

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

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (iii): Editing and imputation of administrative and census data

**DETECTION OF POTENTIAL INFLUENTIAL ERRORS IN VAT TURNOVER DATA
USED FOR SHORT-TERM STATISTICS**

Supporting Paper

Prepared by Jeffrey Hoogland, Koert van Bommel, and Peter-Paul de Wolf, Statistics Netherlands

Abstract: Until now, Statistics Netherlands has used questionnaires to produce short-term statistics on turnover. VAT data were used only as an auxiliary variable. To reduce respondent burden, we are trying to use VAT turnover data instead of questionnaires, for small and medium-sized companies. This is not straightforward, because it can be difficult to link fiscal units to enterprises. Another challenge is that data become available in waves, because companies can report either by month, by quarter, or by year. A VAT-based statistical process was developed and tested for retail trade. Score functions are used to detect potential influential errors at the most detailed publication level for each wave. These errors can be either VAT errors, population frame errors, or linking errors. Data characteristics and aggregates are examined and selective editing is integrated in a macro-editing approach. The effect of selective editing on the microdata and aggregates is illustrated.

I. INTRODUCTION

1. There is a lot of pressure from the Dutch Government and the business community in the Netherlands to reduce the response burden for companies, especially for short-term statistics (STS). Statistics Netherlands has therefore developed a statistical process for STS which uses VAT data for small and medium-sized companies and questionnaires for large enterprises. This required 5 months of preliminary examination, 16 months of research, development and programming, and 8 months of implementation. In the end, the new process was only implemented for retail trade. For this branch the expected reduction in the number of questionnaires was relatively high, namely about 100,000 questionnaires on a yearly basis. Until now, the new process is not used to produce short-term statistics, because of a political decision.

2. In this paper we focus on the use of score functions to detect influential suspicious values for VAT turnover of enterprises. In section II, we discuss the use of VAT data for STS and the wave model. Section III deals with score functions and possible causes for errors. In section IV, we discuss a macro-editing approach and give examples of the effect of selective editing. Recent developments are discussed in section V.

II. SHORT TERM-STATISTICS BASED ON VAT-DATA

3. VAT data are very useful to produce statistics about turnover. However, using Dutch tax data is not a straightforward process. Tax data are only available for fiscal units and several matching procedures are necessary to obtain tax data for enterprises. Furthermore, tax assessments are not used. For business statistics they are not available in time and do not contain all necessary variables to derive VAT turnover. Therefore we use tax declarations, which have to be edited because of measurement, population frame, and linking errors.

4. VAT declarations may be submitted on a monthly, quarterly, or yearly basis. Enterprises with turnover above a certain threshold are obliged to declare every month, enterprises with only little turnover may declare on a yearly basis. VAT turnover for a specific fiscal unit and time period is derived from variables in the tax declaration. Figure 1 gives the publication schedule for retail trade, which depends on publication demands and the availability of VAT data in a certain wave. For each month there are four waves, which result in a first, second, third, and final estimate. For the last month in a quarter the final estimate becomes available relatively quickly.

Publication	Number of days after end of first month of quarter						
	± 30	± 45	± 60	± 75	± 90	± 105	± 120
Month t	1 st estimate	2 nd estimate	3 rd estimate				final estimate
Month $t + 1$			1 st estimate	2 nd estimate	3 rd estimate		final estimate
Month $t + 2$					1 st estimate	2 nd estimate	final estimate
Quarter							1 st estimate

Figure 1: Wave model as a publication schedule for retail trade.

5. For short-term statistics we are interested in the yearly growth of turnover in a specific period, e.g. a month or a quarter. Different methods were developed for monthly and quarterly statistics to compute yearly growth for a branch. Weighting methods used for monthly and quarterly publications make use of a division of the population into strata. A stratum is a combination of NACE and company size class. We use VAT turnover for an enterprise if it is observed for each fiscal unit related to an enterprise. Furthermore, VAT turnover is not used if a fiscal unit is also related to another enterprise.

6. VAT turnover is used for small and medium-sized enterprises (SME), i.e. enterprises with fewer than 50 employed persons. Large enterprises (LE) are observed by a questionnaire. The current statistical process is used to estimate yearly growth for the branch as a whole, after turnover aggregates for SME have been imported.

7. To estimate yearly growth, turnover for an earlier year is used. This implies a starting problem. Different methods are used to estimate turnover aggregates and edit VAT turnover for a starting year. We detect influential suspicious growth rates and turnover values for consecutive periods within a starting year. This detection method and the estimation methods to obtain turnover aggregates are not discussed. For details, we refer to De Wolf and Van Bommel (2007) and De Wolf and Van Delden (2009) .

III. SCORE FUNCTIONS

A. Introduction

8. We want to detect records with an influential suspicious turnover or yearly growth rate, using the principles of selective editing, cf. Granquist and Kovar (1997). That is, we want to detect potential errors that influence the required output. To detect such errors the influence and the risk associated with a value for turnover in a publication cell is assessed. In this section we discuss score functions related to monthly

turnover. The same score functions can be used for quarterly and yearly turnover. The difference is that we have period $s = k$ and $s = k - 4$ for quarters, or period $s = j$ and $s = j - 1$ for years, instead of period $s = t$ and $s = t - 12$ for months.

B. Transformation of negative values

9. As a result of deductions, VAT turnover can have a negative value. This is a problem for the interpretation of growth rates, because these rates can become negative as well. Suppose, for instance, that an enterprise has a turnover of 100 k€ in month $t-12$ and -50 k€ in month t . The growth rate is then $100 / -50 = -2$. A negative growth rate is difficult to interpret and cannot be easily compared with positive growth rates. However, a negative growth rate can be very suspicious. We therefore transform negative rates to positive rates, using the following transformation (by stratum h):

$$\tilde{O}_{h,j}^s = \begin{cases} Q_{O_h^s}^{(2)} + \frac{Q_{O_h^s}^{(3)} - Q_{O_h^s}^{(2)}}{Q_{O_h^s}^{(2)} - Q_{O_h^s}^{(1)}} (Q_{O_h^s}^{(2)} - O_{h,j}^s) & \text{if } O_{h,j}^s < 0 \\ O_{h,j}^s & \text{if } O_{h,j}^s \geq 0 \end{cases}, \quad (1)$$

where $Q_{O_h^s}^{(p)}$ is the p -th quartile¹ of turnover $O_{h,j}^s$ in stratum h for period s . A turnover of zero is excluded for the computation of the quartiles. By means of formula (1) negative values are converted to positive values that have the same distance from the median, where we take possible skewness of the distribution into account. See figure 1 for an example.

10. A negative turnover can be a deduction of an earlier erroneous declaration or a correction of earlier estimated declarations. The correction or deletion of a negative value may be correct for a monthly estimation. However, it may be incorrect for a quarterly estimation if a negative value corrects for overestimated monthly turnover within the same quarter.

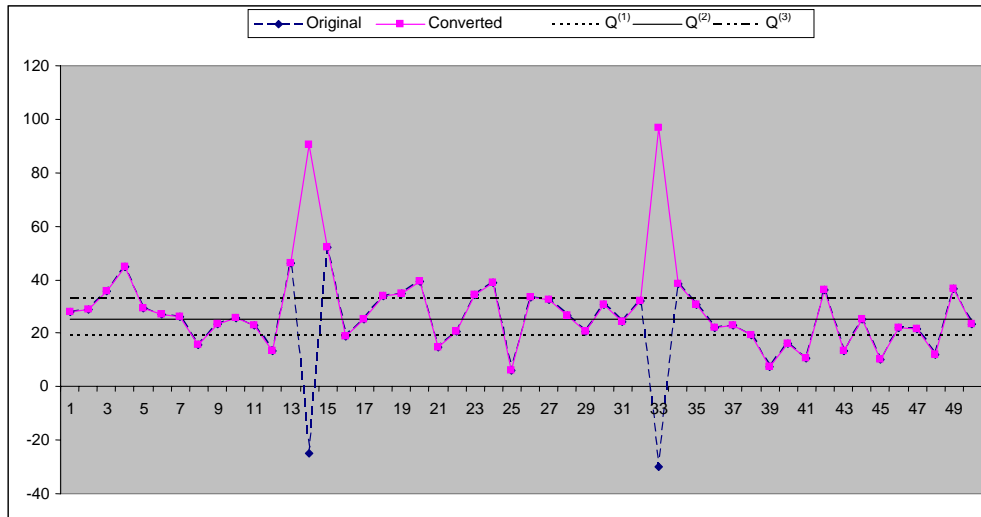


Figure 1: Conversion (1) applied to a set of 50 data points.

¹ The first quartile is the value for which 25% of the observations are lower, the second quartile is the value for which 50% of the observations are smaller (the median), and the third quartile is the value for which 75% of the observations are smaller.

C. Influence

11. Here we show how the influence of monthly turnover is compiled. We break down the influence of a value for turnover into three components. These components are

- a) the influence of small and medium-sized enterprises within the publication cell
- b) the influence of a turnover value within a stratum
- c) the influence of a stratum

12. The influence of small and medium-sized enterprises within the publication cell

The contribution of small and medium-sized enterprises to the total turnover for publication cell p can be formulated as

$$t_1^p = \frac{\hat{O}_{p,SME}}{\hat{O}_{p,SME} + \hat{O}_{p,LE}} = \frac{1}{1 + \hat{R}^p} \quad (2)$$

where $\hat{O}_{p,SME}$ and $\hat{O}_{p,LE}$ are estimates of total turnover in publication cell p , for SME and LE respectively

$$\hat{R}^p = \hat{O}_{p,LE} / \hat{O}_{p,SME} .$$

Ideally \hat{R}^p is determined on the basis of the period under review. Assuming that \hat{R}^p is fairly stable in time, it can also be approximated on the basis of an earlier period.

13. The influence of a turnover value within a stratum

A turnover value in period t or period $t-12$ may be erroneous. We do not want to underestimate influence of turnover due to an error. To assess influence of a turnover value in period t we therefore compute

$$\tilde{O}_j^t = \max\{\tilde{O}_j^t, \tilde{O}_j^{t-12} \hat{Q}_{G_h^{t-12}}^{(2)}\} ,$$

where $\hat{Q}_{G_h^{t-12}}^{(2)}$ is the median of the growth rates of enterprises in stratum h .

To assess influence of a turnover value in period $t-12$ we compute

$$\tilde{O}_j^{t-12} = \max\{\tilde{O}_j^{t-12}, \tilde{O}_j^t / \hat{Q}_{G_h^{t-12}}^{(2)}\}$$

The contribution of enterprise j to the total used VAT turnover in stratum h for publication cell p is

$$t_2^p(s, j) = \frac{\tilde{O}_j^s}{\sum_{j \in \eta_h^s} \tilde{O}_j^s} \quad (3)$$

where η_h^s is the set of indices of enterprises with VAT turnover in period s and stratum h ; and \tilde{O}_j^s the observed turnover after transformation (1) of enterprise j .

14. The influence of a stratum

A publication cell can exist of different strata. The contribution of stratum h to the total turnover is

$$t_3^p(s, j) = \frac{\hat{O}_h^s}{\sum_{k \in p} \hat{O}_k^s} \quad (4)$$

where \hat{O}_h^s is an estimate of the total turnover in stratum h (containing enterprise j) for period s . The summation in the denominator concerns all strata in publication cell p . For $s = t - 12$ the final estimated total turnover can be used:

$$\hat{O}_h^{t-12} = \hat{O}_{h,def}^{t-12}. \quad (5)$$

For $s = t$ we use a robust estimator that is not influenced by suspicious turnover values:

$$\hat{O}_h^t = \hat{Q}_{G_h^{t,t-12}}^{(2)} \hat{O}_{h,def}^{t-12} \frac{N_h^t}{N_h^{t-12}}, \quad (6)$$

with $\hat{O}_{h,def}^{t-12}$ the final estimated total turnover for stratum h in period $t - 12$, and N_h^s the population total for period s in stratum h .

15. The influence measures are combined to one measure for enterprise j in publication cell p :

$$I_{t,t-12}^p(j) = t_1^p \max\{t_2^p(t, j), t_3^p(t, j), t_2^p(t-12, j), t_3^p(t-12, j)\}, \quad (7)$$

The first term used to compute the maximum value gives the influence of enterprise j on the estimate for period t , while the second term gives the influence of the enterprise on the estimate for period $t - 12$.

D. Suspicious values

16. Our aim is to detect enterprises that have a suspicious turnover or growth rate.

For practical reasons we only consider enterprises with a large deviant turnover to determine which enterprises have a suspicious turnover. The reason is that a deviant small turnover will either have

- a) a small influence measure, because of a small turnover in t and $t-12$; In this case an enterprise is not detected anyway.
- b) a large influence measure, because of a large turnover in $t-12$. In this case an enterprise has a suspicious growth rate and will be detected.

17. For enterprises that are suspicious on the basis of turnover the following holds

$$\tilde{O}_{h,j}^s > \hat{Q}_{O_h^s}^{(3)} + C_1 (\hat{Q}_{O_h^s}^{(3)} - \hat{Q}_{O_h^s}^{(2)}), \quad (8)$$

where $\hat{Q}_{O_h^s}^{(p)}$ is the p -th stratum quartile based on turnover values in stratum h and period s , where zero values are excluded and $C_1 > 0$. In the editing phase for period t criterion (8) is applied for $s = t$ and $s = t - 12$, where parameter C_1 can be different in both periods.

18. The degree of suspiciousness of the turnover of an enterprise is given by

$$v_1^s = \begin{cases} 1 + \frac{\xi_1 (\tilde{O}_{h,j}^s - z_1)}{z_1} & \text{als } \tilde{O}_{h,j}^s > z_1 \\ 1 & \text{als } \tilde{O}_{h,j}^s \leq z_1 \end{cases} \quad (9)$$

where z_1 equals the right side of inequality (8) en $\xi_1 > 0$. This is applied for both $s = t$ and $s = t - 12$ and parameter ξ_1 can be chosen differently in both periods. This parameter can be used to scale the degree of suspiciousness. Note that for new enterprises there is no information for period $t - 12$ and for ceased enterprises there is no information for period t . In these cases the degree of suspiciousness equals 1 for period $t - 12$ and period t respectively.

19. We also assess suspiciousness for both the growth rate and the inverse of the growth rate. The growth rate of transformed turnover is given by

$$\tilde{G}_{h,j}^{t,t-12} = \frac{\tilde{O}_{h,j}^t}{\tilde{O}_{h,j}^{t-12}} \quad (10)$$

and the inverse growth rate by

$$\frac{1}{\tilde{G}_{h,j}^{t,t-12}} = \tilde{G}_{h,j}^{t-12,t} = \frac{\tilde{O}_{h,j}^{t-12}}{\tilde{O}_{h,j}^t}. \quad (11)$$

These growth rates can therefore both be written as $\tilde{G}_{h,j}^{s,u}$, where $s = t$ and $u = t - 12$, respectively $s = t - 12$ and $u = t$.

20. Enterprises that are suspicious on the basis of the (inverse) growth rate, are enterprises for which

$$\tilde{G}_{h,j}^{s,u} > \hat{Q}_{\tilde{G}_h^{s,u}}^{(3)} + C_2 \left(\hat{Q}_{\tilde{G}_h^{s,u}}^{(3)} - \hat{Q}_{\tilde{G}_h^{s,u}}^{(2)} \right) \quad (12)$$

where $\hat{Q}_{\tilde{G}_h^{s,u}}^{(p)}$ is the p -th stratum quartile based on the growth rates for period u to s in stratum h of transformed turnover (if $\tilde{O}_{h,j}^s \neq 0$ and $\tilde{O}_{h,j}^u \neq 0$) and $C_2 > 0$. Parameter C_2 can be chosen differently in both cases (growth rate and inverse growth rate).

21. The degree of suspiciousness of an enterprise based on the (inverse) growth rate, is given by

$$v_2^{s,u} = \begin{cases} 1 + \frac{\xi_2 (\tilde{G}_{h,j}^{s,u} - z_2)}{z_2} & \text{als } \tilde{G}_{h,j}^{s,u} > z_2 \\ 1 & \text{als } \tilde{G}_{h,j}^{s,u} \leq z_2 \end{cases} \quad (13)$$

where z_2 equals the right side of inequality (12) and $\xi_2 > 0$. For the growth rate and inverse growth rate the scale parameter ξ_2 can be chosen differently. For ceased and new enterprises $v_2^{s,u} = 1$.

22. Lastly, the suspiciousness measures are combined to one measure for enterprise j :

$$V_{t,t-12}^p(j) = \max\{v_1^t, v_1^{t-12}\} v_2^{t,t-12} v_2^{t-12,t} - 1. \quad (14)$$

The maximum of v_1^t and v_1^{t-12} is taken, because turnover that is suspicious in period t and period $t - 12$ is not considered more questionable than turnover that is only suspicious in period t or period $t - 12$. For $v_2^{t,t-12}$ and $v_2^{t-12,t}$ it is not necessary to compute a maximum because these measures cannot be larger than one at the same time.

23. various suspiciousness measures are either based on turnover level or on growth rate. The length of the right tail of the distribution of these measures may vary. This length must be comparable across measures, because values are considered suspicious when they are in the right tail of a measure. We have to prevent one measure dominating the detection of suspicious values. The scale parameters ξ_1 and ξ_2 are therefore chosen in such a way that these right tails are comparable.

D. Risk indicator

24. To arrange enterprises according to editing priority each enterprise is assigned a Potential Influential Error (PIE) score. This score indicates the risk for an incorrect yearly growth, in a publication cell and period, if the turnover for that enterprise is left unchanged. The PIE score is given by:

$$P_{t,j-12}^p(j) = V_{t,j-12}^p(j) I_{t,j-12}^p(j) \quad (15)$$

The threshold is set at 0.01 and enterprises with the highest PIE scores are edited first. This process consists of checking whether the NACE, size class and VAT turnover are correct.

25. Some extreme turnover values in retail trade are caused by frame errors. Quite often the enterprise is actually a wholesaler, and has a much higher turnover than a retailer with the same number of employed persons. To check the NACE, an internet search is done for the company to find out its main activity. Whether turnover is considered to be correct depends on the turnover for other periods, the seasonal effect expected for the branch, and possible differences between VAT turnover and 'statistical' turnover.

26. There are several causes for errors:

- a) an enterprise may be wrongly classified in the population frame;
- b) an enterprise may be wrongly matched to VAT data;
- c) a fiscal unit may make a typing error while filling in the VAT declaration form;
- d) a fiscal unit may only declare VAT in the last month of a quarter, or last quarter of a year;
- e) a fiscal unit may have a 4-week administration and declares VAT for the first 4-week period for January, VAT for the second 4-week period for February etc, and the sum of VAT for the last two 4-week periods for December;
- f) a fiscal unit may declare VAT turnover which includes a correction related to an earlier period;
- g) a fiscal unit may not declare turnover which is relevant for Statistics Netherlands;
- h) a fiscal unit may declare turnover which is relevant for the Dutch tax authority, but not relevant for Statistics Netherlands.

IV. MACRO-EDITING

27. The aim is to use score functions for VAT data as part of a macro-editing approach. For a publication cell the following data characteristics are examined:

- a) Aggregates and yearly growth for all enterprises, for SME and LE separately, and for each size class;
- b) Population dynamics. E.g. if enterprises change from SME to LE, this may have an effect on yearly growth of turnover of SMEs;
- c) Distributional characteristics, such as minimum, maximum, and quartiles. Graphical methods, such as histograms, scatter plots, and box plots can be useful. The aim is to detect outliers, which can represent frame errors, match errors or VAT errors. These outliers are not edited directly, but they serve as a check for the PIE score. For instance, it is useful to examine a scatter plot with turnover for month t versus turnover for month $t-12$, where enterprises with a PIE score above the threshold have a different colour or sign.

28. Examining data characteristics gives an impression of the quality of the data. The next step is to edit large enterprises and small and medium-sized enterprises with a deviant PIE score. If the NACE or turnover of an enterprise is found to be incorrect, the turnover value is deleted. In the case of an incorrect size class the turnover value is considered correct. Step a) can then be repeated to assess the impact of selective editing on the output. If there is time and manpower left, more records with a deviant PIE score

should be edited. In practice, there may not be enough time or manpower. In this case we should have an idea of the potential impact of remaining potential influential errors.

29. Table 1 gives the number of enterprises with VAT turnover, the number of potential influential errors, and the number of VAT errors and NACE errors for each period in the first quarter of 2009. VAT turnover has already been edited for 2008 (the starting year). We do not show results for wave 1, as only a few VAT turnover values are indicated as a PIE. For waves 2 and 3, too, only a small part of the VAT turnover is indicated as a potential influential error. For January 2009, 27.5% of the potential influential errors in wave 2 are assessed to be VAT errors or NACE errors. Turnover values found to be incorrect are deleted. As a result, the estimated total turnover for retail trade in January 2009 decreased by 5.5%. For other periods and waves these percentages are smaller. Editors were less motivated to check records with potential influential errors, because it was decided that the new statistical process was not going to be used to produce monthly statistics for retail turnover. The reason was that the Ministry of Finance decided that from July 2009, all companies do not have to submit a VAT declaration every month.

Table 1. Number of enterprises with VAT turnover, potential influential errors, VAT errors, and NACE errors, for each period in the first quarter of 2009.

period	wave 2				wave 3			
	VAT	PIE	VAT error	NACE error	VAT	PIE	VAT error	NACE error
January 2009	17022	222	25	36	18063	71	2	7
February 2009	17138	116	9	7	17884	76	1	6
March 2009	17428	73	3	3	17694	99	4	4
1st quarter 2009					68868	75	6	6

V. RECENT DEVELOPMENTS

30. The statistical process and software was almost ready for production in May 2009. Just before the new process was introduced and the distribution of STS questionnaires was stopped for small and medium-sized enterprises, some bad news arrived. To help enterprises to survive the financial crisis and to reduce the administrative burden, the Ministry of Finance decided that all enterprises that submit a monthly VAT declaration are permitted to submit quarterly declarations. If most enterprises make use of this possibility, this will make it impossible to produce reliable monthly statistics based on VAT data. It was therefore decided to mothball the new statistical system.

31. In 2008 another project was started to increase the use of VAT turnover for STS; in the framework of this project we can:

- a) produce STS for all branches, not just retail trade;
- b) provide consistent estimates for yearly growth and total turnover per month, quarter, year;
- c) use VAT turnover instead of questionnaires for all enterprises except the 1,900 largest ones;
- d) change the composition of enterprises in such a way that they can be more easily linked to fiscal units;
- e) automatically correct systematic errors in VAT turnover
- f) deal with differences in definition between VAT turnover and our definition of turnover;
- g) macro-editing, including interactive correction of influential errors;
- h) impute missing VAT turnover values;

32. The emphasis of this follow-up project is primarily on obtaining reliable turnover statistics. Parts of the mothballed methods will be used, including the methods for detection of suspicious influential VAT turnover. In autumn 2009 we shall examine how many monthly declarations remain and how many monthly questionnaires are needed. This number may be much higher than before, because monthly VAT turnover is also used as an auxiliary variable to decrease sample size.

References

- 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.
- De Wolf, P.-P., and K. van Bommel, 2007, *Method and process description STS and secondary sources*. Internal paper Statistics Netherlands, The Hague.
- De Wolf, P.-P., and A. van Delden, 2009, *Method description STS and secondary sources version 2*. Internal paper Statistics Netherlands, The Hague.