STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 33
English only

CONFERENCE OF EUROPEAN STATISTICIANS

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

Topic (iv): Impact of new technologies on statistical data editing

SELECTIVE EDITING BY MEANS OF PLAUSIBILITY INDICATORS

Contributed paper

Submitted by Statistics Netherlands[1]

Abstract: At Statistics Netherlands the statistical process for annual structural business statistics has been completely redesigned. Within the scope of IMPECT (IMPlementaion Economic Transformation process) a method for selective editing has been developed. Crucial businesses are always edited manually, because of their large impact on publication totals. However, for non-crucial records plausibility indicators decide whether a record is automatically edited or whether it is edited manually. In this paper the emphasis is on the construction of plausibility indicators.

I. INTRODUCTION

1. At Statistics Netherlands a extensive reorganization has been carried out. For instance, instead of a department for each survey the Division of Business Statistics now has a department for

- **business registers**
- **observation:** adjusting, sending, receiving and editing of questionnaires
- **analysis:** imputing unit non response, weighting, macro editing, publication
- **research and support**.

2. This new organization requires a uniform statistical process for all annual structural business statistics. Furthermore, the same work has to be done with fewer people of a higher educational level. The editing process is often a demanding and costly one. It was therefore decided to switch over to selective editing, cf. Granquist (1995), Granquist and Kovar (1997), Hidiroglou and Berthelot (1986). De Jong (2002) describes the new editing process for annual structural business statistics at Statistics Netherlands.

3. In this paper one step in this new process is highlighted, namely the partition of records for automatic editing and records for manual editing. The emphasis is on selective editing of annual structural business statistics. Business statistics differ from social statistics, because they mainly contain numerical data instead of categorical data. Furthermore, companies receive a questionnaire to fill in, while persons are often visited by an interviewer. Questionnaires that are distributed by mail often contain more errors than questionnaires that are dealt with by an interviewer.

II. SELECTIVE AND AUTOMATIC EDITING

4. The goal of selective editing is to select those records for manual editing that have a large influence on the publication total and/or contain large errors. In our view records that do not satisfy either of these conditions can be edited automatically. When records are aggregated to publication totals small non-systematic errors will largely be cancelled out anyway.

[1] Prepared by Jeffrey Hoogland.

5. Automatic editing is done with the software package SLICE (Statistical Localisation, Imputation, and Correction of Errors) which contains a CherryPi module, cf. de Waal (2000) and de Waal and Wings (1999). CherryPi is a program for automatic editing of numerical data that requires a description of the data, edit rules, confidence weights that give an indication of the reliability of each variable, and imputation rules. When at least one edit rule is violated a record is considered to be incorrect and values of variables are changed by means of imputation.

6. CherryPi uses the generalised Fellegi-Holt paradigm for localisation of errors. This paradigm implies that values of variables within a record should be adjusted such that all edit rules are satisfied and the sum of the confidence weights for adjusted variables is minimal.

III. PLAUSIBILITY INDICATORS

A. Selecting records for manual editing

7. For selective editing, we need a reliable estimate of the expected clean value of a variable in a record. Therefore, records are grouped into homogenous subgroups. Such a group is called an editing cell, which is a intersection of a publication cell and a company size class, see Appendix A. For each editing cell the plausibility of records is assessed.

8. A plausibility indicator has to serve several purposes. The main purpose is to make a good selection of records that have to be edited manually. Records that contain errors that have a significant effect on the population total should be selected for manual editing. Let's assume that records that are not selected are not edited and that we have both raw data and clean data. The selection should be such that the pseudo bias of the (weighted) publication cell total

$$\Delta_j(y,\hat{y},S) = \left| \frac{1}{Y_j} \sum_{i \notin S} w_i (y_{ij} - \hat{y}_{ij}) \right|$$

is minimal, where

$y_{ij}, \hat{y}_{ij}$: clean, respectively raw value of variable j in record i,

$y, \hat{y}$: matrices containing y_{ij} and $\hat{y}_{ij}$, respectively

S : selection of records for manual editing

w_i : weight of record i

Y_j : weighted publication cell total of variable i, $Y_j = \sum_i w_i y_{ij}$.

9. We prefer to have one overall plausibility indicator that determines whether a record is either edited manually, or automatically.

B. Partial plausibility indicators

10. Another purpose of plausibility indicators is to assist a statistical analyst in the interactive editing of a record. Besides an overall plausibility indicator (OPI), seven partial plausibility indicators (PPI) are used to assess the plausibility of a specific part of the questionnaire. Each (partial) plausibility indicator can attain a discrete number between 0 (very implausible) and 10 (completely as expected).

11. For the construction of plausibility indicators we need to distinguish two important aspects: influence and risk. The influence component quantifies the relative influence of a record on a publication total. The risk component quantifies either the extent in which a record is filled in properly or the extent in which it deviates from reference values. These reference values should be close to the expected clean value of a variable.

12. A questionnaire for an annual structural business statistic at Statistics Netherlands has four important parts, namely an employed persons block **A**, a business profit block **B**, a business costs block **C**, and a business results block **D**. For each block a partial plausibility formula (PPF) is constructed that has a range of $[0, \infty)$ and mainly measures the influence component. The reference values are the population median and population total for a subset of variables in a block. These reference values are computed on the basis of clean records of last year for the same editing cell.

13. A record will obtain a high value for the partial plausibility formula for a block, when values of variables in a block differ much from the corresponding population medians. That is, a block PPF has a high value when values of variables in a record are either small or large compared to corresponding population medians. Raw values that are relatively small can then also cause a high PPF. This is advisable, because it might be unjust that a raw value has a relatively small influence on a publication total. In a block PPF the sample weight of a record is also taken into consideration.

14. Besides a PPF for each block (PPF **A**, **B**, **C**, and **D**), three other PPF's (**E**, **I**, and **Q**) that mainly measure the risk component for variables across all four blocks are used. PPF **E** uses external information, like VAT-information and turnover from short-term statistics. Information from clean structural business records of last year is also used for four important variables. This PPF differs from the other six in that the reference values only relate to the specific company. The larger the distance between raw values and reference values, the larger the risk that raw values are wrong.

15. Other important tools for editing records are indicators. An indicator is a quotient of two variables, for instance turnover divided by number of employees. PPF **I** compares an indicator based on raw records with the median of the indicators based on clean records of last year. Seven indicators are used for the comparison. When these indicators differ much from corresponding medians PPF **I** will have a large value.

16. Finally, PPF **Q** is used to assess the quality of filling in. This is the only PPF that considers all variables in the questionnaire. The number of empty entries and the number of violated edit rules are counted for a specific record. When these numbers are high then PPF **Q** attains a large value.

IV. CALIBRATING PARTIAL PLAUSIBILITY INDICATORS

A. Mark limits

17. Partial plausibility formulas have a range of $[0, \infty)$ and they are transformed to corresponding partial plausibility indicators, which can attain discrete values between 0 and 10. In The Netherlands marks that are given on schools vary between 0 and 10, where a 0 is very bad, a 10 is excellent and all marks below 6 are insufficient. Everyone therefore has developed a feeling for these marks. The PPF's are transformed to marks by means of mark lower limits. These lower limits can vary across PPF's and edit cells. In Table 1 an example is given of a set of mark lower limits. The resulting PPI's are PPI **A**, **B**, **C**, **D**, **E**, **I**, and **Q** respectively.

Table 1. Lower mark limits for NACE 52110, number of employees 0-9 and PPI A.

lower limit	mark	lower limit	mark
0	10	8,61	4
1,35	9	9,87	3
1,82	8	13,15	2
2,38	7	14,20	1
3,47	6	16,10	0
6,29	**5**		

18. Partial plausibility indicators are calibrated by the determination of mark limits. These limits are computed using raw and clean data of last year for the same edit cell.

B. Sufficiency limit

19. The lower limit for mark five is important, because it determines whether a PPI has a sufficient mark. This limit is also referred to as the sufficiency limit. For the determination of lower limits we make use of the empirical cumulative distribution function of a PPF for both clean and raw data of last year.

20. We start with the cumulative distribution function of a specific PPF for clean data. The main part of the clean data should have a PPI of at least six. Ideally, clean records are close to reference values and each PPF has a small value for those records. On the other hand, we would like to select influential records, which might be clean. We define $P_{\geq 6}^{clean}\%$ as the percentage clean forms that have at least mark six for a specific PPI. In practice we often choose $P_{\geq 6}^{clean}\% = 90\%$.

21. The procedure below is followed for each edit cell provided that there are at least 50 clean records available. Otherwise, some publication cells are combined into one new publication cell and the limits are determined for the resulting aggregated edit cells.

22. For every PPF:

- Determine the quantile corresponding to $P_{\geq 6}^{clean}\%$ on the basis of clean data, that is, $CDF_{clean}^{-1}\left(P_{\geq 6}^{clean}\%\right)$, where CDF_{clean}^{-1} is the inverse cumulative distribution function of a PPF for clean data. This quantile equals the sufficiency limit.
- Determine $P_{\geq 6}^{raw}\%$ from $CDF_{raw}^{-1}\left(P_{\geq 6}^{raw}\%\right) = CDF_{clean}^{-1}\left(P_{\geq 6}^{clean}\%\right)$, where CDF_{raw}^{-1} is the inverse cumulative distribution function of the PPF for raw data. Hopefully, clean data are more close to the reference values than raw data. In that case $P_{\geq 6}^{clean}\%$ will be larger than $P_{\geq 6}^{raw}\%$, because $CDF_{raw}(PPF)$ will lie on the right of $CDF_{clean}(PPF)$. In Figure 1 an example is given; $P_{\geq 6}^{raw}\%$ is about 84% when $P_{\geq 6}^{clean}\%$ is 90%. In practice $P_{\geq 6}^{raw}\%$ is often considerably smaller than 90%.

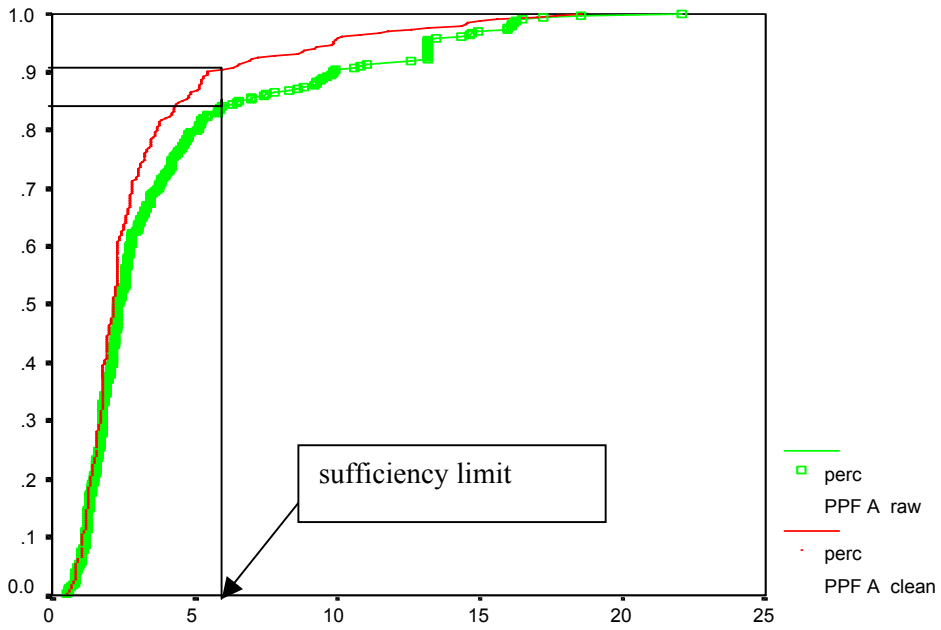

Figure 1. Cumulative distribution functions of PPF **A** for clean and raw data.

23. Suppose that for many records in a specific edit cell the difference between clean data and raw data was large for block **A** in the foregoing year. In that case the sufficiency limit will be such that a large percentage of the raw records has an insufficient mark for PPI **A**. If the raw data for the present year are

of the same quality as last year then in general many records in the edit cell will again have an insufficient mark for PPI **A**.

C. Remaining mark limits

24. For the construction of other limits than the sufficiency limit we solely make use of raw data of last year. For each PPF and edit cell, the sufficient and insufficient marks are distributed uniformly across the available raw forms. That is, the lower limits are determined such that the number of raw forms with marks 0, 1, 2, 3, 4, and 5 is about the same, and that the number of raw forms with marks 6, 7, 8, 9, en 10 is about the same.

25. Given the sufficiency limits, the remaining limits are determined on the basis of the following percentiles:

$$P^{raw}_{\geq c}\% = c / 6\, P^{raw}_{\geq 6}\%,\ \ c=1,2,...,6,$$

$$P^{raw}_{\geq c}\% = P^{raw}_{\geq 6}\% + (\text{c-6}) / 5\ \ (100\% - P^{raw}_{\geq 6}\%),\ \ c=7,8,9,10.$$

The lower limit for mark c now equals $CDF^{-1}_{raw}\left(P^{raw}_{\geq c+1}\%\right)$.

V. COMPUTATION OF THE OVERALL PLAUSIBILITY INDICATOR

26. In our view the mechanism that determines the overall plausibility indicator should satisfy four criteria:

i) It is rather simple
ii) The OPI cannot change dramatically due to small changes in a PPI
iii) An increasing PPI cannot cause the OPI to decrease
iv) A low mark for a PPI has a large influence on the OPI

27. The overall plausibility indicator that is used is a truncated weighted mean of the seven partial plausibility indicators, where the weights are such that a lower mark has a larger weight. Furthermore, the weights are optimised in the sense that a lower mark has a maximal larger weight, without violating criterion iii). The resulting formula is

$$\text{OPI} = \left\lfloor \frac{\sum_{i=0}^{10} i n_i s_i}{\sum_{i=0}^{10} n_i s_i} \right\rfloor,$$

where n_i is the number of PPI's with value i, s_i is the weight of value i, and $\lfloor \ldots \rfloor$ means that the value is truncated. The weights s_i are given in Table 2.

Table 2. Weights s_i

i	s_i	i	s_i
10	1	4	9/3
9	9/8	3	9/2
8	9/7	2	9
7	9/6	1	18
6	9/5	0	36
5	9/4		

VI. CONCLUSIONS

28. At Statistics Netherlands a uniform statistical process for annual structural business statistics has been developed. In this new process, a partition is made between records for automatic editing and records for manual editing. Records that are labelled as crucial are always edited manually. Non-crucial

records with an insufficient overall plausibility indicator are also edited manually. The remaining records are edited automatically by means of SLICE.

29. Partial plausibility indicators are used to assess the plausibility of specific parts of a questionnaire. These indicators assist a statistical analyst in locating errors in a record. Every year the partial plausibility indicators are calibrated before the new annual structural business statistics are edited. For this calibration clean and raw data of annual structural business statistics of last year are used.

References

Jong, A.G., 2002, UniEdit: Standardised processing of structural business statistics in The Netherlands. Statistics Netherlands, Voorburg.

Waal, T. de, 2000, SLICE: generalised software for statistical data editing and imputation. In: Proceedings in computational statistics 2000 (ed. J.G. Bethlehem and P.G.M. van der Heijden), Physica-Verlag, Heidelberg, pp. 277-282.

Waal, T. de, and Wings, 1999, From CherryPi to SLICE. Report BPA-no 461-99-RSM, Statistics Netherlands, Voorburg.

Granquist, L., 1995, Improving the Traditional Editing Process. In: Business Survey Methods (ed. Cox, Binder, Chinnappa, Christianson, and Kott), John Wiley & Sons, pp. 385-401.

Granquist, L. and J. Kovar, 1997, Editing of Survey Data: How Much is Enough? In: Survey Measurement and Process Quality (ed. Lyberg, Biemer, Collins, De Leeuw, Dippo, Schwartz, and Trewin), John Wiley & Sons, pp. 415-435.

Hidiroglou, M.A., and J.-M. Berthelot, 1986, Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12, pp. 73-83.

Appendix A: Different types of cells for selective editing

<table>
<tr><th rowspan="2">company size</th><th rowspan="2">number of employees</th><th colspan="3">NACE</th></tr>
<tr><th colspan="2">publication cell (=NACE 52121+52122)</th><th>publication cell (=NACE 5263)</th></tr>
<tr><td></td><td></td><td>52121</td><td>52122</td><td>5263</td></tr>
<tr><td rowspan="4">Small</td><td>0</td><td rowspan="4" colspan="2">editing cell</td><td></td></tr>
<tr><td>1</td><td></td></tr>
<tr><td>2-4</td><td>sample cell</td></tr>
<tr><td>5-9</td><td></td></tr>
<tr><td rowspan="3">Medium</td><td>10-19</td><td colspan="2"></td><td rowspan="3">editing cell</td></tr>
<tr><td>20-49</td><td>sample cell</td><td>sample cell</td></tr>
<tr><td>50-99</td><td colspan="2"></td></tr>
<tr><td rowspan="3">Large</td><td>100-199</td><td colspan="2"></td><td></td></tr>
<tr><td>200-499</td><td colspan="2"></td><td></td></tr>
<tr><td>> 499</td><td colspan="2"></td><td></td></tr>
</table>