and paper questionnaires.

2. However, the interest groups of the enterprises criticise actuality deficits and inconsistencies compared with the results of other surveys. For the questioned enterprises themselves the reduction of the load originating from the questionnaire is an important issue. An optimised questionnaire should facilitate the participation and raise at the same time the quality of the data.

3. Thus the Federal Statistical Office of Germany (FSO Germany) decided to evaluate the questionnaire by a pretest in 2007. The pretest design combined the analysis of the corrections with focus interviews of subject matter statisticians and representatives of enterprises. The results obtained by the analysis of the corrections contributed to the recommendations as regards the improvement of the questionnaire which were obtained from other pretest methods. The contribution will describe suitable indicators as regards detecting measurement errors derived from the analysis of the corrections.

4. The analysis of corrections causes an additional work load. Hence it is useful to use this information for improving data editing processes and the pricing of survey characteristics. First considerations of these topics will be discussed at the end of this contribution.

II. ANALYSIS OF CORRECTIONS AS A MODULE OF A QUESTIONNAIRE PRETEST

A. Brief overview of the pretest

5. Methodologists of the FSO developed in accordance with the recommendations for the systematic conduction of pretests a concept with different test methods. They bear in mind the specific conditions of the annual structural survey of the wholesale trade as well as the intention to consider as many conditions

---

[1] Prepared by Elmar Wein; elmar.wein@destatis.de; the statements of this contribution reflect the author's opinion.

[2] Michael Schäfer: "eSTATISTIK.core: Collecting Raw Data from ERP Systems", UNECE 2006, http://www.unece.org/stats/documents/2006.09.sde.htm

of the respondents as possible. The paper questionnaire stood on the test bench in the wholesale trade of 2005 which was used - from smaller corrections seen - also in 2006.

6. The main focus lay on qualitative test methods, but also quantitatively straightened procedures were taken into consideration. [3] [4] All persons involved in the survey process (e.g. staff from the subject matter unit as well as users) were also included. In addition, above all, members of the enterprises were interviewed to find out the inter activities with the data collection instrument.
The pretest started with an analysis of the corrections. The information from the analysis was used to prepare semi standardised interviews of reporting enterprises that means employees / owners of enterprises were asked on specific characteristics to complete the information obtained by the analysis of the corrections.

B. Methodological considerations as regards the analysis of corrections

7. Hints as regards the optimisation need of a questionnaire can be obtained by comparing the information given by respondents with the true values. In practice true values are not known, thus plausible values as substitutes had to be used. A possible disadvantage of this modification is that plausible values provide only a consistent *image* of the reality on the basis of (corrected) information. That means a possible discrepancy between the reality and the consistent information may be caused by the person who completes a questionnaire (misunderstanding, incomplete estimations).

8. Another important source of discrepancies may be caused by the subject matter statistician who is responsible for correcting inconsistent data. Discrepancies may especially occur in the case of “signals”, one-sided corrections or by the choice of the inappropriate characteristics to be corrected. Thus false interpretations can be originated from this modified approach concerning the weak spots of the questionnaire. To reduce the amount of misinterpretations the focus will be set on the corrections which were often carried out that means high frequencies of corrections will be interpreted as an indicator for a measurement error. This procedure will be effective if more than one subject matter statistician edit data of the same economic sectors in the case of enterprise surveys. The limitation of this approach is that it does not reveal the real problems of respondents while completing a questionnaire. They can only be found out with classical pretest methods.

9. Another important precondition for a correct interpretation of corrections is the knowledge of the underlying data editing specifications. In surveys with numerous characteristics some of them may serve as “anchors” that means other characteristics are modified in such a way that they are consistent with these survey characteristics.

10. To ensure that all measurement errors will be detected it is assumed that all plausibility checks and statistical error detection methods are specified and employed. In addition to this it is necessary to assume that types of errors will be corrected in the same way.

11. An important precondition for the analysis of corrections is their documentation. An easy way to document them is the comparison of raw data with plausible data. It is a common practice of official statistics of Germany (= FSO and statistical offices of the German states) to enter and edit data in one step. So in many cases there are only plausible data available. Another problem may be that only the majority of the raw data is available whereas the data of enterprises which report very lately are entered and edited in one step. It is assumed that these missing data posses the same statistical properties than the majority of the available raw and plausible data.

12. The primary aim of the analysis is to obtain hints as regards weak parts of the questionnaire. This consideration may first lead to the decision to perform the analysis on the basis of absolute data but this is not an appropriate approach because it becomes obviously that this measure would hinder comparisons of characteristics. As the analysis of corrections induces a heavy workload the benefit of it should be

[3] Australian Bureau of Statistics (ed.) (2001): Pretesting in survey development. An Australian Bureau of Statistics perspective. Research Paper. Canberra, Australia: Australian Bureau of Statistics.
[4] Eurostat (2007): Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys, p. 66ff. (http://edimbus.istat.it/)

maximised in such a way that the analysis should also support comparisons over time, e.g. a monitoring of the effects induced by modifications of a questionnaire, comparisons of similar surveys like the structural business surveys of the EU, and it should improve the preconditions for the optimisation of data editing, e.g. by the provision of information for the development of selective or macro editing methods. To fulfil these enhanced demands the analysis of relative corrections with the plausible values as the basis is introduced. A relative correction c_m of a characteristic m is defined as:

$$c_m = \begin{cases} \frac{v_m^{pl} - v_m^r}{v_m^{pl}} \cdot 100, & v_m^{pl} \neq 0 \\ 100, & v_m^{pl} = 0 \wedge v_m^r < 0 \\ -100, & v_m^{pl} = 0 \wedge v_m^r > 0 \\ ., & v_m^r = v_m^{pl} \end{cases}$$

It shows the amount of a change between a raw value v^r and a plausible one v^{pl} compared to a plausible value v^{pl}. If there is no correction a zero value deters computational analysis. Thus c_m is set to a missing value in these cases. To provide information about the direction of a correction c_m may be positive or negative and is set to ± 100 if the positive value v^{pl} is zero. This information is very helpful with regard to the improvement of a questionnaire because it indicates an under or over coverage during the data collection process. An absolute relative correction c_m^a is similarly defined. The only distinction is that the absolute relative correction is always positive.

13. The relative corrections can be summed up to subtotals and totals. If the sum of the relative corrections for a characteristic is positive than this fact may be used as a hint for an under coverage and otherwise for an over coverage. Another – similar – indicator for under or over coverage may be the median or the number of positive / negative relative corrections.
The definitions above can be similarly applied to aggregated characteristics which are computed on the basis of collected characteristics. The leading sign indicates, whether positive or negative relative corrections dominate an aggregated characteristic. A comparison of these relative corrections with the relative corrections of collected characteristics is only meaningfully for averages.

14. Finally it should be stressed that the aim at this step is the detection of errors during the data collection process that means: no weighting of the corrections because all respondents should be treated equally – regardless their influence on statistical results.

15. The following examples illustrate the definitions:

A		B		C		D	
Raw Data		Plausible Data		Absolute Corrections		Corr. [% of the plausible value]	
Enterprise	Turnover	Enterprise	Turnover	Enterprise	Turnover	Enterprise	Turnover
1	100 000	1	90 000	1	-10 000	1	- 11,1
2	75 000	2	80 000	2	5 000	2	6,3
3	0	3	35 000	3	35 000	3	100,0
4	30 000	4	0	4	-30 000	4	- 100,0
5	-30 000	5	0	5	30 000	5	100,0
6	50 000	6	50 000	6	0	6	.

Section C shows the absolute corrections and section D the relative corrections as mentioned in the previous chapter. The relative corrections made for enterprise 3 to 6 represent the special cases as defined in the previous paragraph.

16. Paragraph 13 touched already the question of the significance of a correction that means “When do corrections indicate a measurement problem?”. To give a more precise answer on this question it is proposed to use the median of the relative corrections and judge it together with the number of the respective corrections. If the median is far away from zero and if there is a considerable number of corrections then there is an evidence for a measurement error. If all characteristics possess the same frequencies absolute numbers are sufficient if this precondition is not fulfilled because of filters in the questionnaire relative frequencies (number of corrections / number of observations of a characteristic) are necessary. To secure the conclusions of this analysis identifying extreme corrections (minimum, maximum) is essential when means are used instead of the recommended median.

C. The analysis of corrections in the context of the pretest

17. The approach described above was implemented for the analysis of the corrections for the structural survey on the wholesale trade 2004. The data editing process of this survey is characterised by workflows of automatic checking, manual selection of the variables to be corrected and manual correction. The dataset with up to 240 survey characteristics is edited by 94 checks, around 52 of them are hard checks and 42 of them are signals (soft checks). 12 persons perform the data editing process, each of them is responsible for around 1 000 enterprises which are dedicated to the section 51 of the NACE Rev. 1.1 in such a way that every person is responsible for enterprises from all NACE branches. Important corrections or questionnaires with lots of errors are corrected by re-contacting the respective enterprises. Some corrections are influenced by general conventions.

18. The analysis is based on corrections which were made by 9,300 enterprises. With the help of table 1 it becomes clear that the characteristics from 1 to 3, 6, 7, 8 and 13 had the biggest changes. All other characteristics which were not mentioned were looked - regardless of her partly very big sums - as exceptions because the case figures or the correction sums are very low. They were not taken into further consideration of the pretest.

Table 1: Absolute corrections in percent of plausible values made by the survey on the structure of the German wholesale trade 2004

No.	Survey characteristics	Absolute corrections		
		Number	Mean [%]	Sum [%]
1.	Expenditures for services	1,322	103	2,002,734
2.	Paid taxes	557	594	5,246,859
3.	Trading goods, beginning of the year	508	240	954,626
4.	Turnover by wholesale trade	494	69	324,768
5.	Number of employees	425	89	450,190
6.	Expenditures for trading goods	331	151	460,994
7.	Trading goods, end of the year	325	345	835,935
8.	Expenditures for commodities	321	6,501	12,885,895
9.	Part time employees	310	99	349,839
10.	Turnover [%] by wholesale trade	244	90	157,886
11.	Owner	231	105	364,044
12.	Personnel, total	214	74	94,342
13.	Commodities, end of the year	205	844	1,325,452
14.	Commodities, beginning of the year	203	192	305,516
15.	Turnover [%] by commission trade	195	104	216,399
16.	Employees, male	181	106	105,665

No.	Survey characteristics	Absolute corrections		
		Number	Mean [%]	Sum [%]
17.	Turnover [%] by retail trade	179	126	266,594
18.	Expenditures for social systems	178	125	350,690
19.	…			

19. To obtain information on the direction of the errors table 2 was computed which contains the relative corrections. The median of the characteristics 1 to 7 are positive and thus indicate an under coverage. Opposed to that the characteristics 8 and 9 indicate an over coverage. Medians with the value of 100 indicate an imputation. To interpret the results of the table correctly additional information as regards the correction of these variables is absolutely necessary. Table 2 contains also the extreme values of the corrections which may be used to decide on measurement errors.

Table 2: Relative corrections in percent of plausible values made by the survey on the structure of the German wholesale trade 2004

No.	Survey characteristics	Corrections			
		Number	Minimum	Maximum	Median
1.	Expenditures for services	1,322	-2,391	100	100
2.	Paid taxes	557	-202,649	100	100
3.	Trading goods, beginning of the year	508	-9,900	100	42
4.	Turnover [%] by wholesale trade	494	-1,000	100	34
5.	Number of employees	425	-15,200	100	29
6.	Expenditures for trading goods	331	-38,051	100	75
7.	Trading goods, end of the year	325	-9,900	100	100
8.	Expenditures for commodities	321	-12,307,872	100	-100
9.	Part-time employees	310	-1,500	100	-100
10.	...				

D. Contribution of the corrections to the pretest

20. The comparison of the raw data with the plausible data gave first hints about weaknesses of the questionnaire. The results of this module were used for preparing the interviews with the enterprises. 14 half-standardised interviews were carried out ("company site visits"). This test enabled to find out reasons for implausible information.

21. As a first result can be held on that beside managers and owners different persons (companion, accountant or tax adviser) from the most different enterprise areas fill out the questionnaire what complicates the optimisation of the questionnaire.
In addition, it became clear as with explanations becomes gone forward: The majority of the test persons does not read them or read them only sporadically. Explanations are only red when no own concept understanding is available. In all other cases the concept understanding finds use which is known from media, everyday life or the accountancy. This behaviour takes place with questions, e.g., part-time employees, gross wages, social expenditures. Are to be thought over and to be tested in the future, nevertheless, the order, the length and the layout of explanations.

22. The analysis of the corrections revealed problems with the characteristics “expenditures for services”, “paid taxes”, “trading goods”, and “commodities”. The interviews confirmed the information as regards the characteristics of trading goods. Some respondents mentioned that they don’t know how to price the inventory on trading goods. Another main reason as regards the characteristics of “trading goods” and “commodities” is the fact that tax accountants fill out in many cases the questionnaires and they don’t know how to distinct both characteristics.

Problems with the characteristic “paid taxes” are caused by the fact that some respondents use their own comprehension and thus include the value added tax. Other respondents include in this characteristic the costs for administrative services.

III. THE ANALYSIS OF CORRECTIONS AND OTHER SURVEY PROCESSES

A. Optimisation of the data editing process

23. The main purpose of the analysis as part of the pretest was to deliver hints on measurement errors. Thus corrections were summed up – regardless their influence on statistics. As the modernisation of the data editing process is another topic for the statistics of wholesale trade the analysis switched from the input oriented approach to an output-oriented one. The aim of the output oriented approach is to identify the most important edits.

24. Moving to an output-oriented approach a weighting of the corrections by the projection factors of the enterprises is necessary. To obtain additional information it useful to bear in mind the frequency of a correction as introduced in paragraph 16.
Influential corrections of characteristics can not be obtained alone by dividing relative corrections by their frequencies. The results will be misleading if the contribution of a characteristic to a published statistics is negligible. So there is a need to weight the amount of the relative corrections with the relative contribution of a characteristic to a statistic. If a collected characteristic is also a published statistic this step may be omitted.

25. Table 3 contains the statistics to be published (3rd column) which are in general sums of weighted plausible data with the turnover as the statistic with the overall interest. Columns 3 and 4 contain the changes of the raw data which are the sum of the corrections and imputations. Variable (10) possesses the biggest portion of imputations among the variables listed beneath.
To establish a hierarchy among the variables it seems to be reasonable to consider earnings and expenditures separately. A comparison of the analysis shows that the expenditures for commodities which played an important role referring to the input oriented analysis switches now to a variable of lower interest because of its minor amount – despite of the high relation between corrections and the plausible sum.

Table 3: Relative absolute corrections and their relations to published statistics by the survey on the structure of the German wholesale trade 2004

No.	Survey characteristic	Sum of weighted plausible data	Change of raw data by ...			Number of corrections
			Total		Absolute corrections	
		1000 Euro	1000 Euro	%	%	
1.	Expenditures for trading goods	530,788,318	21,382,004	4.03	3.17	331
2.	Turnover	658,319,003	15,260,742	2.32	1.29	494
3.	Expenditures for commodities	9,629,530	7,044,410	73.15	72.45	321
4.	Expenditures for services	42,075,988	2,582,344	6.14	4.82	1,322
5.	Trading goods, beginning of the year	40,772,543	2,470,549	6.06	5.30	508
6.	Trading goods, end of the year	43,074,296	2,010,962	4.67	3.95	325
7.	Gross pay	36,186,155	1,341,588	3.71	1.76	129
8.	Commodities beginning of a year	1,387,358	1,103,350	79.53	79.02	203
9.	Commodities end of the year	1,829,809	938,898	51.31	50.93	205

No.	Survey characteristic	Sum of weighted plausible data	Change of raw data by ...			Number of corrections
			Total		Absolute corrections	
		1000 Euro	1000 Euro	%	%	
10.	Paid taxes	3,568,022	543,230	15.22	13.34	557
11.	Expenditures for social systems	7,647,493	247,428	3.24	1.67	178
12.	Hire/Rental	6,440,632	155,387	2.41	1.46	52
13.	...					

26. Compared with the analysis in the context of the pretest the output oriented analysis set other focal points than the analysis in the context of the pretest because the corrected sums were set in relation to the published sums. Where as the analysis related to the pretest prioritises the characteristics “expenditures for services”, “paid taxes”, and “turnover by wholesale trade” the output oriented analysis favours the characteristics “expenditures for trading goods”, “turnover“, „expenditures for services”, and characteristics of trading goods.

27. On the basis of available corrections one can also analyse the respective distribution for each characteristic. The existence of big corrections indicates that a selective editing method may help to set priorities among data editing activities. This analysis was also done and showed clearly the existence of extreme corrections. As a consequence it was decided to develop a selective editing method for the structural business survey on wholesale trade.

B. Corrections and costs of editing characteristics

28. The relation between data quality and costs is nowadays generally accepted. Data quality is achieved by the input of the personnel, the use of IT-equipment combined with software and statistical methods, and the use of infrastructure, e.g. an office building or other office structure. In general official statistics can are supplied for free but many offices perform cost accounting systems to monitor the most resource consuming processes. Other situations which may occur for statistical offices are calculations for funded projects. In these situations statistical offices need reliable information on the occurrence of costs in relation with demanded statistics. Another motivation for this paragraph is to provide an approach which enables statistical office to establish a different price management for enhanced statistical analysis on demand.

29. Data editing is a resource consuming process that influences the accuracy of statistics. In general it’s costs are known that means the costs of needed personnel as well as the use of computers. On the basis of manual and automatic corrections the costs can be assigned to different characteristics of a survey. The assignment of costs should reflect the improvement of the accuracy per characteristic and thus should be based on the relative changes – like it was done in column 2 of table 4. In detail the value of the relative change per characteristic was put in relation to the sum of relative changes over all characteristics (here: 277.11%).
Table 4 contains another basis for assigning costs to characteristics (column 3): The sum of the number of the corrections 3,123. It may be used as an alternative basis for assigning costs to survey characteristics. Compared with the favoured relative changes the main disadvantage of this reference is that it does not bear in mind information on the improvement of characteristics.

Table 4: Two ways to allocate costs to published statistics by the survey on the structure of the German wholesale trade 2004

Survey characteristic	Distributing costs on the basis of ...	
	Relative changes	Number of corrections
Expenditures for trading goods	1,14	10,60
Turnover	0,47	15,82
Expenditures for commodities	26,14	10,28
Expenditures for services	1,74	0,04
Trading goods, beginning of the year	1,91	16,26
Trading goods, end of the year	1,43	10,41
Gross pay	0,64	0,00
Commodities beginning of a year	28,52	6,50
Commodities end of the year	18,38	6,56
Paid taxes	4,81	17,83
Expenditures for social systems	0,60	5,70
Hire/Rental	0,53	0,00
Other turnover	0,25	0,00
Sum of investments	1,24	0,00
Investments for equipment	1,24	0,00
Sale of facilities	4,26	0,00
Investments for estates	4,95	0,00
Subsidies	1,62	0,00
Investments for new buildings	0,12	0,00
Investments for the maintenance of buildings	0,01	0,00
Value of purchased equipment	0,01	0,00

Table 4 shows differences between the two references which correlate to an extent of 0.24.

IV. CONCLUSIONS

30. Corrections can be documented easily on the basis of raw and plausible data. They are useful for improving a questionnaire, a data editing process and pricing data quality if miscellaneous preconditions are fulfilled:

- Corrections are based on a data editing strategy which tries to discover all errors in data.
- Corrections are performed correctly by subject matter statisticians or software.
- Modifications of a given data editing strategy are well documented so that their influence on corrections can be noticed.

To interpret corrections correctly additional information of a given data editing strategy (specified checks and used methods) is necessary. For example it should be found out whether there are characteristics which are corrected generally to obtain consistent data records.

31. It is recommended to analyse relative corrections instead of absolute ones because they enable comparisons of characteristics and over time in the sense of a monitoring. Corrections alone do not possess enough information as regards their significance. Hence they should be combined with their (relative) frequencies.

32. Corrections may help to detect measurement errors. Suitable information for this purpose is the frequency of corrections, the extent of changes and their leading signs. Positive corrections indicate an under coverage and negative ones the opposite. To discover measurement errors with a high certainty numerous corrections of a considerable extent may be used as hints.

33. Due to the aim to discover as many measurement errors as possible the significance of corrections on statistical results - which is an important criterion to identify over editing - is of secondary importance. Opposed to that using corrections for improving a given data editing strategy it is recommended to weight them and compare the contribution of the respective characteristics to demanded statistics as a second weighting.

34. Relative corrections represent improvements of accuracy and thus may be used as a basis of a cost allocation to survey characteristics. Main advantages of this approach may result in a more different pricing of statistical results and better calculations for funding projects.