

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Ottawa, Canada, 16-18 May 2005)

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

**EVALUATION OF SCORE FUNCTIONS FOR SELECTIVE EDITING OF ANNUAL
STRUCTURAL BUSINESS STATISTICS**

Supporting Paper

Submitted by Statistics Netherlands¹

Abstract: Statistics Netherlands uses score functions to select records for manual editing of Annual Structural Business Statistics. Score functions that were used for ASBS 2000 were modified during the past few years. However, it was not clear whether these modifications actually improved error detection. This paper aims to evaluate some of the score functions used so far, using raw and edited data for nine publication cells for Retail trade and Transport. We evaluate three types of score functions that monitor variables within a questionnaire block and two types of score functions that monitor key variables through ratios. The size of errors in records with sufficient scores hardly depends on the type of block score function or the type of ratio score function. The bias due to selective editing does depend on the choice for either a block score function or a ratio score function. The accuracy of reference values seems to play a decisive role.

I. INTRODUCTION

1. Since the statistical year 2000 Statistics Netherlands has a uniform statistical process for most Annual Structural Business Statistics (ASBS). The editing phase is an important part of this process, because filled in questionnaires for ASBS contain many influential errors. One of the main problems is that uniform questionnaires are not always clear for the respondent. Not even the fact that financial variables have to be reported in thousands of euros in stead of in euros. Another problem is that respondents do not always have the data available in the required detail. The motivation of respondents can also be low, which results in questions that are misinterpreted, or not filled in at all.
2. The questionnaires for ASBS contain many important variables that are either published, or supplied to third parties such as Eurostat. However, they also contain a lot of variables that are not published. These additional questions might prevent that respondents forget to report anything, but they are also a burden for them.
3. Until ASBS 1999 all filled in questionnaires were edited by hand, which was very costly. In the nineties it became clear that selective editing of ASBS is possible and necessary. Possible, because not all

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

records contain influential errors in important variables (Granquist and Kovar, 1997; Van de Pol, 1997). Necessary, because Statistic Netherlands has to produce more output with less people.

4. Since a few years, score functions (also called plausibility formulas) are used to select records for manual editing of ASBS (Hoogland, 2002). The remaining records are edited automatically with SLICE 1, which contains the module CherryPi, cf. de Waal (2000), and De Waal and Wings (1999). Evaluation studies show that the score functions for ASBS 2000 do not detect all influential errors (Hoogland and Van der Pijll, 2003; Hoogland, 2005). Furthermore, SLICE does not correct all unseen errors. In the last two years a lot of effort has been put in improvement of SLICE (Van der Pijll and Hoogland, 2003; Bikker, 2003a; Bikker, 2003b; Bikker et al, 2004a; Bikker et al, 2004b). Nonetheless, score functions are still crucial to obtain acceptable quality of automatically edited data.

5. Most of the score functions used for ASBS 2000 were modified during the past few years. These modifications are made because of theoretical insights (Hoogland and Van der Pijll, 2003). However, it is not clear whether these modifications actually improve error detection. This paper aims to evaluate some of the score functions used so far.

6. In section 2 we will summarize the principles of selective editing and the past research on instruments for error detection. Score functions for ASBS and the underlying ideas are discussed in section 3. An evaluation of some of these score functions is described in section 4, evaluation results are given in section 5, and conclusions are made in section 6.

II. Selective editing

7. Selective editing instead of traditional editing of business statistics is one of the major improvements in editing in the past twenty years, cf. Granquist (1995), Granquist and Kovar (1997), and Hidirolou and Berthelot (1986). It is based on the insight that only a part of the records contains errors that can influence publication totals. For selective editing the emphasis is therefore on detection of influential errors.

8. There are two stages in the editing process. The first stage is named either micro editing, significance editing, or input editing. This stage starts as soon as records enter the editing process (Lawrence and McKenzie, 2000). Most of the instruments for error detection can already be used at this stage.

9. The second stage is called macro editing or output editing (Granquist, 1995), and mainly consists of outlier detection (Béguin and Hulliger, 2004; Hidirolou and Berthelot, 1986). It starts when most of the records have entered the editing process.

10. Several types of score functions were introduced to detect records with influential errors. We refer to Hidirolou and Berthelot (1986), Latouche and Berthelot (1992), Lawrence and McDavitt (1994), Farwell and Raine (2000), Farwell (2002), Farwell, Poole, and Carlton (2002), and Hedlin (2003).

11. Other instruments for error detection are also available, such as forward search (Chambers, Hentges and Zhao, 2004), classification and regression trees (Chambers, Hentges and Zhao, 2004; Sanders, 2002), logistic regression models (Madsen and Larsen, 2000; Van Langen, 2002), and neural networks (Nordbotten, 1995). Within the Euredit project many methods for detection of errors have been compared (Charlton, 2002; Charlton, 2004).

12. In this paper we will confine ourselves to score functions for ASBS. The degree of error detection of score functions is promising (Hedlin, 2003; Hoogland, 2005; Latouche and Berthelot, 1992; Van de Pol, 1997). Furthermore, score functions can be computed when a record enters the micro editing phase. This is important, because editors should do their job early to realize timeliness of the output.

III. Score functions

13. In the editing process the data have several stages. A value of a variable is *raw* when it has not been checked at all. Obvious errors in ASBS, such as uniform 1000-errors (all financial variables are a factor 1000 too large), erroneous negative values, and empty (sub)totals are edited automatically in an early stage. The resulting data are *raw data without obvious mistakes*. We assume that a record is *correct* when it has been edited manually.

14. At Statistics Netherlands score functions are used to assess raw data without obvious mistakes. They are used to estimate two characteristics of a record, namely influence and risk (Hoogland, 2002). It is important to quantify the relative influence of a record on an estimated publication total. It is also necessary to quantify the risk, that is, the extent in which the observed values deviate from what we would expect. We therefore have to estimate the correct values of variables that we publish, or at least estimate characteristics of groups of edited records. These estimates are referred to as reference values. Medians are often used as a robust estimate of the expected value of a variable for a group of records.

15. Score functions that assess both influence and risk are also proposed by Farwell (2002), Hedlin (2003), Hidioglou and Berthelot (1986), Latouche and Berthelot (1992), Lawrence and McDavitt (1994). According to Hedlin (2003) the type of score function has a small influence on the bias due to selective editing of the Monthly inquiry for the Distribution and Services Sector. However, some preliminary analyses of score functions for ASBS 2000, 2001, and 2002 of Statistics Netherlands indicated that changes in score functions for ASBS can have noticeable effects on the selection of records for manual editing. Latouche and Berthelot (1992) also show that the type of score function does matter.

16. The main aim of this paper is to compare current and earlier score functions for ASBS of Statistics Netherlands. So far, changes in score functions used for ASBS were hardly evaluated.

17. In Hoogland (2002) and Hoogland (2005) the performance of score functions is already investigated. The aim of these papers is to evaluate score functions in a production environment for ASBS 2000. In this setting there were a lot of problems, such as bad reference values, suboptimal parameters of score functions, and calibration errors. These were mainly start up problems, however, because of the new statistical process.

18. In this paper we also use ASBS 2000 to compute score functions, because this is the only statistical year for which we have fully manually edited data obtained with the new statistical process. However, this time we have a different approach with a more optimal environment that approaches the current production environment. Score functions now have different parameters and improved reference values. We will also calibrate score functions ourselves, because of calibration errors made in the production environment for ASBS 2000. Due to these changes the results can differ from the results in Hoogland (2002) and Hoogland (2005) for the same score functions.

IV. Design of evaluation of score functions for ASBS Retail trade and Transport

A. Publication cells

19. We consider nine publication cells for ASBS Retail trade and Transport. The classification of publication cells depends on the SIC (Standard Industrial Classification; Dutch version of the NACE). Publication cells contain companies of one or more SICs, see appendix A.

20. The publication cells used for evaluation are given in table 1. They contain 3429 companies that responded for ASBS 2000. According to the score functions used in production about 43% of them were suitable for manual editing. Records that were edited automatically were later edited by hand for evaluation purposes.

Table 1. Publication cells selected for evaluation.

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

21. The enterprises in each publication cell are categorised by size class in three edit cells: size class 1 (less than 10 employees), size class 2 (from 10 to 100 employees), and size class 3 (at least 100 employees). An edit cell is an intersection of publication cell and size class, see Appendix A. 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 they should have high quality data. There are therefore 18 edit cells in our evaluation where records are edited selectively.

B. Available data

23. We have raw data without obvious mistakes, selectively (either manually or automatically) edited data, and correct data for ASBS Retail trade and Transport 2000. We also have reference values from ASBS 1999 and 2000 and other information about companies, such as inclusion probabilities, final weights, SIC, and size of the company.

24. We use improved reference values compared to those used in production. There were a lot of problems with the reference values for ASBS Trade and Transport 2000 (Hoogland and Van der Pijll, 2003). These reference values were based on data of ASBS 1999. These data had to be transformed to the 2000 format, because of new questionnaires for 2000. However, a one-to-one transformation did not exist for some variables. Furthermore, some transformation errors were made. We want to evaluate score functions under conditions where reference values are both of acceptable quality and realistic. In this paper the reference values are therefore based on selectively edited data from ASBS 2000.

C. Score functions

25. We assess several types of score functions. They are characterised as either *block score functions* or *ratio score functions*. Furthermore, block score functions monitor different sets of variables for ASBS 2000 Retail trade and Transport.

Block score functions

26. A questionnaire for an annual structural business statistic at Statistics Netherlands has four important blocks of variables, 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 we want to calculate several score functions and assess which score function performs best at error detection.

27. The reference values for block score functions are estimated population medians and totals for a subset of variables in a block. These values are estimated on the basis of selectively edited records of ASBS 2000 for the same edit or median cell. A median cell is an intersection of a publication cell and company size, and therefore a subdivision of an edit cell, see Appendix A.

28. A record will obtain a high value for a block score function, when values of variables in a block differ much from the corresponding weighted medians. These weighted medians serve as estimates of the population medians. Furthermore, a block score function 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.

29. We investigate three variants of block score function $SF_i(p, k_g)$ in Appendix B.

1. $SF_i(\mathbf{p}, l_d)$, used for ASBS 2000, with weighted medians per edit cell d .
2. $SF_i(\mathbf{p}, m_c)$, used for ASBS 2001-2002, with weighted medians per median cell c .
3. $SF_i(\mathbf{p}^2, m_c)$, used for ASBS 2003, with a squared inclusion probability in the denominator and weighted medians per median cell c . Small companies have a relatively large influence for this score function, because small companies have a relatively small inclusion probability.

Example

30. We give an example of the differences in the effect of inclusion probability and weighted medians for the above block score functions. Table 2 shows that the average effect of inclusion probability on $SF_i(p, k_g)$ decreases when company size increases. Furthermore, weighted medians in a median cell can easily be twice as small or large than weighted medians in an edit cell.

Table 2. Weighted median of net turnover and average value of the effect of inclusion probability on score function $SF_i(p, k_g)$ for publication cell 152110 and company size 0-6.

Company size	Average value of		Weighted median of Net turnover (in millions of euros)	
	$1 / \sqrt{\mathbf{p}_i}$	$1 / \mathbf{p}_i$	Median cell	Edit cell
0	4.5	21.7	0.1	0.3
1	4.2	17.9	0.2	0.3
2	3.1	10.9	0.4	0.3
3	2.5	9.5	0.8	0.3
4	2.0	6.0	2.5	4.2
5	1.8	3.1	5.3	4.2
6	1.0	1.0	9.7	4.2

31. The parameters of a score function can have a substantial effect on the resulting value of this score function. From now on we refer to this value as the *score*. The importance weights \mathbf{a}_{dj} for variables y_j in edit cell d vary from 0 to 100. The empty entry factors e_{dj} have values that range from 0 to 20. An empty value for a variable with an empty entry factor of 20 and an importance weight of 100 will often result in an insufficient score for $SF_i(p, k_g)$ for that record. In table 3 we give the input for score function $SF_i(\mathbf{p}^2, m_c)$ for a record that has a sufficient score for publication cell 152110. In the business profit block of a small questionnaire for Trade and Transport there are five entries: net turnover (n.t.), net principal turnover (n.p.t.), net sideline turnover (n.s.t.), remaining operating profit (r.o.p.), and total operating profit (t.o.p.). Table 3 shows that the contribution of a variable to a score can differ substantially across variables.

Ratio score functions

32. Ratios can be a powerful tool for error detection, because they tend to vary less than the underlying variables. For example, the net income per employed person will vary less than the net income for a branch of industry.

Table 3. A record that has a sufficient score for $SF_i(p^2, m_c)$ and the operating profit block.

Variable	Parameters		Value	Reference values		Contribution to $SF_i(p^2, m_c)$
	Importance weight	Empty field factor		Weighted median (in thousands of euros)	Weighted total (in millions of euros)	
n.t.	100	20	510	232	3691	1.02
n.p.t.	100	20	510	228	3492	1.17
n.s.t.	1	0	0	12	199	0.01
r.o.p.	1	0	20	3	34	0.43
t.o.p.	100	20	530	232	3725	1.00

33. We will evaluate two score functions that use ratios based on raw values without obvious mistakes (w.o.m.). Score function (8) in appendix B used for ASBS 2000-2002 and score function (9) used for ASBS 2003. Weighted medians of ratios based on selectively edited values are used as reference values. We discuss the reasons for using (9) instead of (8) in Appendix C.

D. Calibration

34. We need to calibrate score functions to determine the cut-off point for manual editing. Normally, edited data for last year is used, but we have transformation problems for ASBS 1999. We will therefore use edited values of variables for ASBS 2000 Retail trade and Transport. It suffices to set the sufficiency limit (Hoogland, 2002; Hoogland, 2005). This is done by computing the 90% quantile of the empirical cumulative distribution function of a score function with edited data as input.

E. Evaluation criteria

35. ASBS Retail trade and Transport contain a lot of important variables. We compute several evaluation criteria for each score function, important variable and edit cell for the nine publication cells in table 1. We use the fact that we have correct values of variables for each record.

36. First, for each score function and edit cell we compute the percentage of raw w.o.m. records that have a sufficient score. We then look at the contribution of raw w.o.m. records that have a sufficient score to the weighted total of variable y_j for the edit cell. We compute this contribution as follows

$$B_{B_{bf}}^{dj} = \frac{\sum_{i: SF_{bf}^i \leq h_{bf}^d} w_i \cdot y_{ij}^c}{\sum_i w_i \cdot y_{ij}^c}, \quad (1)$$

where

- w_i is the weight for record i ;
- y_{ij}^c is the correct value of variable y_j for record i ;
- B_{bf} is a block score function for block b of type $f=1,2,3$;
- $SF_{B_{bf}}^i$ is the score for B_{bf} for record i ;
- $h_{B_{bf}}^d$ is the sufficiency limit for B_{bf} in edit cell d .

$$B_{R_f}^{dj} = \frac{\sum_{i: SF_{R_f}^i \leq h_{R_f}^d} w_i \cdot y_{ij}^c}{\sum_i w_i \cdot y_{ij}^c}, \quad (2)$$

where

- $R_f, f=1,2$, is a ratio score function;
- $SF_{R_f}^i$ is the score for R_f in record i ;
- $h_{R_f}^d$ is the sufficiency limit for R_f in edit cell d .

37. We also compute evaluation criteria of Hoogland (2005). We discuss these criteria in detail now. We are mainly interested in evaluation criteria that contain the size of errors for variable y_j in records with a sufficient score. We can simultaneously take the number of records with a sufficient score into account. We first define the error in variable j for record i as

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

For score functions $B_{1f}, B_{2f}, B_{3f}, B_{4f}$ (for respectively Block A, B, C, and D) we then use the following evaluation criterion for variable y_j and edit cell d .

$$C_{B_{bf}}^{dj} = \frac{\sum_{i: SF_{B_{bf}}^i \leq h_{B_{bf}}^d} w_i \cdot |e_{ij}| / v_{B_{bf}}^d}{\sum_i w_i \cdot y_{ij}^c / r^d} , \quad (3)$$

where

- $v_{B_{bf}}^d$ is the number of records in edit cell d that has a sufficient score for B_{bf} ;
- r^d is the response in edit cell d .

38. For ratio score functions R_1 (8) and R_2 (9) we use a different criterion, which considers errors in ratios for records with a sufficient score for R_f . For each ratio r_j and edit cell d we compute

$$C_{R_f}^{dj} = \frac{1}{v_{R_f}^{dj}} \cdot \sum_{i: SF_{R_f}^i \leq h_{R_f}^d} \left(\left| \frac{r_{ij}^c}{r_{ij}^r} \right| + \left| \frac{r_{ij}^r}{r_{ij}^c} \right| - 2 \right) , \quad (4)$$

where

- r_{ij}^c is the value for ratio j in record i for manually edited data;
- r_{ij}^r is the value for variable j in record i for raw data without obvious mistakes;
- $v_{R_f}^{dj}$ is the number of records in edit cell d with a sufficient score for R_f where both r_{ij}^c and r_{ij}^r can be calculated.

39. For both $SF_{B_{bf}}$ and SF_{R_f} we also compute the *absolute pseudo bias* for each y_j and edit cell d

$$D_{B_{bf}}^{dj} = \left| \frac{\sum_{i: SF_{B_{bf}}^i \leq h_{B_{bf}}^d} w_i \cdot e_{ij}}{\sum_i w_i \cdot y_{ij}^c} \right| \quad \text{and} \quad (5)$$

$$D_{R_f}^{dj} = \left| \frac{\sum_{i: SF_{R_f}^i \leq h_{R_f}^d} w_i \cdot e_{ij}}{\sum_i w_i \cdot y_{ij}^c} \right| . \quad (6)$$

Evaluation criteria (5) and (6) assess the bias in a weighted total of an edit cell due to selective editing, where records with a sufficient score are not edited (except for obvious mistakes).

V. Results of evaluation of score functions

A. Percentage of sufficient scores

40. The percentage of raw w.o.m. records in an edit cell that have a sufficient score varies from 69% for score functions B_{12} and B_{13} to 92% for score functions B_{42} and B_{43} . This percentage is on average 86% for raw w.o.m. records. There is a small structural difference between score functions, because it is in general somewhat higher for block score function B_2 and indicator score functions R_1 and R_2 . However, there is no structural difference across edit cells. For edited records it is 90% by definition, due to the way we calibrate score functions.

B. Contribution of sufficient scores to a weighted total

41. The contribution (to a weighted total for an edit cell) of raw w.o.m. records that have a sufficient score varies across variables, score functions, and edit cells. For instance, for number of employees there is no clear difference in the contribution of sufficient scores for different score functions (see table 4). However, for net turnover $B_{B_{21}}$ is in general larger than $B_{B_{22}}$ and $B_{B_{23}}$. This means that $SF_i(\mathbf{p}, l_d)$ is less suitable for selecting influential records for manual editing than $SF_i(\mathbf{p}, m_c)$ and $SF_i(\mathbf{p}^2, m_c)$. The same holds for other variables that represent totals of financial variables, such as total operating profit and total operating costs.

Table 4. Contribution $B_{B_{ij}}$ for number of employees and net turnover. Contributions larger than 90% are printed in boldface and those smaller than 70% are underlined.

Publication cell	Size class	Number of employees			Net turnover				
		B_{11}	B_{12}	B_{13}	R_2	B_{21}	B_{22}	B_{23}	R_2
152110	1	82	80	77	83	87	79	75	83
	2	78	76	84	82	80	74	84	87
152121C	1	86	90	90	87	94	87	84	86
	2	75	74	74	78	78	80	78	78
152121E	1	<u>68</u>	<u>68</u>	<u>67</u>	87	<u>69</u>	<u>56</u>	<u>55</u>	86
	2	82	85	90	82	82	85	86	81
160220	1	91	77	80	82	95	89	83	82
	2	82	75	85	84	77	<u>65</u>	80	79
161100	1	78	80	74	91	78	<u>67</u>	73	84
	2	<u>60</u>	<u>59</u>	74	87	86	<u>56</u>	72	<u>68</u>
161200	1	72	85	78	79	74	87	77	89
	2	<u>43</u>	73	75	96	<u>54</u>	<u>46</u>	86	91
163110	1	77	98	83	93	88	79	72	96
	2	82	84	81	95	96	76	78	89
163300	1	89	80	85	75	96	<u>64</u>	76	73
	2	73	81	90	73	92	<u>64</u>	81	<u>60</u>
163400	1	78	91	85	77	90	85	82	<u>61</u>
	2	73	80	84	83	89	<u>60</u>	<u>70</u>	<u>65</u>

C. Size of errors in records with sufficient scores

42. There are large differences in the size of errors in records with sufficient scores $C_{B_{ij}}$ across variables and edit cells (see table 5 for net turnover and number of employees). However, there are hardly any structural differences in $C_{B_{ij}}$ across block score functions. Score function B_{b1} seems to be somewhat better in error detection than score functions B_{b2} and B_{b3} for a number of important variables in block B and C and companies of size class 1. We will not go into details here, but this difference is caused by an unexpected side effect of the computation of weighted medians in median cells that contain many zeros. Due to this side effect score function B_{b1} used for ASBS 2000 seems to be in favour compared to the ‘improved’ score functions used for ASBS 2001-2003. However, we claim that there will be no structural difference in the quality of error detection of block score functions when this side effect is solved.

Table 5. Size of undetected errors $C_{B_{ij}}$ and C_{R_j} for number of employees and net turnover. $C_{B_{ij}}$ larger than 0.1 and C_{R_j} larger than 0.2 are printed in boldface.

Publication cell	Size class	Number of employees			Net turnover			Net turnover / Number of employees	
		B_{11}	B_{12}	B_{13}	B_{21}	B_{22}	B_{23}	R_1	R_2
152110	1	0.10	0.16	0.18	0.02	0.02	0.02	0.03	0.03
	2	0.06	0.06	0.06	0.01	0.01	0.01	0.02	0.02
152121C	1	0.10	0.10	0.10	0.02	0.02	0.02	0.09	0.09
	2	0.09	0.05	0.05	0.00	0.00	0.00	0.00	0.00
152121E	1	0.10	0.08	0.08	0.02	0.02	0.02	0.05	0.05
	2	0.07	0.04	0.04	0.05	0.04	0.04	0.00	0.00
160220	1	0.05	0.06	0.06	0.00	0.00	0.00	0.01	0.01
	2	0.15	0.12	0.13	0.00	0.00	0.00	0.04	0.04
161100	1	0.09	0.09	0.09	0.01	0.01	0.01	0.04	0.04
	2	0.06	0.07	0.07	0.04	0.03	0.03	0.00	0.00
161200	1	0.16	0.13	0.12	0.01	0.01	0.01	0.09	0.09
	2	0.04	0.07	0.07	0.03	0.03	0.03	0.58	0.58
163110	1	0.06	0.06	0.06	0.00	0.00	0.00	0.08	0.08
	2	0.05	0.01	0.01	0.05	0.05	0.05	0.26	0.26
163300	1	0.05	0.05	0.05	0.93	0.76	0.92	1.17	1.15
	2	0.03	0.03	0.03	0.36	0.39	0.49	1.09	1.06
163400	1	0.03	0.03	0.03	0.01	0.01	0.01	0.09	0.15
	2	0.04	0.05	0.05	0.02	0.02	0.02	0.28	0.13

43. There are large differences in C_{R_j} across ratios and edit cells, but hardly any differences in C_{R_j} across ratio score functions (see table 5 for ratio net turnover / number of employees). For each ratio monitored by the two ratio score functions, error detection is better for Retail trade than for Transport. For Retail trade there are no large errors in any of the ratios, when a ratio score function has a sufficient score.

44. Table 6 shows that there can be large difference between $D_{B_{ij}}$ and D_{R_j} . For number of employees and net turnover the absolute pseudo bias is in general larger for a block score function than for a ratio score function. However the opposite holds for some other variables of ASBS that are monitored by both a block score function and a ratio score function. The accuracy of reference values, that is, weighted medians of block variables and ratios, seems to play a decisive role. In general, the absolute pseudo bias increases when the accuracy of reference values decreases. A median is inaccurate as a reference value when it is based on edited values with a large variation (Hoogland and Van der Pijll, 2003).

Table 6. Absolute pseudo bias $D_{B_{ij}}$ and D_{R_j} (in percentages) for number of employees and net turnover. $D_{B_{ij}}$ and D_{R_j} larger than 5% are printed in boldface.

Publication cell	Size class	Number of employees					Net turnover				
		B_{11}	B_{12}	B_{13}	R_1	R_2	B_{21}	B_{22}	B_{23}	R_1	R_2
152110	1	1.7	9.9	10.7	8.3	8.4	1.4	1.4	1.3	10.1	10.1
	2	4.7	4.7	4.6	11.0	11.0	0.5	0.5	0.5	6.8	6.8
152121C	1	6.6	7.6	7.6	8.4	8.4	1.9	1.9	1.9	0.5	0.5
	2	7.6	3.3	3.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0
152121E	1	0.5	1.3	1.1	1.1	1.1	1.6	1.6	1.6	0.3	0.3
	2	6.1	3.4	3.4	1.5	1.5	2.9	1.9	1.9	0.0	0.0
160220	1	1.6	1.6	1.6	6.9	6.9	0.0	0.0	0.0	2.8	2.8
	2	8.3	8.4	7.8	12.1	12.1	0.0	0.0	0.0	3.6	2.7
161100	1	5.6	5.6	5.6	1.0	1.0	0.6	0.6	0.6	128.7	128.7
	2	4.4	4.4	4.4	15.8	15.8	3.2	2.0	2.0	0.3	0.3
161200	1	1.5	2.0	1.5	12.5	12.5	0.5	0.4	0.5	0.1	0.1
	2	3.1	0.1	0.1	20.1	20.1	1.8	1.9	1.9	0.0	0.0
163110	1	2.8	2.8	2.8	2.6	2.6	0.0	0.4	0.4	0.0	0.0
	2	4.4	1.2	1.2	3.1	3.1	3.1	3.1	3.1	2.4	2.4
163300	1	4.2	4.2	4.2	6.1	6.1	67.1	51.9	67.2	49.6	49.3
	2	2.2	2.2	2.2	3.1	3.1	21.1	20.6	28.2	25.6	25.5
163400	1	2.0	1.3	1.3	1.7	0.4	0.4	0.4	0.5	0.4	0.5
	2	2.7	2.3	2.3	5.6	3.7	0.8	0.8	0.8	1.1	0.9

VI. CONCLUSIONS

45. The editing phase is crucial for Annual Structural Business Statistics of Statistics Netherlands, because raw data contain many influential errors. Furthermore, ASBS contain many important variables that are either published, or supplied to third parties.

46. ASBS are edited selectively since the statistical year 2000. Score functions are used to select records for manual editing of ASBS. The remaining records are edited automatically. Score functions that were used for ASBS 2000 were modified during the past few years. This paper aims to evaluate some of the score functions used so far. We distinguish three types of block score functions that monitor variables within a questionnaire block and two types of score functions that monitor key variables through ratios. Block score functions compare raw values of variables with estimated population medians and weighted totals for a subset of variables in a block. Ratio score functions compare raw ratios with estimated population medians of ratios.

47. For our evaluation of score functions we use raw and edited data for nine publication cells for ASBS Retail trade and Transport 2000. Each record has been edited manually for evaluation purposes. Score functions are first computed for edited data to determine a cut-off point for manual editing. Score functions are then computed for raw data without obvious mistakes. Finally, several evaluation criteria assess the effect of the selection made by score functions on a set of important variables.

48. For some important variables the contribution of raw w.o.m. records that have a sufficient score depends on the score function. However, the size of errors in records with sufficient scores hardly depends on the type of block score function or the type of ratio score function. The bias due to selective editing does depend on the choice for either a block score function or a ratio score function. The accuracy of reference values seems to play a decisive role.

References

- Béguin, C., and B. Hulliger, 2004, Multivariate outlier detection in incomplete survey data: the epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society A* 167, pp. 275-294.
- Bikker, R., 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., 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.
- Bikker, R., J. Daalmans, J. Hoogland, and A. de Jong, 2004a, *Automatic editing of production statistics Building, Trade and Transport: possible gain in efficiency and improvements (In Dutch)*. Internal paper BPA-no 1384-04-TMO, Statistics Netherlands, Voorburg.
- Bikker, R., J. Daalmans, J. Hoogland, P. Muyrers, and M. Seip, 2004b, *Automatic editing of production statistics Industry and Commercial Services: possible gain in efficiency and improvements (In Dutch)*. Internal paper BPA-no 2-04-TMO, Statistics Netherlands, Voorburg.
- Chambers, R., A. Hentges, and X. Zhao, 2004, Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society A* 167, pp. 323-339.
- Charlton, J., 2002, *First results from the Euredit project – Evaluating methods for data editing and imputation*. Paper for UNECE Work Session on Statistical Data Editing, Helsinki, Finland.
- Charlton, J., 2004, *Evaluating new methods for data editing and imputation – Results from the Euredit project*. Paper for UNECE Work Session on Statistical Data Editing, Madrid, Spain.
- Farwell, K., 2002, *Fundamentals and suggestions for ranking and scoring edited unit responses*. Internal paper, Australian Bureau of Statistics.
- Farwell, K., R. Poole, and S. Carlton, 2002, *A technical framework for input significance editing*. Paper for Dataclean 2002, Jyväskylä, Finland.
- Farwell, K., and M. Raine, 2000, *Some current approaches to editing in the ABS*. Paper for the International Conference of Economic Statisticians, Buffalo, U.S.A.
- 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.
- Hedlin, D., 2003, Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics* 19, pp. 177-199.
- Hidiroglou, M., and J.-M. Berthelot, 1986, Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology* 12, pp. 73-83.
- Hoogland, J., 2002, *Selective editing by means of Plausibility Indicators*. Paper for Dataclean 2002, Jyväskylä, Finland.
- Hoogland, J., 2005, Selective editing using Plausibility Indicators and SLICE. In: *Statistical Data Editing Volume No. 3, Quality*.
- Hoogland, J., and E. van der Pijll, 2003, *Evaluation of the plausibility indicator for production statistics 2000 Trade and Transport (In Dutch)*. Internal paper BPA-no 1971-03-TMO, Statistics Netherlands, Voorburg.

- Langen, S., 2002, *Selective editing with logistic regression (In Dutch)*. Research paper no. 0206, Statistics Netherlands, Voorburg.
- Latouche, M., and J.-M. Berthelot, 1992, Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics* 8, pp. 389-400.
- Lawrence, D., and C. McDavitt, 1994, Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics* 10, pp. 437-447.
- Lawrence, D., and R. McKenzie, 2000, The General Application of Significance Editing. *Journal of Official Statistics* 16, pp. 243-253.
- Madsen, B., and B. Larsen, 2000, *The uses of neural networks in data editing*. Paper for the International Conference of Economic Statisticians, Buffalo, U.S.A.
- Nordbotten, S., 1995, Editing Statistical Records by Neural Networks. *Journal of Official Statistics* 11, pp. 391-411.
- Pol, F. van de, 1997, Selective editing in The Netherlands annual construction survey. In: *Statistical Data Editing, Volume No. 2, Methods and Techniques*, pp. 8-12.
- Sanders, S., 2002, *Selective editing by means of classification and regression trees (In Dutch)*. Internal paper BPA-no 599-02-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 H. 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

company size class	company size	number of employees	SIC		
			publication cell 163400		publication cell 160220
			63401	63402	6022
1	0	0	edit cell		median cell
	1	1			
	2	2-4			
	3	5-9			
2	4	10-19	median cell		Edit cell
	5	20-49			
	6	50-99			
3	7	100-199			
	8	200-499			
	9	> 499			

Figure 1. Cells for selective editing of ASBS, which are combinations of SIC and company size.

Appendix B: Score functions

The score functions below are evaluated. They are computed for each record in each edit cell. Score function B_b for Block b , $b = A, B, C$, and D is computed as follows

$$SF_{B_b}^i(p, k_g) \text{ Block } b = \sqrt{\frac{\sum_{j=1}^{J_b} \mathbf{a}_j \cdot \left(\frac{y_{ij} - k_{gj}}{Y_{dj}} \right)^2}{p_i \cdot \sum_{j=1}^{J_b} \mathbf{a}_{dj} \cdot \left(\frac{k_{gj}}{Y_{dj}} \right)^2}}, \quad (7)$$

where

- J_b is the number of variables that is monitored for block b ;
- y_{ij} , $j = 1, 2, \dots, J_b$, are entries for business unit i and year t ;
- k_{gj} , $j = 1, 2, \dots, J_b$, are the corresponding weighted medians for year $t-1$ and cell g containing business unit i in year t ;
- Y_{dj} , $j = 1, 2, \dots, J_b$, are the corresponding weighted totals for edit cell d and year $t-1$;
- \mathbf{a}_{dj} , $j = 1, 2, \dots, J_b$, denotes the importance of variable y_j for edit cell d ;
- p_i is a function of the inclusion probability of business unit i in year t .

If an entry y_{ij} is equal to 0 or empty then $y_{ij} = e_{dj} \cdot k_{gj}$ is used instead, where e_{dj} is the empty entry factor of variable y_j in edit cell d . An empty entry factor is large when it is considered unlikely that the corresponding entry equals 0.

The score functions with ratios are computed as follows

$$SF_{R_t}^i = \frac{\sum_{j=1}^J \mathbf{a}_{dij} \cdot \left(\frac{1}{2} \cdot \left(\frac{r_{ij}}{v_{cj}} \right)^{2 \cdot s_{dj}} + \frac{1}{2} \cdot \left(\frac{v_{cj}}{r_{ij}} \right)^{2 \cdot s_{dj}} \right)^{\frac{1}{2 \cdot s_{dj}}}}{\sum_{j=1}^J \mathbf{a}_{dij}}, \quad (8)$$

$$SF_{R_t}^i = \frac{\sum_{j=1}^J \mathbf{a}_{dij} \cdot \left(\text{MAX} \left(\left| \frac{v_{cj}}{r_{ij}} \right|, \left| \frac{r_{ij}}{v_{cj}} \right| \right) \right)^{s_{dj}}}{\sum_{j=1}^J \mathbf{a}_{dij}}, \quad (9)$$

where

- r_{ij} , $j = 1, 2, \dots, J$, are ratios for business unit i and year t ;
- v_{cj} , $j = 1, 2, \dots, J$, are the corresponding population medians for year $t-1$ and median cell c containing business unit i in year t ;

- s_{dj} , $j = 1, 2, \dots, J$, are exponents for edit cell d ;
- \mathbf{a}_{dj} , $j = 1, 2, \dots, J$, denotes the importance of ratio r_j in edit cell d .

$\mathbf{a}_{dj} = 0$ if r_{ij} or v_{cj} is zero or empty, and $\mathbf{a}_{dj} = \mathbf{a}_{dj}$ otherwise.

Appendix C: Details about change in score functions for indicators

In the editing process for ASBS, the parameters s_{dj} for (8) equal one and s_{dj} for (9) equal two for each ratio r_j and edit cell d . However, varying s_{dj} across ratios seems to have theoretical advantages. There is a large difference in the dispersion of the various ratios based on edited data. A deviant value for one ratio is therefore more serious than a deviant value for another ratio. We want to level the influence of deviant values for a ratio to assure that relatively large errors are detected for each ratio.

Score functions (8) and (9) are of the form

$$\frac{\sum_{j=1}^J a_{dj} b_{dj}}{\sum_{j=1}^J a_{dj}}$$

To level the influence of ratios r_j for edit cell d we level the 95% percent quantile of the empirical cumulative distribution function of b_{dj} , $j=1,2,\dots,J$, for edited data.

It is found that score function (8) is unsuitable to level out the different ratios, because the limit

$$\lim_{s_{dj} \rightarrow \infty} b_{dj}^{R_s} = \text{MAX} \left(\left| \frac{r_{ij}}{v_{cj}} \right|, \left| \frac{v_{cj}}{r_{ij}} \right| \right) .$$

for one ratio r_j can still be smaller than $b_{dij}^{R_s}$ with $s_{dij} = 1$ for another ratio r_j .

This is the main reason for considering the alternative ratio score function (9). The parameters s_{dj} for (9) can be univocally determined by dictating that $Q_{.95}(b_{dj}^{R_s}) = 9$, for all $j=1,2,\dots,J$, where

$$b_{dj}^{R_s} = \left(\text{MAX} \left(\left| \frac{r_{ij}}{v_{cj}} \right|, \left| \frac{v_{cj}}{r_{ij}} \right| \right) \right)^{s_{dj}} .$$

Example

We use the 95% quantile of the empirical cumulative distribution of $b_j^{R_s}$ for edited data to determine the exponents s_j for ratios r_j in edit cell 163400 size class 1. The resulting exponents s_j for score function (9) are given in table 7.

Table 7. Exponents s_j , $j=1,2,\dots,7$, for score function (9) and edit cell 163400 size class 1.

j	$Q_{.95}(b_j^{R_s})$	s_j
1	18.5	0.8
2	1.0	45.0
3	43.2	0.6
4	2.9	2.1
5	3.1	2.0
6	3.5	1.8
7	1.6	4.9