CRP.2
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

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

SELECTIVE EDITING USING PLAUSIBILITY INDICATORS AND SLICE

Chapter

Submitted by Statistics Netherlands, the Netherlands[1]

Abstract: At Statistics Netherlands the statistical process for annual structural business statistics has been redesigned. Within the scope of IMPECT (IMPlementation 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. Plausible records are edited automatically with the computer program SLICE 1. In this chapter we first discuss the construction and calibration of plausibility indicators. We then examine 12 of the 54 publication cells of the annual structural business statistics 2000 Trade and Transport that have been edited selectively. Records that were edited automatically have later been edited manually for evaluation purposes. We check for differences between raw, manually edited, and automatically edited data and see the influences of selective editing on the publication totals. The plausibility indicators do not detect all influential errors. However, differences between manually and automatically edited data are small for most variables. Sometimes there are greater differences, also for key variables. In some publication cells for Transport we found major deviations. We expect that these differences can be reduced by improving the plausibility indicators and SLICE.

I. INTRODUCTION

1. At Statistics Netherlands a large reorganisation has been carried out. As a result, 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**.

An important reason for creating a new organisation is that a uniform statistical process was required for structural business statistics, cf. De Jong (2002). Furthermore, the same work had to be done by fewer people of a higher educational level. The editing process is often a demanding and costly process. It was therefore decided to switch over to selective editing, cf. Granquist (1995), Granquist & Kovar (1997),

[1] Prepared by Jeffrey Hoogland (jhgd@cbs.nl)

Hidiroglou & Berthelot (1986). De Jong (2002) gives a detailed description of UniEdit 1, the new micro editing process for annual structural business statistics (ASBS) at Statistics Netherlands.

2. In this chapter two steps in this new process are highlighted, namely the partition of records for automatic and manual editing, and automatic editing itself. The emphasis is on selective editing of ASBS. 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, because in the latter case edit rules are checked directly.

3. In Section 2 the principles of selective editing and automatic editing are discussed and the new process for micro editing is highlighted. In Section 3 the purposes of plausibility indicators are stated. Partial plausibility indicators (PPI) need to be calibrated on the basis of raw and edited data of last year, before they can be used to make a selection of records that are implausible. In Section 4 the technique that is used to calibrate PPI's is explained. The computation of the overall plausibility indicator (OPI) is discussed in Section 5. Records with a low OPI are edited manually. In Section 6 the evaluation of plausibility indicators on the basis of raw and edited data is considered. The evaluation of SLICE 1 is treated in Section 7. Deviations as a result of selective automatic editing are treated in Section 8. Finally, some conclusions are drawn in Section 9.

II. Selective and automatic editing

A. Principles

4. First, some terminology is introduced based on Rivière (2002). The *right* value is the value that would be obtained with an ideal measurement process. A value is *correct* if a subject-matter specialist would agree to leave it unchanged in the database. We therefore assume that a record is correct when it has been edited manually. A value is *acceptable* if the data-editing program accepts it in the database without any requirement to check it manually or adjust it automatically. A record is acceptable when it has been edited. A value is *raw* when it has not been checked at all.

5. Some types of errors are obvious, such as uniform 1000-errors (all financial variables are a factor 1000 too large), erroneous negative values, and empty (sub)totals. Ideally these types of errors should bother neither an editor nor an advanced program for automatic editing. These errors should be edited automatically in an early stage. Since they are obvious a simple program suffices to do the job.

6. In our view 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. 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. So, we use the principle of selective editing, cf. Granquist (1995), Granquist & Kovar (1997), Hidiroglou & Berthelot (1986). Plausibility indicators are used to decide whether a record can be edited automatically.

7. To obtain acceptable data it is sufficient to edit records automatically. When at least one edit rule is violated a record is *doubtful* and values of variables should be changed by means of imputation. At Statistics Netherlands automatic editing of ASBS is done with SLICE 1. It requires a description of the data, edit rules, confidence weights that give an indication of the reliability of each variable, and imputation rules. The generalised Fellegi-Holt paradigm for localisation of errors is used. 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.

B. UniEdit

8. UniEdit 1 is the latest statistical process for selective micro-editing of ASBS of Statistics Netherlands. For many branches of industry it was first applied to the ASBS for the year 2000. UniEdit aims at a uniform editing process that is identical for all branches of industry so that the efficiency of production can be optimised.

9. After removing obvious mistakes, it is determined whether a record must be edited by hand or automatically. For this partition several plausibility indicators are used. The records are then changed by either editors or by SLICE 1 in such a way that all editing rules are satisfied.

10. UniEdit 2 is the latest statistical process for weighting and macro-editing of ASBS of Statistics Netherlands. It includes unit-imputation of large companies that did not respond, automatic outlier detection, weighting with auxiliary information, and validation of figures. Weights of records can be adjusted manually in the validation step when figures do not seem plausible.

III. Plausibility indicators

A. Selecting records for manual editing

11. For selective editing, we need reliable estimates of correct values of variables in a record. Records are therefore grouped into relatively homogeneous subgroups. Such a group is called an edit cell, which is an intersection of a publication cell and a company size class, see figure 2 in Appendix A. For each edit cell the plausibility of each record is assessed.

12. A plausibility indicator (PI) 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 estimated population total should be selected for manual editing. Let's assume that *P%* of the records is to be selected and *(100-P)%* of the records is not edited. The selection should be such that a distance function, based on raw and correct values of variables for plausible records, is minimal. We first compute a sum of weighted differences between raw and correct values for variable y_j, relative to the (weighted) publication cell total for variable y_j

$$\Delta_j(y^c, y^r, S_{PI}) = \left| \frac{1}{Y_j} \sum_{i \notin S_{PI}} w_i (y_{ij}^r - y_{ij}^c) \right| \quad , \qquad (1)$$

where

y_{ij}^c, y_{ij}^r : correct, respectively raw value (without obvious mistakes) of variable j in record i ;

y^c, y^r : matrices containing y_{ij}^c and y_{ij}^r, respectively ;

S_{PI} : selection of records for manual editing made by a plausibility indicator;

w_i : weight of record i ;

Y_j : estimated publication cell total of variable j, $Y_j = \sum_i w_i y_{ij}^c$.

The main problem is that y_{ij}^c is not known beforehand. It can be estimated using y_{ij}^c from last year, for instance multiplied by an estimate for the economic growth or inflation rate. The final weight voor record i is also not known at the beginning of the editing process. It can be estimated by the inverse of the inclusion probability π_i.

13. We prefer to have an overall plausibility indicator that determines whether a record is either edited manually, or automatically. This OPI should for example minimise

$$\sum_{j=1}^{J} a_j \Delta_j(y^c, y^r, S_{OPI}), \qquad (2)$$

subject to the constraint that *P%* of the records are to be selected, where

a_j : denotes the importance of variable y_j ;

J : is the set of variables on a questionnaire.

In this chapter we will make use of (1), not of (2). Furthermore, we will only use (1) for evaluation purposes when the editing process is finished. The selection percentage *P* is then not a predetermined but an observed value.

B. Partial plausibility indicators

14. Another purpose of plausibility indicators is to assist an editor in the interactive editing of a record. Besides an overall plausibility indicator, 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).

15. 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 an estimated publication total. The risk component quantifies either the extent in which a questionnaire is filled in properly or the extent in which it deviates from reference values. These reference values should be close to the correct value of a variable.

16. 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, ∞) and measures both influence and risk. The resulting PPF equals formula (9) in Appendix B. The reference values are the estimated population median and total for a subset of variables in a block. The risk is incorporated by the difference between observed values y_{ij} and the corresponding estimated population medians m_{cj}. The influence is mainly included through the inverse of the inclusion probability π_i and estimated population totals. These totals are estimated on the basis of edited records of last year for the same edit cell. Population medians are estimated on the basis of edited records of last year for the same median cell. A median cell is an intersection of a publication cell and company size, see Appendix A.

17. 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 block PPF. This is advisable, because it might be unjust that a raw value for one company is much smaller than raw values for other companies within a supposedly homogeneous group. Furthermore, a block PPF has a high value when a record contains many empty or zero entries for variables within this block for which this is considered to be unlikely.

18. Besides a PPF for each block (PPF **Block A**, **B**, **C**, and **D**), three other PPF's (**External**, **Indicators**, and **Quality**) are used that measure risk for variables across all four blocks. PPF **External** uses external information, like VAT-information and turnover from short-term statistics. Information from edited structural business records of last year is also used for four important variables. PPF **External** equals formula (10) in Appendix B. This PPF differs considerably from the other six, because reference values only relate to the specific company. It is assumed that the larger the ratio of raw values and reference values, the more likely that raw values are wrong.

19. Other important tools for editing records are ratios of two variables, for instance turnover divided by number of employees. PPF **Ratios** compares ratios within a raw record with medians for corresponding ratios based on edited records of last year. Seven ratios are used for this comparison. PPF **Ratios** equals formula (11) in Appendix B. If the values of ratios within a record differ much from corresponding medians then PPF **Ratios** will have a large value.

20. Finally, PPF **Quality** (formula (12) in Appendix B) is used to assess the quality of filling in questionnaires. 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 raw record before obvious mistakes are corrected. If these numbers are high then PPF **Quality** attains a large value.

IV. Calibrating partial plausibility indicators

A. Mark limits

21. Partial plausibility formulas have a range of [0, ∞) 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 in schools vary between 0 and 10, where 0 is very bad, 10 is excellent, and all marks below 6 are insufficient. Everyone therefore has developed a feeling for these marks. PPF's are transformed to marks by means of mark upper limits. These upper limits can vary across PPF's and edit cells. In table 1 an example is given of a set of mark upper limits. Note that if PPF **A** is lower than 6,29 then PPI **A** has a sufficient mark. When the upper limits are determined for each PPF and edit cell the resulting PPI's are PPI **Block A**, **B**, **C**, **D**, and PPI **External**, **Indicators**, and **Quality** respectively.

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

upper limit	Mark	upper limit	mark
1.35	10	8.61	5
1.82	9	9.87	4
2.38	8	13.15	3
3.47	7	14.20	2
6.29	**6**	16.10	1
		¥	0

22. Partial plausibility indicators are calibrated by the determination of mark limits. Mark limits for year *t* are computed using raw data of year *t-1* and edited data of both year *t-1* and year *t-2* for the same edit cell, as is shown below.

B. Sufficiency limit

23. The upper limit for mark six is important, because it determines whether a PPI obtains a sufficient mark. This limit is also referred to as the *sufficiency limit*. For determination of upper limits we make use of the empirical cumulative distribution function (*ECDF*) of a PPF for both edited and raw data of year *t-1*. Note that manually edited data is correct by definition and data that has been edited selectively is a mixture of correct and acceptable data.

24. For calibration of PPF **Block**, **External**, and **Indicators** we start with the *ECDF* of edited data for year *t-1*. The main part of this data should have a PPI of at least six. Ideally, values of variables in edited records are close to reference values and each PPF has a small value for those records. However, some outliers will often be present, because financial variables have skew distributions. Furthermore, we would like to select influential records in any case, although they might be correct. We define $P_{\geq 6}^{edited}\%$ as the percentage edited forms that have at least mark six for a specific PPI. For calibration of year 2000 we have chosen $P_{\geq 6}^{edited}\% = 90\%$ for PPI **Block**, **External**, and **Indicators**.

25. For calibration of PPF **Quality** we only use raw data of year *t-1*. It will be very low for edited data, because violations of edit rules are not permitted for these data. The sufficiency limit is therefore based on $P_{\geq 6}^{raw}\% = D\%$, where *D%* is the average value of $P_{\geq 6}^{raw}\%$ for the other six PPF's within the same edit cell.

26. The procedure below is followed for each edit cell, provided that there are at least 50 edited records available. Otherwise, some publication cells are combined into an aggregated publication cell and mark limits are determined for the resulting aggregated edit cells. For every PPF, except PPF **Quality**:

1. Determine the quantile corresponding to $P_{\geq 6}^{edited}\%$ on the basis of edited data of year *t-1*, that is, $ECDF_{edited}^{-1}\left(P_{\geq 6}^{edited}\%\right)$, where $ECDF_{edited}^{-1}$ is the inverse empirical cumulative distribution function of a PPF for edited data. This quantile equals the sufficiency limit.

2. Determine $P_{\geq 6}^{raw}\%$ from $ECDF_{raw}^{-1}\left(P_{\geq 6}^{raw}\%\right) = ECDF_{edited}^{-1}\left(P_{\geq 6}^{edited}\%\right)$, where $ECDF_{raw}^{-1}$ is the inverse cumulative distribution function of the PPF for raw data of year *t-1*. Hopefully, edited data are more close to the reference values than raw data. In that case $P_{\geq 6}^{edited}\%$ will be larger than $P_{\geq 6}^{raw}\%$, because $ECDF_{raw}(PPF)$ will lie on the right of $ECDF_{edited}(PPF)$.

27. In Figure 1 an example is given; The upper curve is the *ECDF* of PPF Block A with edited *t-1*-data for NACE 52110. The quantile for which $P_{\geq 6}^{edited}\%$ is 90% equals 6.29. The lower curve is the *ECDF* of PPF Block A with raw *t-1*-data for NACE 52110. When we take 6.29 as the sufficiency limit then $P_{\geq 6}^{raw}\%$ is about 84%. That is, 84% of the raw data of year *t-1* for NACE 52110 has a PPI Block A of at least six. In practice $P_{\geq 6}^{raw}\%$ is often considerably smaller than 90%.

28. Suppose that the difference between edited data and raw data is large for many records in an edit cell, block A, and year *t-1*. Furthermore, let us assume that reference values are reliable. In that case the sufficiency limit will be such that a large percentage of raw records in year *t-1* has an insufficient mark for PPI **Block A**. Many records will then have an insufficient mark for PPI **Block A** in year *t* if the following two conditions hold

- raw data for year *t* are of the same quality as raw data for year *t-1*, that is, the differences between raw and edited data are about the same for both years
- differences between edited data for year *t* and year *t-1* are comparable with differences between edited data for year *t-1* and year *t-2*.

C. Remaining mark limits

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

$$P_{\geq c}^{raw}\% = P_{\geq 6}^{raw}\% + \frac{6-c}{c}\,(100\% - P_{\geq 6}^{raw}\%), \quad c=1,2,...,5,$$

$$P_{\geq c}^{raw}\% = \frac{11-c}{5} P_{\geq 6}^{raw}\%, \quad c=7,8,9,10.$$

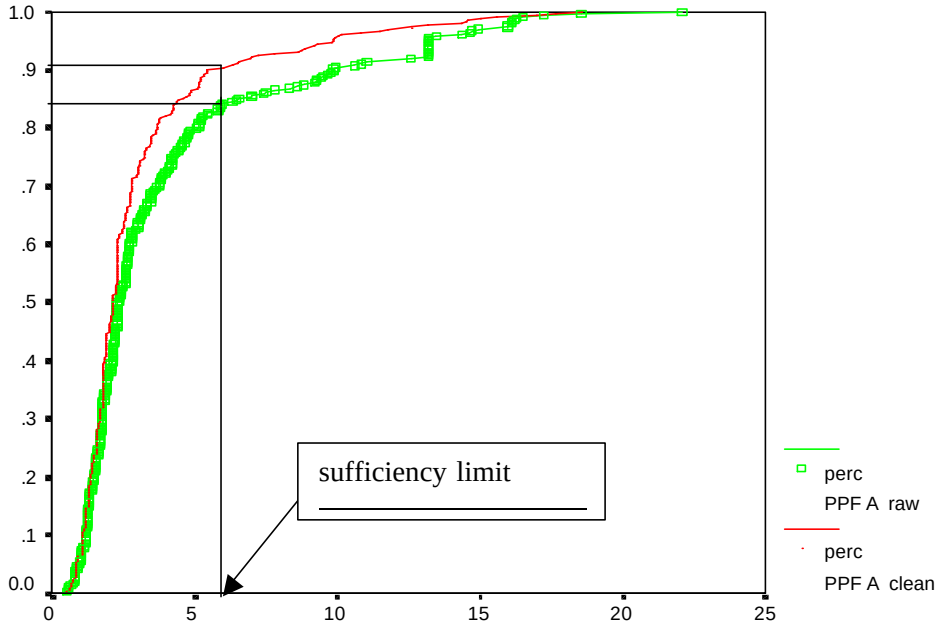

Figure 1. Empirical cumulative distribution functions of PPF **Block A** for edited and raw data (without obvious mistakes) for NACE 52110.

The upper limit for mark *c* now equals $ECDF_{raw}^{-1}\left(P_{\geq c}^{raw}\%\right)$. When all upper limits are determined the PPI's are calibrated for year *t*.

V. Computation of overall plausibility indicator

30. The overall plausibility indicator serves to select records that have to be edited manually. In our view the mechanism that determines the overall plausibility indicator should satisfy four criteria:

1. It is rather simple;
2. The OPI cannot change dramatically due to small changes in a PPI;
3. An increasing PPI cannot cause the OPI to decrease;
4. A low mark for a PPI has a large influence on the OPI.

31. The overall plausibility indicator for a record is a truncated weighted mean of the values of the seven partial plausibility indicators for this record, where the weights are such that a lower mark has a larger weight. The resulting formula is

$$\text{OPI} = \left\lfloor \frac{\sum_{v=0}^{10} in_v s_v}{\sum_{v=0}^{10} n_v s_v} \right\rfloor , \qquad (3)$$

where n_v is the number of PPI's with value *v* for a specific record, s_v is the weight of value *v*, and $\lfloor \ldots \rfloor$ means that the expression between brackets is truncated. The weights s_v are given in Table 2.

Table 2. Weights s_v

v	s_v	V	s_v
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. Evaluation of plausibility indicators

A. Data requirements

32. If we have raw and correct values for all records then plausibility indicators can be evaluated. However, as soon as we make use of selective editing we do not have correct values for all records. Records that have been edited automatically are acceptable, but they are not always correct. Ideally, in the traditional editing process raw and correct values were kept. However, at Statistics Netherlands raw values were often not saved in the past. In the year before selective editing was introduced, editors were therefore instructed to save raw values.

33. The ASBS of year 2000 are produced with a new statistical process. An important problem is that variables are not always comparable between year 1999 and year 2000, because of new questionnaires for year 2000. Variables of 1999 have to be transformed to the 2000 format. However, a one-to-one transformation does not exist for some variables. Raw and correct data of 1999 can therefore be unavailable in the right format. Furthermore, raw data can be missing because some editors ignored instructions to save raw values. Because of data problems and time pressure an extensive evaluation of plausibility indicators by means of 1999 data has not been performed.

34. Plausibility indicators are therefore evaluated by means of 2000 data. The main problem is that records that are edited automatically cannot be considered as correct. Records that have been edited automatically are therefore also edited manually. This can only be done for a small selection of publication cells due to the availability of editors. The performance of SLICE, Statistics Netherlands` module for automatic editing, can then also be evaluated for these publication cells. Furthermore, when records that have been edited manually are also edited automatically, it can be investigated whether the percentage of records that is edited automatically should be adjusted.

B. Evaluation criteria

35. The performance of OPI as a selector of influential errors for variable y_j and a specific edit or publication cell can be assessed by computing (1) when correct values of variables are either available or estimated. Suppose that $P\%$ of the records has to be edited manually. When (1) is small for that edit or publication cell then OPI makes a good selection regarding variable y_j. To determine whether OPI makes a good selection for all variables in a record we can compute (2). However, in this chapter we only discuss the performance of OPI for each variable separately. Note that we only consider undetected errors and that we do not notice errors that OPI does detect.

36. The performance of a PPI as a selector of influential errors for variable y_j can also be assessed by computing (1). S is then defined as the selection of records that have an insufficient mark for this PPI. Furthermore, we can compute (2) to assess the performance of a PPI as a selector of influential errors for a set of variables. J is then defined as the set of variables that is monitored by a specific PPI.

37. A disadvantage of (1) is that it does not take the number of records into account that are edited automatically. The less records are sent to SLICE, the better the value for (1). We therefore also use

evaluation criteria that take the number of records with a sufficient PPI or OPI into account. We first define the error in variable *j* for record *i* as

$$e_{ij} = y_{ij}^{r} - y_{ij}^{c} \quad .$$

For PPI **Block A, B, C, and D** we then use the following evaluation criterion for variable y_j

$$C_b^j = \frac{\sum_{i:PPI_b \geq 6} w_i \cdot |e_{ij}| / v_b}{\sum_i w_i \cdot y_{ij}^{c} / r} \qquad (4)$$

where

- PPI_b is PPI **Block** b , b = **A, B, C, or D**;
- v_b is the number of records in an edit cell that has a sufficient PPI **Block** b ;
- r is the response in the edit cell.

38. For each variable y_j and edit cell we consider the records with a sufficient PPI **Block** and we compute the average of the weighted absolute error. This average is compared with the average contribution of a correct value in a record to the estimated total of an edit cell. The score on criterion (4) is considered as bad if larger than 0.1. In this case we speak of *large unseen errors*.

39. For PPI **External** we use a different criterion, because this PPI serves to detect large relative deviations of raw values with external data. For each variable y_j and edit cell we consider the records with a sufficient PPI **External** and we compute the average relative deviation between raw and correct values

$$C_E^j = \frac{1}{v_E^j} \cdot \sum_{i:PPI_E \geq 6} \left(\frac{y_{ij}^{c}}{y_{ij}^{r}} + \frac{y_{ij}^{r}}{y_{ij}^{c}} - 2\right) \quad , \qquad (5)$$

where

- PPI_E is PPI **External**;
- v_E^j is the number of records with a sufficient PPI **External** where both y_{ij}^{c} and y_{ij}^{r} are unequal to zero and not empty.

40. For PPI **Ratios** we also use a different criterion, which considers the errors in ratios for records with a sufficient PPI **Ratios**. For each ratio r_j and edit cell we compute

$$C_R^j = \frac{1}{v_R^j} \cdot \sum_{i:PPI_R \geq 6} \left(\frac{r_{ij}^{c}}{r_{ij}^{r}} + \frac{r_{ij}^{r}}{r_{ij}^{c}} - 2\right) \quad , \qquad (6)$$

where

- PPI_R is PPI **Ratios**;
- r_{ij}^{c} is the value for ratio *j* in record *i* for manually edited data;
- r_{ij}^{r} is the value for variabele *j* in record *i* for raw data without obvious mistakes;
- v_R^j is the number of records with a sufficient PPI **Ratios**, where both r_{ij}^{c} and r_{ij}^{r} can be calculated.

41. Criteria (5) and (6) take relative sizes of errors into account. Small companies therefore have the same impact as large ones. The score on criteria (5) and (6) is considered as bad if larger than 0.2. In this

case we speak of *large unseen errors*. The value of 0.2 is for example attained on criterion (5) if 40% of the plausible records have a ratio that is a factor two too high.

42. Unlike the other partial plausibility indicators, PPI **Quality** is not calculated on the basis of raw values of variables, but on the basis of two characteristics of the questionnaire as a whole. It is therefore not useful to investigate unseen errors in separate variables and a criterion such as (4), (5) or (6) cannot be derived.

43. For OPI we use a comparable evaluation criterion as (4) for PPI **Block A, B, C, and D**. For each variable y_j and edit cell we compute

$$C_O^j = \frac{\sum_{i:OPI \geq 6} w_i \cdot |e_{ij}| / v_b}{\sum_i w_i \cdot y_{ij}^c / r} \quad . \qquad (7)$$

C. Selected publication cells

44. The returned questionnaires are divided into a number of publication cells depending on the SIC (Standard Industrial Classification; Dutch version of the NACE) of the respondent. These publication cells contain one or more SICs. During the editing process we no longer distinguish records with different SICs within a publication cell. This implies that the publication cells must be as homogenous as possible for the editing process to work well. In contrast with this is the demand that a publication cell must be big enough to allow the application of statistical methods.

45. The twelve publication cells used for evaluation are given in table 3. They contain 4651 records in total for ASBS 2000. Some 45% of these records have a sufficient PI. Because the PI was not ready in time for production, records were not edited automatically the first few months. Records that were edited automatically were later edited by hand in the course of this study. The records that were edited by hand during production were later edited automatically.

46. The enterprises in each publication cell are categorised by size class in three edit cells: size class 1 (less than 10 employees), 2 (from 10 to 100 employees), and 3 (at least 100 employees). Records in edit cells of size class three are completely edited by hand. These *crucial edit cells* contain major enterprises, which attribute so much to the published total, that it is very important that these enterprises have optimal quality data.The questionnaires consist of cover pages and inserted pages. Cover pages are less extensive for size class 1 than for size class 2-3, but they do not contain specific questions on publication cells. These are on the inserted pages. In the publication cells we studied there are five different inserted pages.

47. We encountered a problem when we re-edited the wholesale trade records in size class 1 by hand. We could not produce all necessary files for the interactive editing program. So for the corresponding edit cells we only have part of the records available that contain both raw and manually edited data. These edit cells were therefore eliminated from the study. After removing the records of wholesale trade size class 1 we had 4162 records left. All weighted totals were determined on the basis of these records.

Example

48. We give an example of the use of criterion (4) for PPI **Block B**. There are five important variables in this block of questions concerning operating profit. These variables have an importance weight a_j of 100, the remaining variables in Block B have an importance weight a_j of 0, see formula (9) in Appendix B.

49. In publication cell 163300 there are large (by PPI **Block B)** unseen errors for each important variable, see table 4. In some edit cells the average weighted unseen absolute error is even larger than the average contribution of a correct record to the estimated population total of the edit cell. That is, the

nominator of criterion (4) is larger than the denominator of criterion (4). For publication cells Wholesale trade and Retail trade there are only large unseen errors for other operating profit.

Table 3. The publication cells selected for evaluation.

	Wholesale trade
151200B	in flowers and plants
151300C	in food, beverages and tobacco excl. fruit, vegetables and potatoes
151600C	in tool, machinery for agriculture/textile production
	Retail trade
152110	in food, beverages and tobacco in shops; super markets
152121C	in furniture, household textile, lights, and household articles
152121E	in hardware, tools, paint and construction materials
	Transport
160220	Irregular transport of people by taxi
161100	Shipping at sea
161200	Inland shipping
163110	Loading, unloading, warehousing
163300	Travel organisation and mediation; information for tourism
163400	Shipping agents, cargo insurance and chartering brokers; weighing and measuring

Table 4. Results for criterion (4) for PPI **Block B**, for each important variable and non-crucial edit cell.

Publication cell	Size class	Net turnover	Net turnover for main activity	Net turnover for other activities	Other operating profit	Total operating profit
151200B	2	0.04	0.04	0.00	0.02	0.04
151300C	2	0.03	0.04	0.04	0.05	0.04
151600C	2	0.02	0.02	0.01	0.10	0.02
152110	1	0.01	0.01	0.00	0.00	0.01
	2	0.00	0.00	0.00	**0.11**	0.01
152121C	1	0.03	0.03	0.00	0.02	0.03
	2	0.00	0.00	0.00	**0.90**	0.01
152121E	1	0.02	0.02	0.00	0.00	0.02
	2	0.04	0.03	0.00	**0.24**	0.03
160220	1	0.00	0.00	0.05	0.05	0.00
	2	0.00	0.00	0.00	**0.10**	0.00
161100	1	0.01	0.01	0.00	0.02	0.01
	2	0.04	0.03	0.01	**0.30**	0.04
161200	1	0.01	0.03	0.02	0.00	0.00
	2	0.03	0.05	0.06	0.00	0.04
163110	1	0.00	0.01	**0.24**	0.00	0.01
	2	0.05	0.07	**0.16**	**0.24**	0.05
163300	1	**1.47**	**1.51**	0.00	0.00	**1.46**
	2	**0.38**	**0.38**	**0.51**	**0.33**	**0.38**
163400	1	0.01	0.01	0.00	0.02	0.01
	2	0.01	0.02	**0.12**	**1.52**	0.01

D. Evaluation results for plausibility indicators

50. The emphasis of our evaluation is on errors that are not detected by plausibility indicators PPI **Block**, PPI **External**, PPI **Ratios**, and OPI. We use evaluation criteria (1), (4), (5), (6) and (7) to assess their performance. We now summarize the performance of plausibility indicators according to the above mentioned criteria.

51. Table 4 shows that PPI **Block B** does not detect all large errors in variables for that block of questions. The same holds for other PPI **Block**. PPI **External** and PPI **Ratios** also miss some large errors for variables considered in these PPI's. PPI **External** mainly has a bad performance when external information is unavailable. However, most of the large unseen errors for one PPI are detected by some of the other six PPI's. The computation of OPI is such that a low mark for one PPI has a relatively large influence. A low mark for at least one PPI therefore often results in an insufficient OPI and the detected errors can then be solved by an editor.

52. The performance of a plausibility indicator varies across publication cells and variables. In general, it is better for important variables. These variables also have a larger weight in the computation of a plausibility indicator. Furthermore, important variables are often (sub)totals of entries in a questionnaire. These (sub)totals contain less substantial errors, because errors in underlying entries are often relatively small or cancelled out. In Table 7 in Appendix C results are given for OPI for three important variables: total number of employed persons, net turnover, and operating result. For total number of employed persons criteria (1) and (7) are acceptable. However, for publication cell 163300 criteria (1) and (7) have a large value for net turnover and operating result. This means that some large errors have not been detected by OPI. Furthermore, a large value for criteria (1) indicates that some of those errors are influential on the publication level and that they do not compensate each other.

53. In general, there are more unseen influential errors for publication cells Transport than for Trade. One of the reasons is that reference values are less accurate for Transport (Hoogland & Van der Pijll, 2003). Raw data for Transport also contain more errors than raw data for Trade. Transport companies seem to have more difficulties with understanding variable definitions on the uniform cover pages of questionnaires Trade & Transport. The main question becomes now whether SLICE 1 tackles unseen influential errors or not.

VII. EVALUATION OF SLICE 1

A. SLICE 1

54. When a record is considered plausible on the basis of the calculated OPI, it is edited with the SLICE 1 program developed at Statistics Netherlands. SLICE 1 contains the module CherryPi, cf. de Waal (2000) and De Waal & Wings (1999). This module localises errors in a record on the basis of a number of edit rules that show the mathematical relationships between variables. It also contains an imputation module, which can apply mean, ratio, or regression imputation in the case of erroneous values. Finally, it contains a module that corrects imputed values when they violate edit rules.

55. The edit rules violated by a record have to be satisfied by CherryPi by changing the values of one or more variables. Usually there are several possibilities for editing a record. This requires a choice, which is made according to the *principle of Fellegi and Holt*, see Fellegi & Holt (1976). The principle states that it is more likely that there is one major error in one variable, than that there are smaller errors in more than one variable. The best solution for editing a record is to change as few variables as possible.

56. In practise some variables contain fewer errors than others, or are not allowed to be changed as frequently as other variables. That is why each variable in CherryPi is weighted for reliability. This implies that a change in one variable can weight a number of times heavier than a change in another

variable. The best solution for the record is obtained by minimising the sum of the reliability weights of the variables changed, such that all edit rules are satisfied. The resulting constrained minimisation problem is solved with the Chernikova algorithm, cf. Chernikova (1965).

57. When several variables have the same weight, CherryPi often finds several equally good solutions. The output given by CherryPi contains all these solutions, from which one must be selected. In this case the first solution is always chosen. When many edit rules are violated, many variables must be changed. In such cases CherryPi cannot find a solution because there are too many possibilities to check and the record must be edited by hand anyway.

B. Available data

58. We have both manually and automatically edited data available for 4162 records, except for 247 records for which SLICE 1 could not find a solution. This is mainly due to the poor quality of these records, which means too many variables had to be adjusted. Although it may be possible to edit these records automatically anyway by improving SLICE 1, this is not the way we would go. The quality of these records is so low that they should be checked by an editor in any case. Table 5 shows that most of these records received an insufficient PI and would not end up with SLICE 1 during production. The percentage of records with a PI of four or more for which SLICE 1 could not find a solution was very small.

Table 5. Records for which SLICE 1 cannot find a solution, by OPI value.

OPI	Total number of records	No solution	
		number	percentage
0	222	66	29.7%
1	560	76	13.6%
2	367	30	8.2%
3	303	20	6.6%
4	364	14	3.8%
5	373	11	2.9%
6	395	12	3.0%
7	593	14	2.4%
8	685	4	0.6%
9	292	0	0%
10	8	0	0%
Total	4162	247	5.9%

59. The crucial edit cells contain 216 enterprises. Together with the group that cannot be edited automatically, there are 449 records (11% of the total in the twelve publication cells) that have to be edited by hand in any case.[2] This means that the percentage of automatically edited records cannot exceed 90%.

C. Publication totals in evaluation versus actual publication totals

60. Part of the manually edited data in this evaluation come from the production of ASBS 2000 and part were edited especially for this evaluation. We assume that the editing quality of the records later edited by hand is as good as the editing of the records that were originally edited by hand. There are still

[2] Because the two groups of records overlap somewhat, the total number is smaller than the sum of the sizes of the two individual groups.

differences between the ASBS 2000 production and our evaluation. This is mainly due to the following points:

1. The implementation of the plausibility indicators was not ready when manual editing was started for production. So at the start of the editing process, records with unknown OPI's went to the editors. Some of these may have had a sufficient OPI, and could have been edited automatically. In the analysis we assume that all records that would have had a sufficient OPI went to SLICE 1. So this means we could overestimate the inaccuracy of the published totals for ASBS 2000.

2. In the production of ASBS 2000 we used a new weighting method. However, during this analysis we used an old weighting method, because we were only interested in the trend breach as a consequence of the introduction of selective editing. In the old method direct weighting is done per SIC and company size on the basis of the population and survey sample sizes, excluding outliers. A number of records were already identified as outliers in the editing phase. These records were given a weighting factor of 1.

3. For the evaluation we used the production database as it was directly after micro-editing and before the UniEdit 2 process. So we did not have the outlier indications that were determined in that process.

61. The weighted totals we calculated are therefore not identical to the published totals. Furthermore, aggregated deviations between records that were edited manually and automatically, which are calculated in the next section, can differ from these deviations in practice.

62. Because more records were edited manually than was necessary on the basis of the OPI, and because the greatest deviations from the ASBS 1999 figures were smoothed out in the validation step, the bias in the published totals will be lower than the values we calculated. We are mainly interested in the effect of automatic editing as it will take place in the future. We therefore feel that it is more important to look at the ideal case, in which the PI is available from the very beginning, and no work has to be done during validation.

D. Evaluation criteria

63. The bias of a weighted total after selective automatic editing is the expected deviation from the real population total. Because this population total is unknown we cannot determine this bias. However, we are concerned with the difference between automatic and manual editing. So we can approximate the bias with the difference between the weighted total of 100% manually edited data and that of partly automatically edited data. We call this value the *pseudo-bias* due to selective automatic editing. It is expressed as a percentage of the manually edited weighted total.

64. The pseudo bias of SLICE 1 for variable y_j is assessed for a publication cell by computing

$$pb(y_j) = \frac{\sum_{\bar{S}} w_i (y_{ij}^a - y_{ij}^c)}{\sum_{C} w_i y_{ij}^c} \quad , \qquad (8)$$

where

$\bar{S}$: is the set of records that were edited automatically in a publication cell;
C : is the set of all records in a publication cell;
y_{ij}^a : is an automatically edited raw value;
y_{ij}^c : is an interactively edited raw value that was edited automatically in production.

Note that (8) looks similar to the expression between absolute signs in (1). The only difference is that we now use automatic edited (acceptable) values instead of raw values. In section VIII we will present the nominator in (8) as a percentage of the denominator in (8).

65. The pseudo-bias of SLICE 1 depends on the selection of records that are edited automatically (and therefore on the OPI) and on the quality of automatic editing with SLICE 1. The aim is to keep the influence of automatic editing on the weighted totals to a minimum. So the pseudo-bias must be small. However, it will differ across variables. The main reasons are:

- Some variables are given much more attention when the plausibility of a record is assessed. Therefore, SLICE 1 will mainly have to correct small errors for these variables.
- The number of times that SLICE 1 adjusts a raw value for variable y_j depends on the specified confidence weight, the number of empty/zero values for variable y_j and the number of times that variable y_j is involved in violated edit rules.

66. For variable y_j the percentage of records that are edited automatically might be increased when $|pb(y_j)|<d_j$, where d_j is determined on the basis of quality requirements. For a set of variables the percentage of records that are edited automatically might be increased when

$$\sum_{j=1}^{J} a_j \left| pb(y_j) \right| < \sum_{j=1}^{J} a_j d_j \cdot$$

67. On the other hand, we might argue that none of the *important* variables are allowed to have an unacceptable pseudo bias. In this case the percentage of records that are edited automatically can only be increased for an edit cell when

$$\forall\, y_j, j = 1,2,...,J, \quad \text{if } a_j \geq e_j \text{ then } |\, pb(y_j) \,| < d_j \ ,$$

where e_j denotes the minimal importance weight for a variable to be important.

VIII. DEVIATIONS AS A RESULT OF AUTOMATIC EDITING

A. Pseudo-bias as a result of selective editing

68. The pseudo-bias is calculated for all variables and twelve publication cells on the basis of the OPI as it could have been used in 2000. Due to problems, specified above, with the OPI calculation during production of ASBS 2000, the selection of automatically edited records in our study is not identical to the selection of automatically edited records during production.

69. There is a great deal of variation in pseudo-bias. For most variables the pseudo-bias is close to zero. This is because many variables are hardly changed by editors or SLICE 1. This is mainly true for less important variables. Variables that show major deviations, however, are also usually unimportant ones. For important variables the pseudo-bias is at most 15%. The pseudo-bias of 15% is caused by errors in the program that corrects obvious mistakes. These errors have been removed for ASBS 2001. Details of this research are given in Van der Pijll & Hoogland (2003).

70. Pseudo-biases in publication cells for transport are usually greater than those for trade. This is because questionnaires for transport are not filled in as well as those for wholesale and retail trade. One way to explain this is that cover pages of the questionnaires are the same for all publication cells trade and transport, while they are based on questionnaires for trade ASBS 1999. This may confuse respondents from transport, because definitions in variables may differ.

71. For some variables we found the cause of deviant editing by SLICE 1. These deviations were mainly found in the publication cells for transport. It will be possible to remove major deviations in these variables in the future by

- improving questionnaires (Bikker, 2003a; Bikker, 2003b; Van der Pijll & Hoogland, 2003);
- splitting heterogeneous publication cells (Van der Pijll & Hoogland, 2003);
- adapting the software that corrects obvious mistakes (Bikker, 2003a);
- improving plausibility indicators (Hoogland & Van der Pijll , 2003);
- adding a number of edit rules (Bikker, 2003a; Bikker, 2003b; Van der Pijll & Hoogland, 2003);
- adjusting reliability weights (Bikker, 2003a; Bikker, 2003b; Van der Pijll & Hoogland, 2003);
- improving the error localisation module within SLICE (De Waal & Quere, 2000; Quere, 2000);
- building an extra step in the statistical proces before SLICE, which removes systematic mistakes that do not follow the Fellegi-Holt principle (Bikker, 2003a; Bikker, 2003b).

B. The effect of more automatic editing

72. One key question is whether the percentage of records that is edited automatically can be increased or must be decreased. We can study this by varying the selection of records to be edited automatically. We can again calculate the pseudo-bias for each selection. This will generally be larger as the percentage of automatically edited records increases. In some cases the errors of the added automatically edited records can cancel out some of the existing deviation, so that the pseudo-bias is reduced. However, these are only incidental cases, and we should not count on them.

73. We varied the threshold for sufficient grades. This means we varied the number of records for which the OPI is considered sufficient. Table 6 shows the percentage of records that is automatically edited at the given threshold. When the threshold is six, as was the case for ASBS 2000, the percentage of automatically edited records is between 43 and 62 percent for most publication cells. It is impossible to determine the percentage in advance because the PPI are calibrated on the basis of raw and edited values of the previous year. Therefore the percentage of automatically edited records fluctuates for each publication cell. Apparently the PI was very severe for publication cell 152110, because only 28% of the records were deemed plausible enough for automatic editing. When the threshold is four the percentage of automatically edited records will exceed 60% in most publication cells, whereas a threshold of two will generally correspond to an automatic editing percentage from 72 to 90 percent.

74. In Appendix D we show how the pseudo-bias depends on the threshold for three key variables. Table 8 shows that it is very difficult to automatically edit the number of employed persons in publication cell 161200 (inland shipping) when the threshold is set at six. The deviation between automatic and manual editing is over 5% here. This is partly due to a systematic error made by the respondents, which is not always detected by the overall plausibility indicator. This systematic error cannot be corrected by SLICE 1. There are no significant problems in the other publication cells. When the percentage of automatically edited records exceeds 60%, the pseudo-bias in the total number of employed persons in publication cells 152121C and 163300 reaches more than 2%.

Table 6. Percentage of automatically edited records per threshold.

Publication cell	1	2	3	4	5	6	7
151200B	83%	72%	64%	60%	55%	48%	38%
151300C	83%	79%	73%	65%	58%	53%	44%
151600C	85%	74%	62%	52%	46%	43%	36%
152110	81%	65%	54%	46%	37%	28%	22%
152121C	88%	77%	68%	62%	53%	43%	41%
152121E	91%	82%	74%	65%	55%	48%	45%
160220	83%	73%	64%	59%	55%	44%	36%
161100	94%	90%	85%	80%	73%	62%	43%
161200	93%	86%	80%	77%	71%	62%	53%
163110	89%	81%	77%	77%	70%	61%	48%
163300	87%	72%	66%	59%	49%	38%	34%
163400	83%	72%	67%	62%	53%	44%	30%
All twelve cells	86%	76%	68%	61%	54%	45%	36%

75. The variable net turnover (table 9) is correctly edited automatically in almost all publication cells. Raising the percentage of automatically edited records to 80% causes virtually no problems for these variables. The only publication cell in which there is a major difference between manual and automatic editing is cell 163300 (travelling organisations and travel intermediation). This is because respondents often fill in some purchase value while the variable in this publication cell should almost always be zero. The editors usually remove the purchase value and balance it with net turnover, whereas SLICE 1 leaves the records unchanged. This leads to major pseudo-bias in these variables, because the overall plausibility indicator fails to detect some of those errors. The resulting bias cannot be avoided by applying extra edit rules. The problem will continue to show up in future, because the Fellegi-Holt principle does not hold. One long-term solution for this problem is an extra editing round focusing on specific errors such as these.

76. SLICE 1 does a fairly good edit of total operating result (table 10) for all publication cells. However, when SLICE 1 has to edit more records some difficulties show up in various publication cells. SLICE 1 cannot cope with all influential errors that are unseen by the overall plausibility indicator. For publication cell 163300 pseudo-bias occurs when the threshold is set at five or less. This corresponds to a percentage of records to be edited automatically of over 50%. In publication cell 163110 (loading, unloading, warehousing) the problems start around 75%.

77. The pseudo-bias for most variables and publication cells is small. It does not get much higher either when the number of records edited by SLICE 1 increases up to 70%. This is not true for variables with a large pseudo-bias in a publication cell. For these variables and publication cells we can see a rapid increase in pseudo-bias when the number of automatically edited records increases.

78. Tables 8-10 show several high percentages. This does not mean that these major deviations ended up in the published totals, though. The weighted totals of these variables may have been corrected by the automatic outlier detection, the new weighting method, or during validation, substantially reducing these deviations.

IX. CONCLUSIONS

79. At Statistics Netherlands a uniform statistical process for annual structural business statistics has been developed. In this new process, some obvious mistakes in raw data are edited automatically in an early stage. A partition is then made between records for further 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.

80. Partial plausibility indicators are used to assess the plausibility of specific parts of a questionnaire. These indicators assist an editor in locating errors in a record. Every year the partial plausibility indicators have to be calibrated before the new annual structural business statistics are edited. For this calibration correct and raw data of annual structural business statistics of last year are needed. Calibration of PPI's is not straightforward due to possible lack of correct and raw data.

81. When records that have been edited automatically are also edited manually and vice versa, the performance of the overall and partial plausibility indicators, and SLICE can be assessed. Furthermore, it can be investigated whether the percentage of records that is edited automatically has to be adjusted. We examined differences between manually and automatically edited weighted totals of ASBS 2000 prior to validation. We assumed that the plausibility indicator was operative during the entire editing period. The evaluation was made for twelve publication cells in the wholesale and retail trade, and in transport.

82. The effect of selective editing differs per variable. Most variables are hardly changed during manual or automatic editing. This is true both for less important variables and for some key variables such as total operating costs, total labour costs and total operating profits. The weighted totals for these variables hardly change in most publication cells when selective editing is used. Even when the percentage of automatic editing increases to 80% the deviation in the weighted totals for these variables stays under 2%.

83. However, for some variables the deviations of selective editing are large. These are mainly variables from the results block of the questionnaire, such as the result before taxes and the financial result. The current 45% threshold for automatic editing already yields many deviations of more than 5% in the weighted totals of these variables. When the percentage of records for automatic editing increases, the quality of these variables will plummet.

84. We found the greatest deviations in publication cells in transport. The problems are such that these publication cells will have to be edited more by hand, rather than less. The deviations in wholesale and retail trade are smaller. It depends on the level of bias in the published figures that is considered acceptable whether we can gain in efficiency by more automatic editing. The biases mentioned in this chapter are based on how the PI and SLICE 1 worked during ASBS 2000. There may well be less bias in most variables in the future when the PI and SLICE 1 are improved.

85. We have found room for improvement on a number of points. By adding a few edit rules, by developing software that removes systematic errors, and by improving the questionnaires we can come up with considerable improvements for the variables. Some of these improvements can be applied to large numbers of publication cells, also well beyond the twelve cells we studied here. Other improvements are publication cell specific (for instance pertaining to inland shipping and travel organisations). When these improvements are implemented those publication cells where SLICE 1 currently produces major deviations may well be edited without major problems.

References

Bikker, R.P., 2003a, *Evaluation of automatic versus manual editing of Annual Structural Business statistics 2000 Trade & Transport – additional explanations (In Dutch)*. Internal paper BPA-no 1900-03-TMO, Statistics Netherlands, Voorburg.

Bikker, R.P., 2003b, *Automatic editing of Annual Structural Business statistics 2000 Building & Construction branche: four structural problems with solutions (In Dutch)*. Internal paper BPA-no 2263-03-TMO, Statistics Netherlands, Voorburg.

Chernikova, N.V., 1965, Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. *USSR Computational Mathematics and Mathematical Physics*, **5**, pp. 228-233.

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.

Hoogland, J. J. en E.C. van der Pijll, 2003, *Evaluation of the plausibility indicator for production statistics 2000 Trade & Transport (In Dutch)*. Internal paper BPA-no 1971-03-TMO, Statistics Netherlands, Voorburg.

Jong, A.G., 2002, *UniEdit: Standardised processing of structural business statistics in The Netherlands*. Invited paper for UNECE Work Session on Statistical Data Editing, 27-29 May 2002, Helsinki, Finland.

Quere, R., 2000, *Automatic Editing of Numerical Data*. Report BPA-no 2284-00-RSM, Statistics Netherlands, Voorburg.

Rivière, P., 2002, *General principles for data editing in business surveys and how to optimize it*. Contributed paper for UNECE Work Session on Statistical Data Editing, 27-29 May 2002, Helsinki, Finland.

Van der Pijll, E.C. en J. J. Hoogland, 2003, *Evaluation of automatic versus manual editing of annual structural business statistics 2000 Trade & Transport (In Dutch)*. Internal paper BPA-no 286-03-TMO, 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 Quere, R., 2000, *Error localisation in Mixed Data Sets*. Report BPA-no 2285-00-RSM, Statistics Netherlands, Voorburg.

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

Appendix A. Different types of cells for selective editing of ASBS

<table>
<tr><td rowspan="3">company size class</td><td></td><td rowspan="3">number of employees</td><td colspan="3">NACE</td></tr>
<tr><td rowspan="2">company size</td><td colspan="2">publication cell (=NACE 52121+52122)</td><td>publication cell (=NACE 5263)</td></tr>
<tr><td></td><td>52121</td><td>52122</td><td>5263</td></tr>
<tr><td rowspan="4">Small</td><td>0</td><td>0</td><td rowspan="4" colspan="2">edit cell</td><td>median cell</td></tr>
<tr><td>1</td><td>1</td><td></td></tr>
<tr><td>2</td><td>2-4</td><td>sample cell</td></tr>
<tr><td>3</td><td>5-9</td><td></td></tr>
<tr><td rowspan="3">Medium</td><td>4</td><td>10-19</td><td colspan="2"></td><td rowspan="3">edit cell</td></tr>
<tr><td>5</td><td>20-49</td><td>sample cell</td><td>sample cell</td></tr>
<tr><td>6</td><td>50-99</td><td colspan="2"></td></tr>
<tr><td rowspan="3">Large</td><td>7</td><td>100-199</td><td colspan="2">median cell</td><td></td></tr>
<tr><td>8</td><td>200-499</td><td colspan="2"></td><td></td></tr>
<tr><td>9</td><td>> 499</td><td colspan="2"></td><td></td></tr>
</table>

Figure 2. Cells for selective editing of ASBS, which are combinations of NACE and company size.

Appendix B: Partial plausibility formulas

The partial plausibility formulas below are computed for each record in each edit cell. The quantities in the formulas can vary across edit cells.

PPF **Block** b , b = **A**, **B**, **C**, and **D** are computed as follows

$$\text{PPF}_i\ \textbf{Block} = \sqrt{\frac{\sum_{j=1}^{J_b} a_j \cdot \left(\frac{y_{ij} - m_{cj}}{Y_j}\right)^2}{p_i^{\,2} \cdot \sum_{j=1}^{J_b} a_j \cdot \left(\frac{m_{cj}}{Y_j}\right)^2}}\ , \qquad (9)$$

where

- y_{ij} , $j = 1, 2, \ldots, J_b$, are entries for business unit i and year t that are considered for a specific block;
- m_{cj} , $j = 1, 2, \ldots, J_b$, are the corresponding population medians for year *t-1* and median cell c containing business unit i in year t;
- Y_j , $j = 1, 2, \ldots, J_b$, are the corresponding weighted edit cell totals for year *t-1;*
- a_j , $j = 1, 2, \ldots, J_b$, denotes the importance of variable y_j, which can differ across edit cells;
- p_i is the inclusion probability of business unit i in year t.

If an entry y_{ij} is equal to 0 or empty then $y_{ij} = e_{cj} \cdot m_{cj}$ is used instead, where e_{cj} is the empty entry factor of variable y_j. If m_{cj} = 0 then m_{cj} is first given the value 1 and y_{ij} is then given the value e_{cj} . An empty entry factor is large when it is unlikely that the corresponding entry equals 0.

The PPF for external information is computed as follows

$$\text{PPF}_i\ \textbf{External} = \frac{\sum_{j=1}^{J_E} a_{ij} \cdot \sqrt{\frac{1}{2}\left(\frac{y_{ij}}{x_{ij}}\right)^2 + \frac{1}{2}\left(\frac{x_{ij}}{y_{ij}}\right)^2}}{\sum_{j=1}^{J_E} a_{ij}}\ , \qquad (10)$$

where

- y_{ij} , $j = 1, 2, \ldots, J_E$, are entries for business unit i and year t that are compared with external sources;
- x_{ij} , $j = 1, 2, \ldots, J_E$, are external sources for business unit i and year t ;
- a_{ij} , $j = 1, 2, \ldots, J_E$, denotes the importance of y_{ij} .

If y_{ij} or x_{ij} is equal to 0 or empty then a_{ij} = 0, otherwise a_{ij} = a_j . This is necessary because external information is often incomplete. If y_{ij} or x_{ij} is equal to 0 or empty for each $j = 1, 2, \ldots, J_E$ then PPF_i **External** = (upper limit for mark 6 + upper limit for mark 7) / 2 . That is, if there is insufficient information to calculate PPF **External** then PPI **External** equals six.

The PPF for ratios is computed as follows

$$\text{PPF}_i \textbf{ Ratios} = \frac{\sum_{j=1}^{J_R} a_j \cdot \sqrt{\frac{1}{2}\left(\frac{r_{ij}}{k_{cj}}\right)^2 + \frac{1}{2}\left(\frac{k_{cj}}{r_{ij}}\right)^2}}{\sum_{j=1}^{J_R} a_j} \quad , \qquad (11)$$

where

- r_{ij}, $j = 1, 2, ..., J_R$, are ratios for business unit i and year t;
- k_{cj}, $j = 1, 2, ..., J_R$, are the corresponding population medians for year t-1 and median cell c containing business unit i in year t;
- a_j, $j = 1, 2, ..., J_R$, denotes the importance of ratio r_j, which can differ across edit cells.

The PPF for quality of filling-out is computed as follows

$$\text{PPF}_i \textbf{ Quality} = \sqrt{\frac{w_{EMPTY} \cdot \left(\frac{\#EMPTY_i}{\#EMPTY_{MAX}}\right)^2 + w_{HARD} \cdot \left(\frac{\#VEDIT_i}{\#VEDIT_{MAX}}\right)^2}{w_{EMPTY} + w_{HARD}}} \quad , \qquad (12)$$

where

- w_{EMPTY} and w_{HARD} are weights for the two quality aspects;
- $\#EMPTY_i$ is the number of empty entries for business unit i in year t and $\#EMPTY_{MAX}$ is the total number of entries for an edit cell;
- $\#VEDIT_i$ is the number of violated edit rules for business unit i in year t and $\#VEDIT_{MAX}$ is an estimate of the maximum number of violated edit rules for an edit cell that will occur in practice.

Appendix C. Some evaluation criteria for the Overall Plausibility Indicator

Table 7. Evaluation criteria (1) and (7) for OPI, for total number of employed persons, net turnover, and operating result. Criterion (1) is printed in boldface if larger than 0.05 and criterion (7) is printed in boldface if larger than 0.1.

Publication cell	Size class	Number of employees		Net turnover		Operating result	
		Criterion (1)	Criterion (7)	Criterion (1)	Criterion (7)	Criterion (1)	Criterion (7)
151200B	2	0.01	0.06	0.00	0.00	0.00	0.00
151300C	2	0.01	0.04	0.00	0.00	0.01	0.00
151600C	2	0.00	0.00	0.00	0.00	0.01	0.00
152110	1	0.00	0.02	0.00	0.00	0.00	0.00
	2	0.00	0.01	0.00	0.00	0.00	0.00
152121C	1	0.01	0.10	0.00	0.00	0.03	0.00
	2	0.01	0.07	0.00	0.00	0.01	0.00
152121E	1	0.01	0.03	0.00	0.00	0.02	0.00
	2	0.00	0.00	0.00	0.00	0.00	0.00
160220	1	0.00	0.03	0.00	0.00	0.01	0.00
	2	0.02	0.08	0.00	0.00	0.00	0.00
161100	1	0.01	0.06	0.00	0.01	0.00	0.01
	2	0.00	0.01	0.00	0.00	0.00	0.00
161200	1	0.04	0.09	0.00	0.00	0.00	0.00
	2	0.00	0.03	0.00	0.05	0.00	0.04
163110	1	0.00	0.08	0.00	0.00	0.03	0.00
	2	0.01	0.05	0.00	0.01	**0.10**	0.00
163300	1	0.00	0.00	0.03	**0.70**	0.00	**0.69**
	2	0.00	0.01	**0.05**	**0.32**	**0.07**	**0.31**
163400	1	0.00	0.04	0.00	0.01	0.00	0.01
	2	0.01	0.05	0.00	0.00	0.05	0.00

Appendix D. Pseudo-bias in publication totals for several key variables

A column in table 8-10 shows the pseudo-bias resulting from automatic editing of records with a PI equal to the threshold or higher. When the pseudo-bias exceeds 5% it is printed in bold character. Pseudo-biases between 2% and 5% are underlined.

Table 8. Pseudo-bias in the total number of employed persons (in Block A).

	Threshold for plausibility indicator						
Publication cell	1	2	3	4	5	6	7
151200B	1.1%	1.1%	0.5%	0.6%	0.6%	0.1%	0.3%
151300C	0.3%	0.1%	0.2%	0.3%	0.2%	0.2%	0.1%
151600C	0.8%	0.2%	0.1%	0.1%	0.1%	0.1%	0.1%
152110	0.8%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%
152121C	3.6%	2.4%	2.4%	2.5%	1.8%	1.5%	1.4%
152121E	0.1%	0.5%	0.4%	0.5%	0.8%	1.0%	1.1%
160220	2.5%	0.7%	1.2%	1.0%	1.0%	0.9%	0.8%
161100	1.5%	1.4%	1.4%	1.4%	1.4%	1.4%	1.3%
161200	**7.7%**	**5.4%**	**5.1%**	4.1%	5.0%	**5.4%**	4.2%
163110	1.2%	1.4%	1.4%	1.4%	1.4%	1.4%	1.1%
163300	2.3%	2.1%	2.1%	0.5%	0.2%	0.0%	0.0%
163400	0.1%	0.2%	0.3%	0.1%	0.2%	0.3%	0.0%

Table 9. Pseudo-bias in net turnover (in Block B).

	Threshold for plausibility indicator						
Publication cell	1	2	3	4	5	6	7
151200B	0.2%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
151300C	0.2%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
151600C	0.2%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
152110	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
152121C	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
152121E	0.6%	0.2%	0.2%	0.1%	0.0%	0.0%	0.0%
160220	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
161100	**5.1%**	0.2%	0.0%	0.1%	0.1%	0.1%	0.1%
161200	0.8%	0.4%	0.4%	0.4%	0.4%	0.5%	0.0%
163110	0.6%	0.3%	0.3%	0.3%	0.2%	0.2%	0.2%
163300	**34.6%**	**21.6%**	**18.0%**	**16.8%**	**13.1%**	**8.7%**	**7.7%**
163400	0.4%	0.4%	0.4%	0.4%	0.2%	0.1%	0.1%

Table 10. Pseudo-bias in the operating result (in Block D).

	Threshold for plausibility indicator						
Publication cell	1	2	3	4	5	6	7
151200B	0.8%	0.2%	1.3%	1.1%	0.7%	0.6%	0.2%
151300C	0.4%	0.5%	0.5%	0.9%	0.1%	0.1%	0.1%
151600C	0.6%	0.2%	0.5%	0.3%	0.3%	0.2%	0.2%
152110	1.5%	1.7%	1.4%	0.3%	0.2%	0.2%	0.0%
152121C	**10.1%**	1.0%	2.4%	1.9%	1.9%	1.7%	1.7%
152121E	**12.7%**	0.8%	0.8%	0.3%	0.3%	0.1%	0.1%
160220	1.4%	1.3%	1.3%	1.2%	1.2%	2.1%	2.0%
161100	**8.2%**	2.7%	0.0%	0.3%	0.2%	0.5%	0.5%
161200	0.5%	0.2%	1.0%	1.2%	1.2%	0.3%	0.5%
163110	4.9%	**5.3%**	**5.3%**	**5.3%**	1.0%	0.9%	0.6%
163300	**8.0%**	**7.8%**	**8.0%**	**8.2%**	**8.3%**	0.6%	0.6%
163400	**39.5%**	0.8%	1.1%	1.1%	0.2%	1.1%	0.2%
