Seminar on 'Using Administrative data in the Production of Business Statistics – Member States experiences'

Italy, Rome, 18-19 March 2010

Session 1: Methods for editing and controlling quality of Admin Data

EDITING STRATEGIES FOR VAT DATA

Prepared by Jeffrey Hoogland, Statistics Netherlands

Abstract: Statistics Netherlands uses questionnaires to produce short-term statistics on turnover. VAT data are mainly used as an auxiliary variable. To reduce respondent burden, we are trying to use VAT turnover data instead of questionnaires. This is not straightforward, because it can be difficult to link VAT units to enterprises. A statistical process is developed, using both VAT data and questionnaires. A top-down editing approach is followed. Turnover aggregates are examined first and score functions are used to detect potential influential errors at the publication level. These errors are caused by measurement errors, population frame errors, or linking errors. The number of potential influential errors is relatively small and they are edited manually.

1. INTRODUCTION

There is a lot of pressure from the Dutch government and business community to reduce the response burden for companies, for both short-term statistics (STS) and structural business statistics (SBS). Statistics Netherlands is therefore developing a statistical process which uses VAT data for most companies and questionnaires for the 1900 largest groups of enterprises. The aim is to produce both reliable yearly turnover growth rate and reliable quarterly turnover level. Estimated turnover aggregates for four quarters added up and used for SBS.

The resulting mixed mode statistical process consists of nine parts. Parts 1, 2, 3, and 5 constitute the input data phase. Parts 4, 6, 7, 8, and 9 constitute the combined data phase.

1. Preparation and receipt of input data
2. Technical standardization of input data
3. Editing of generic systematic errors in input data
4. Linking VAT units to statistical units
5. Editing of branch specific systematic errors in input data
6. Estimation of totals
7. Top-down editing of totals
8. Authorization of publication totals
9. Publication of figures

In this paper we focus on editing strategies for VAT data. That is, we focus on parts 3, 5 and 7 for VAT data. In section 2, we discuss the use of VAT data for STS. Section 3 concerns systematic errors in input data. Section 4 deals with score functions to detect influential errors for statistical units. In section 5, we discuss top-down editing. Conclusions are given in section 6.

2. USE OF VAT-DATA FOR SHORT TERM-STATISTICS

VAT data are collected by the tax authorities. They may be very useful to produce turnover statistics. However, using Dutch tax data is not a straightforward process. Tax data are only available for VAT units and several matching procedures are necessary to obtain tax data for enterprises. Furthermore, tax data for enterprises have to be edited because of measurement, population frame, and linking errors. VAT declarations become available in waves, because they may be submitted on a monthly, quarterly, or yearly basis. Only enterprises with little turnover may declare on a yearly basis. A substantial part of the VAT turnover is therefore available soon. VAT turnover for a specific VAT unit and time period is derived by Statistics Netherlands from variables in the VAT declaration.

For short-term statistics we are interested in the yearly growth of turnover in a specific period, e.g. a month or a quarter. In this paper we focus on quarterly statistics. We make use of a division of the population into strata. A stratum is a combination of NACE and company size classes. So far, we use VAT turnover for an enterprise if it is observed for at least one related VAT unit. Otherwise, the turnover for an enterprise is imputed. Methods are developed to impute VAT turnover for missing VAT units. We take into account that a VAT unit can be exempted from taxation.

VAT turnover and other tax information can not be used if a VAT unit is related to several enterprises. Respondent burden can be reduced if administrative data can be used more often. SN therefore changed the composition of enterprises. If a VAT unit is related to several enterprises they are combined. A disadvantage is that enterprises may become more diverse in their economical activities. It can therefore be more difficult for an enterprise to fill in a questionnaire for a business statistic.

For some branches it is expected that VAT can not be used, because of VAT regulations. For each NACE the usability of VAT data is therefore assessed. For past years the yearly VAT turnover of an enterprise is compared with yearly turnover that is observed by SBS. A decision tree is used to decide whether VAT turnover is used for a specific branch. For instance, VAT turnover is not used when a regression analysis shows there is no linear relationship between VAT turnover and SBS turnover, and VAT regulations apply. For example, mortuaries have a VAT dispensation. However, in some cases mortuaries do declare VAT, because of sidelines such as photo reports. There is no linear relationship between sidelines of mortuaries and principal turnover. VAT turnover of mortuaries is therefore not used to estimate net turnover.

From the year 2011 questionnaires are sent to enterprises where VAT turnover is not usable or enterprises that belong to the 1900 largest groups of enterprises. When enterprises do not respond in time, VAT data may also be used to impute missing data for large groups of enterprises. Until then filled-in questionnaires for short-term statistics based on the current statistical process are used.

2. DETECTION OF SYSTEMATIC ERRORS

For Statistics Netherlands (SN) it is important that a VAT declaration for a specific period relates to that period. For the Dutch tax authorities it is less important. For instance, it happens quite often that a VAT unit states that it has zero turnover for the first three quarters and a positive turnover for the last quarter. It can be realistic that a unit was not active for three quarters. However, it is also possible that such a VAT unit actually had a positive turnover in several quarters. For the sake of convenience the VAT unit may declare its yearly turnover in the last quarter. In this case the yearly turnover is correct, but the turnover in the first three quarters is underestimated.

We apply detection rules to detect VAT units with suspicious monthly and quarterly turnover patterns (Ouwehand, 2010). Some of the quarterly patterns are only corrected when they occur for two years in a row. For instance, a VAT unit is detected when the turnover pattern for eight following quarters is {(0, 0, 0, x), (0, 0, 0, y), where $x>0$ and $y>0$}. A correction rule is then used such that quarterly turnover is not

available and only the yearly turnover can be used. As soon as a corrected VAT unit does not follow this pattern anymore, quarterly turnover is used again. Table 1 shows that suspicious quarterly turnover patterns occur frequently in VAT data, especially patterns with zero values or the same turnover values. For patterns 1 till 8 a correction rule is applied only if it occurs two years in a row for a VAT unit. For most patterns the correction rule implies that only the yearly turnover is used.

Table 1. Suspicious turnover patterns in quarterly VAT declarations of a VAT unit.

		Year		
No	Pattern	2007	2008	2009
1	(0,0,0,x) , x > 0	1,59%	1,53%	1,58%
2	(0,0,x,0) , x > 0	0,42%	0,45%	0,45%
3	(0,x,0,0) , x > 0	0,52%	0,52%	0,55%
4	(x,0,0,0) , x > 0	0,99%	0,97%	0,97%
5	(w,x,y,0), w,x,y>0	1,27%	1,27%	1,26%
6	(w,x,0,z), w,x,z>0	1,05%	1,04%	1,13%
7	(w,0,y,z), w,y,z>0	0,81%	0,79%	0,85%
8	(0,x,y,z), x,y,z>0	1,99%	1,99%	1,97%
9	(x,x,x,x), x ≠ 0	2,46%	1,89%	2,28%
10	(x,x,x,y), x ≠ 0 and y ≠ 0	0,98%	0,73%	0,84%
11	(x,x,y,x), x ≠ 0 and y ≠ 0	0,16%	0,12%	0,15%
12	(x,y,x,x), x ≠ 0 and y ≠ 0	0,17%	0,13%	0,16%
13	(y,x,x,x), x ≠ 0 and y ≠ 0	0,51%	0,42%	0,48%
14	(w,x,y,z), z<0	0,17%	0,17%	0,17%
15	(w,x,y,z), y<0	0,08%	0,08%	0,08%
16	(w,x,y,z), x<0	0,05%	0,06%	0,06%
17	(w,x,y,z), w<0	0,04%	0,04%	0,05%
	Other	86,74%	87,81%	86,96%

Some turnover patterns are only suspicious for specific branches. For instance, a VAT unit with the same turnover for each month or quarter is not considered suspicious when it leases out real estate. In the input data phase the only information about the economic activity of a company is a ‘branch code’ that the tax authorities assign to each VAT unit. This branch code is considered unreliable by SN. Branch specific correction rules are therefore not applied in the input data phase.

In the combined data phase SN links (groups of) enterprises to legal units, and legal units to VAT units. Legal units are constructed by means of information of the Chamber of Commerce. Each legal unit is assigned a NACE and a VAT unit obtains the NACE of the legal unit it is linked to. Branch specific correction rules can then be applied for each VAT unit.

3. DETECTION OF INFLUENTIAL ERRORS

A. Introduction

In the combined data phase several data sources are linked to assess publication figures and underlying strata, and detect influential errors. We want to detect enterprises 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 suspiciousness associated with a value for net turnover in a publication cell is assessed. In this section we discuss score functions related to quarterly turnover.

VAT turnover can have a negative value. A negative turnover can be a deduction of an earlier erroneous declaration or a correction of earlier estimated declarations. 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 quarter *t-4* and -50 k€ in quarter *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 turnover values to positive turnover values (Hoogland, 2009).

B. Influence

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

a) the influence of a turnover value within a stratum
b) the influence of a stratum

The influence of a turnover value within a stratum
A turnover value in period *t* or period *t-4* 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

$$\breve{O}_j^t = \max\{\tilde{O}_j^t, \tilde{O}_j^{t-4}\hat{Q}^{(2)}_{G_h^{t,t-4}}\} \ ,$$

where $\hat{Q}^{(2)}_{G_h^{t,t-4}}$ is the median of the growth rates of observed enterprises in stratum *h*, and $\tilde{O}_j^s$ the observed turnover (after transformation of negative values) for enterprise *j*.
To assess influence of a turnover value in period *t-4* we compute

$$\breve{O}_j^{t-4} = \max\{\tilde{O}_j^{t-12}, \tilde{O}_j^t \big/ \hat{Q}^{(2)}_{G_h^{t,t-4}}\}$$

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

$$\iota_1^p(s,j) = \frac{\breve{O}_j^s}{\sum_{j\in\eta_h^s}\tilde{O}_j^s} \ , \qquad (1)$$

where η_h^s is the set of indices of enterprises with VAT turnover in period *s* and stratum *h*.

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

$$\iota_2^p(s,j) = \frac{\hat{O}_h^s}{\sum_{k\in p}\hat{O}_k^s} \ , \qquad (2)$$

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 $s = t - 4$ the estimated total turnover is used:

$$\hat{O}_h^{t-4} = \hat{O}_{h,def}^{t-4} \ . \qquad (3)$$

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

$$\hat{O}_h^t = \hat{Q}^{(2)}_{G_h^{t,t-4}}\hat{O}_{h,def}^{t-4}\frac{N_h^t}{N_h^{t-4}} \ , \qquad (4)$$

with $\hat{O}_{h,def}^{t-4}$ the estimated total turnover for stratum *h* in period *t – 4*, and N_h^s the population total for period *s* in stratum *h*.

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

$$\mathrm{I}_{t,t-12}^{p}(j) = \max\left\{\imath_1^p(t,j)\imath_2^p(t,j), \imath_1^p(t-4,j)\imath_2^p(t-4,j)\right\} \ , \qquad (5)$$

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* – 4.

C. Suspicious values

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*-4; In this case an enterprise is not detected anyway.
b) a large influence measure, because of a large turnover in *t*-4. In this case an enterprise has a suspicious growth rate and will be detected.

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

$$\tilde{O}_{h,j}^{s} > \hat{Q}_{O_h^s}^{(3)} + C_1\left(\hat{Q}_{O_h^s}^{(3)} - \hat{Q}_{O_h^s}^{(2)}\right) \ , \qquad (6)$$

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 (6) is applied for $s = t$ and $s = t - 4$, where parameter C_1 can be different in both periods. The default value for C_1 is 4.

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

$$v_1^s = \begin{cases} 1 + \dfrac{\xi_1\left(\tilde{O}_{h,j}^{s} - z_1\right)}{z_1} & \text{als } \tilde{O}_{h,j}^{s} > z_1 \\ 1 & \text{als } \tilde{O}_{h,j}^{s} \le z_1 \end{cases} \qquad (7)$$

where z_1 equals the right side of inequality (6) en $\xi_1 > 0$. This is applied for both $s = t$ and $s = t - 4$ 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 $t - 4$ and for ceased enterprises there is no information for period *t.* In these cases the degree of suspiciousness equals 1 for period $t - 4$ and period *t* respectively.

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

$$\tilde{G}_{h,j}^{t,t-4} = \frac{\tilde{O}_{h,j}^{t}}{\tilde{O}_{h,j}^{t-4}} \qquad (8)$$

and the inverse growth rate by

$$\frac{1}{\tilde{G}_{h,j}^{t,t-4}} = \tilde{G}_{h,j}^{t-4,t} = \frac{\tilde{O}_{h,j}^{t-4}}{\tilde{O}_{h,j}^{t}} . \qquad (9)$$

These growth rates can therefore both be written as $\tilde{G}_{h,j}^{s,u}$, where $s = t$ and $u = t - 4$, respectively $s = t - 4$ and $u = t$.

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

$$\tilde{G}_{h,j}^{s,u} > \hat{Q}_{\tilde{G}_h^{s,u}}^{(3)} + C_2\left(\hat{Q}_{\tilde{G}_h^{s,u}}^{(3)} - \hat{Q}_{\tilde{G}_h^{s,u}}^{(2)}\right) \qquad (10)$$

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). The default value for C_2 is 4.

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

$$v_2^{s,u} = \begin{cases} 1 + \dfrac{\xi_2\left(\tilde{G}_{h,j}^{s,u} - z_2\right)}{z_2} & \text{als } \tilde{G}_{h,j}^{s,u} > z_2 \\ 1 & \text{als } \tilde{G}_{h,j}^{s,u} \leq z_2 \end{cases} \qquad (11)$$

where z_2 equals the right side of inequality (10) and $\xi_2 > 0$. For the growth rate and inverse growth rate the scale parameter ξ_2 can be chosen differently. For new and disappeared enterprises in the publication cell $v_2^{s,u} = 1$.

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

$$V_{t,t-4}^{p}(j) = \max\{v_1^t, v_1^{t-4}\} v_2^{t,t-4} v_2^{t-4,t} - 1. \qquad (12)$$

The maximum of v_1^t and v_1^{t-4} is taken, because turnover that is suspicious in period t and period $t - 4$ is not considered more questionable than turnover that is only suspicious in period t or period $t - 4$. For $v_2^{t,t-4}$ and $v_2^{t-4,t}$ it is not necessary to compute a maximum because these measures cannot be larger than one at the same time.

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. PIE score

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:

$$P_{t,t-4}^{p}(j) = V_{t,t-4}^{p}(j) \mathrm{I}_{t,t-4}^{p}(j) \qquad (13)$$

The threshold for (13) is set at 0.01 and the threshold for (12) is set at 0.005. Enterprises with the highest PIE score on the so called PIE list are edited first. This process consists of checking whether the NACE, size class and VAT turnover are correct.

E. Causes for influential errors

Some extreme turnover values are caused by frame errors. For instance, an enterprise that is considered to be a retailer, but is actually a wholesaler. A wholesaler can have a 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, possible differences between VAT turnover and 'statistical' turnover, and values of related variables for

There are several causes for influential errors:

a) an enterprise is wrongly classified in the population frame;
b) an enterprise makes a typing error while filling in the questionnaire;
c) a VAT unit is wrongly matched to an enterprise;
d) a VAT unit makes a typing error while filling in the VAT declaration form;
e) a VAT unit declares VAT turnover which includes a correction related to an earlier period;
f) a VAT unit does not declare turnover which is relevant for Statistics Netherlands;
g) a VAT unit declares turnover which is relevant for the Dutch tax authority, but is not relevant for Statistics Netherlands.

5. TOP-DOWN EDITING

At SN we want to increase the use, reliability, and reproducibility of top-down editing of business statistics. For a publication cell the following data characteristics are examined to decide whether turnover growth and turnover level is plausible:

a) Population dynamics within and between publication cells. E.g. if a large enterprise moves to a different publication cell, this may have an effect on yearly growth of net turnover;
b) Turnover aggregates and yearly growth for all enterprises, for the Top 1900 and the remainder separately, for each NACE, and for each size class;
c) Aggregates for wages, employees (fte), and working persons (fte) for the same groups of enterprises as for b). These aggregates are related to turnover and are estimated by means of edited pay roll declarations.
d) Distributional characteristics of turnover and ratios, such as minimum, maximum, and quartiles. E.g. a useful ratio is turnover divided by working persons (fte). Graphical methods, such as histograms, scatter plots, and box plots can be used. 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 list. For instance, it is useful to examine a scatter plot with turnover for quarter *t* versus turnover for quarter *t-4*, where enterprises on the PIE list have a different colour or sign;
e) Maximum value of PIE-score and number of enterprises on PIE list.

Examining data characteristics gives an impression of data quality. The next step is to edit enterprises with suspicious turnover. An editing tool is built showing relevant data for an enterprise, such as

- Series of quarterly turnover for the enterprise
- Series of wages, employees (fte), and working persons (fte) for the enterprise
- Name of enterprise
- Name and turnover of underlying VAT units

If turnover of an enterprise is found to be incorrect the turnover value is deleted and imputed. If the enterprise is classified in the wrong publication cell the NACE is adjusted. Step b) is repeated to asses the impact of selective editing on the output. If there is time and manpower left, more suspicious enterprises

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.

Table 2 gives the number of potential influential errors, for each quarter of 2009. Only a small part of the enterprises has as a potential influential error in VAT turnover. For wholesale trade the number of PIE is relatively high. About 21,6% of the PIE occur in wholesale trade, while only 9,5% of the enterprises with observed turnover is a wholesaler. For wholesale trade there are several VAT regulations that can result in a suspicious turnover. Furthermore, the relation between turnover and number of working persons is weak in this branch. The number of working persons is used to form strata.

Table 2. Number of enterprises based on observed VAT turnover, number of imputed enterprises, and number of potential influential errors in quarterly turnover, for each quarter of 2009.

	All branches			Retail trade			Wholesale trade		
Period	observed	imputed	PIE	observed	imputed	PIE	observed	imputed	PIE
1st quarter 2009	608.428	114.836	810	73.066	15.209	64	59.441	11.004	192
2nd quarter 2009	618.177	171.815	848	73.101	13.365	50	59.542	9.712	169
3rd quarter 2009	635.933	206.676	928	74.856	13.486	58	59.721	9.724	198
4th quarter 2009	631.463	211.463	896	74.252	14.035	52	58.782	10.013	192

V. CONCLUSIONS

Statistics Netherlands aims to increase the use of VAT turnover for STS and SBS. We therefore changed the composition of enterprises in such a way that they can be linked to VAT units more easily. Several editing strategies are implemented for VAT data to improve the quality of estimates for turnover aggregates and yearly growth of turnover. Systematic errors in VAT units are corrected automatically and top-down editing is used to detect influential errors in turnover for (groups of) enterprises.

In 2011 we want to

a) provide consistent estimates for yearly growth and total turnover per month, quarter, year;
b) use VAT turnover instead of questionnaires for all enterprises, except enterprises for the 1900 largest groups of enterprises or for branches where VAT turnover is not usable;
c) automatically correct systematic errors in VAT turnover and filled-in questionnaires;
d) impute missing turnover values for VAT units and enterprises;
e) deal with differences in definition between VAT turnover and ‘statistical’ turnover;
f) use top-down editing, including interactive correction of influential errors.

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.

Hoogland, J., 2009, *Detection of potential influential errors in VAT turnover data used for short-term statistics*. Paper for Work Session on Statistical Data Editing in Neuchâtel, Switzerland, 5-7 October 2009.

Ouwehand, P., 2010, *Detection and correction of systematic errors in VAT declarations (In Dutch).* Internal note, Statistics Netherlands.