Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Modelling measurement errors to enable consistency between monthly and quarterly turnover growth rates

Arnout van Delden, Sander Scholtus, and Nicole Ostlund[1]

Abstract

For a number of economic sectors, Statistics Netherlands (CBS) produces turnover growth rates of businesses: monthly figures based on a sample survey and quarterly figures mainly based on administrative data. CBS aims to benchmark the monthly growth rates on the quarterly ones in order to produce consistent output. Preliminary results of benchmarking showed that the quarterly administrative turnover growth rates turned out to be relatively large in the fourth quarter of the year compared to the survey data whereas the opposite was true in the first quarter. This effect is probably caused by quarterly patterns in measurement errors, for instance due to administrative processes within businesses. We present a methodology, based on a mixture regression model, that aims to automatically detect such measurement errors.

Key Words: Mixture models; Measurement errors; Reporting errors; Tax data; Seasonal patterns.

1. Introduction

For a number of economic sectors, Statistics Netherlands (CBS) produces two turnover time series: a monthly series based on sample survey turnover and a quarterly series based on census data. The census data consist of a combination of Value Added Tax data (VAT) for the smaller and simple enterprises and of survey data for the more complex enterprises. The smaller and simple enterprises are referred to as non-top X units and the more complex ones as top X enterprises.

The monthly time series is used to publish output for the short-term statistics and it is input for early releases of the quarterly national accounts. The sum of the quarterly level estimates based on the census data is used to calibrate the outcomes of the annual structural business statistics, which in turn is input for later releases of the annual national accounts. This way, differences between the two time series contribute to differences between early and late releases of the national accounts figures. To improve the quality of our output, we aim to benchmark the monthly time series upon the quarterly one, using a Denton method (Bikker et al., 2013; Denton, 1971).

CBS aims to benchmark the two series from 2015 onwards. However, preliminary results of benchmarking of the 2015 Retail trade data showed that the year-on-year growth rates of quarterly turnover from the survey were adjusted downwards in the first quarter of the year and upwards in the fourth quarter of the year (see Van Delden and Scholtus, 2017). Depending on the quarter, those adjustments were close to or exceeded the 95 per cent margins for the year-on-year growth rates of Retail trade of 0.7 per cent points (Scholtus and de Wolf, 2011). Based on a preliminary analysis, Van Delden and Scholtus (2017) found that the original quarter-on-quarter growth rate in the third quarter of 2015 was 8.3 per cent points whereas it reduced to 7.9 without the estimated seasonal effect. This difference corresponded to more than 100 million euros turnover. CBS considered this effect to be too large. This result led to the first question of the current paper: to what extent are there systematic seasonal reporting differences between the survey and tax data?

[1]Arnout van Delden, Statistics Netherlands, P.O. Box 24500 2490 HA, The Hague, The Netherlands (a.vandelden@cbs.nl); Sander Scholtus, Statistics Netherlands, P.O. Box 24500 2490 HA, The Hague, The Netherlands (s.scholtus@cbs.nl); Nicole Ostlund, Statistics Netherlands, P.O. Box 24500 2490 HA, The Hague, The Netherlands

Applying automatic methods of benchmarking requires that the two series have been corrected for large and for systematic measurement errors. Correction of measurement errors in business statistics often consists of a combination of automatic and manual editing. Manual editing is often restricted to a limited number of most influential records. A second aim of the current paper is to find out whether the observed seasonal differences are caused by large measurement errors in a limited set of influential units or due to systematic errors in a larger set of units. That determines whether the error correction can be done by manual editing or by applying a generic correction method.

2. Empirical data

We compared survey with VAT turnover of non-top X units on a quarterly basis, using 2014, 2015 and 2016 data of the economic sectors Manufacturing, Construction, Retail trade and Job placement. Manufacturing, Construction and Retail trade are sectors with a monthly survey. Until recently, output of Job placement was on a quarterly basis and it was based on a quarterly sample survey. Nowadays its output is completely based on VAT. We included Job placement in the study because preliminary results showed that this sector might have clear seasonal effects, which helps to understand effects in the other sectors. All our analysis are based on microdata that are classified by economic activity according to the NACE classification. The term 'economic sector' roughly corresponds to the first-digit NACE code, whereas the term 'industries' refers to more detailed NACE codes. Seasonal effects were analysed for each economic sector separately, rather than for each industry, because the effects were too subtle to be estimated accurately with the amount of data available at industry level.

The VAT and survey micro data that have been used to produce the output of the two time series were linked at the level of the statistical units, the enterprises, using a unique enterprise identification number. Within those linked data, four categories of units were omitted:

1. units that were likely to have a 'thousand error' (see section 2.3 in Van Delden and Scholtus, 2017);
2. units that were not present in both data sets for all four quarters within a year;
3. units that did not report their turnover in all four quarters of the year;
4. industries for which the turnover level estimates or change estimates based on VAT are considered unreliable because of definitional differences between VAT and survey turnover.

We refer to the final set of units as 'selected' units. We have applied those four selections to ensure that the seasonal effects that we find are not due to other factors. Van Delden and Scholtus (2017) showed that the seasonal effects were not very sensitive to those selections. Some basic figures on the non-top X population are given in Table 2-1.

Table 2-1
Basic figures on the non-top X population per economic sector: total turnover (T in 10^9 euros) based on VAT, population size (N in 10^3 enterprises) and total number of selected units (n enterprises)

Year	Manufacturing			Construction			Retail trade			Job placement		
	T	N	n	T	N	n	T	N	n	T	N	n
2014	21.5	56.6	2296	12.3	143.3	863	12.9	110.4	2070	3.6	12.2	1290
2015	22.4	58.5	2187	13.1	149.7	740	13.5	115.1	1590	4.0	12.5	936
2016	23.3	60.3	2271	14.3	156.5	735	14.1	117.8	1627	4.4	12.8	1086

3. Are there seasonal reporting differences?

3.1 Methodology

Van Delden and Scholtus (2017) showed that the relationship between quarterly survey and VAT turnover can be described well by a simple linear model, where the slope varies with the quarter of the year in combination with a common intercept. We applied a regression analysis with VAT turnover as the independent variable and sample survey turnover as the dependent variable. We are aware that both sources may contain measurement errors, and errors in the independent variable may cause an underestimation of the slopes of the regression analysis. In section 4.2 we describe results for an extended model, including a group of units without a quarterly effect, i.e. with a yearly slope. The slope of those units was very close to 1, suggesting that the effect of underestimating the slopes was nearly negligible. Note that, after exclusion of industries with definitional differences between VAT and survey turnover (i.e., the fourth category mentioned above), we would expect the true slope to be 1 in the absence of random measurement errors.

We therefore applied the following linear model within a given year. Let x_i^q denote the VAT turnover for quarter q of enterprise i and let y_i^q be its sample survey turnover. Further, let α be the common intercept, $\beta^{q=1}$ be the slope for quarter 1 and let $d\beta^{q=q*}$ stand for the difference in the slope between quarter $q = q^*$ and quarter 1. Finally, let $\delta_{q*}^q \in \{0,1\}$ be a dummy variable that indicates whether $q = q^*$, with $q^* \in \{2,3,4\}$. We used the following basic model:

$$y_i^q = \alpha + \left(\beta^{q=1} + d\beta^{q=2}\delta_2^q + d\beta^{q=3}\delta_3^q + d\beta^{q=4}\delta_4^q\right)x_i^q + \varepsilon_i^q \qquad (1)$$

Here, ε_i^q is a disturbance term. We assumed that ε_i^q is normally distributed with mean 0 and its variance varies with weights ω_i^q of the units, according to $\tilde{\sigma}^2/\omega_i^q$. These weights account for heteroscedasticity in the data.

We extended the model (1) to account for the presence of outliers in the data. To that end, we used a finite mixture model comparable to Di Zio and Guarnera (2013). We assumed that the data have been generated from a mixture of two sets of units: one set with a small error variance and another set with a larger error variance. We will expand this model to more groups of units in section 4. This two-group mixture model (M2) was given by:

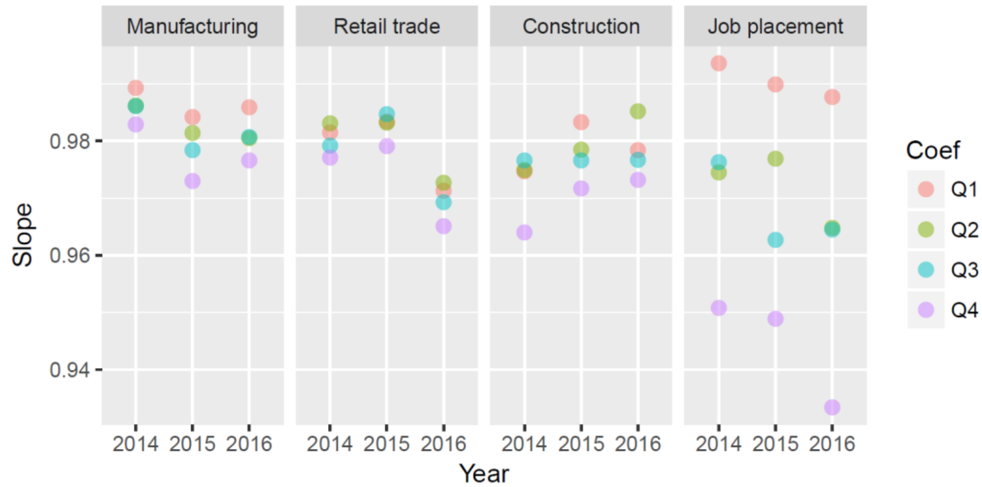
$$y_i^q = \alpha + \left(\beta^{q=1} + d\beta^{q=2}\delta_2^q + d\beta^{q=3}\delta_3^q + d\beta^{q=4}\delta_4^q\right)x_i^q + \varepsilon_i^q + z_i e_i^q \qquad (2)$$

where $z_i \in \{0,1\}$ denotes an unobserved indicator with $P(z_i = 1) = \pi$, and e_i^q is an additional, normally distributed disturbance with mean 0 and a variance $(\vartheta - 1)\tilde{\sigma}^2/\omega_i^q$ that only affects units with $z_i = 1$. The conditional expectation of z_i given the observed data for unit i is denoted by τ_i. This can be interpreted as a group membership probability. It is assumed that ε_i^q, z_i and e_i^q are mutually independent. Under this model, the variance of the disturbance term for a given unit is inflated by a factor ϑ when $z_i = 1$. Note that we assumed that units are assigned to the same group for a whole year.

We used the set of selected units, as described above, to estimate formula (2). We used an estimator for (2) that includes a calibration weight, where this weight is defined as the ratio of the population size to the size of the set of selected units per sampling stratum. A sampling stratum is given by a combination of an industry by a 1-digit enterprise size class. The parameters of the model were estimated by an Expectation Conditional Maximisation (ECM) algorithm similar to that of Di Zio and Guarnera (2013). Details can be found in Van Delden and Scholtus (2017).

3.2 Results

Figure 3.2-1
Estimated slopes based on non-top X units that report both to the survey and the VAT data. Quarters are numbered from the first quarter of 2014 onwards



For all four economic sectors and all three years, the estimated slope in the fourth quarter was smaller than that of the first quarter (see Figure 3.2.-1). The absolute size of $d\beta^{q=4}$ was largest for Job placement (range: -0.054 to -0.041), followed by Manufacturing (range: -0.006 to -0.011), Construction (range: -0.005 to -0.008) and Retail trade (range: -0.004 to -0.006). For each slope effect coefficient ($d\beta^{q=q*}$ in equation (2)) we computed the p value of the hypothesis that its value is 0. The effects were strong for Manufacturing (p values for all years < 0.01) and weaker for Construction and Retail trade where part of the p values were between 0.05 and 0.10. For all economic sectors and all years, the p value of the slope effect coefficient $d\beta^{q=q*}$ was smallest for the fourth quarter of the year.

4. Are reporting differences due to a limited set of units?

In a preliminary analysis we computed the contribution of each of the units to the estimated slopes of the two-group mixture model. We sorted the units according to the absolute value of this contribution and found that the quarterly slope differences could not be explained by just a limited set of units. We wanted to better understand which units contribute to the seasonal reporting differences by extending the mixture model. This extension is described in the next section.

4.1 Methodology

We tried out a number of extensions for the two-group model. In those extensions we allowed for more groups of units, where each group has its own quarterly or yearly slope and its own variance. We modelled this by introducing a dummy z_{gi} which is 1 when unit i belongs to group g and 0 otherwise. The symbol τ_{gi} stands for the expectation of z_{gi} given the observed data for unit i. Furthermore, we compared three structures for the variance-covariance matrix of the four quarterly disturbances for the same unit (Σ): diagonal, banded and free. In case of a diagonal structure, all diagonal elements have the same, positive, value and all other elements are zero. Note that this variance-covariance structure was assumed for the above two-group model. In case of a banded structure, all elements of Σ on a sub-diagonal at the same distance to the main diagonal have a common value. In case of a free structure the only restriction is that Σ is a symmetric, positive semidefinite matrix.

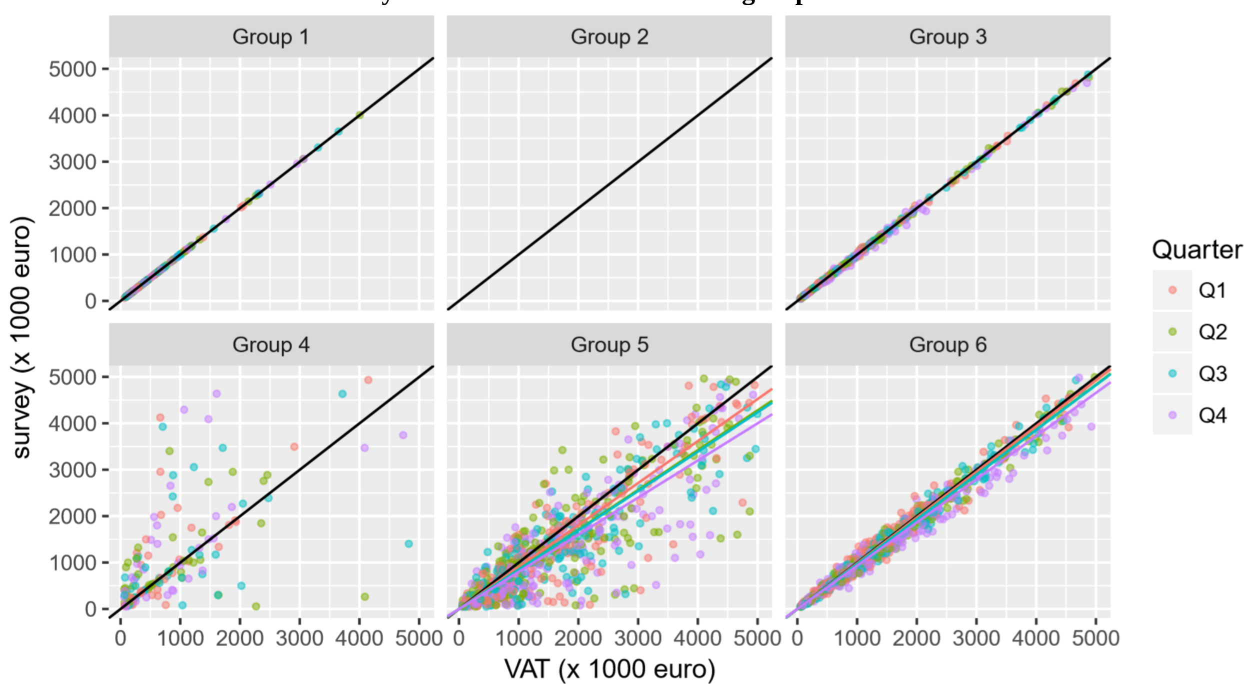
An extension of the ECM algorithm was used to estimate these mixture models. All models were started with a range of starting values and the solution with the best likelihood value was selected. To compare the performance of different models we computed the Akaike and Bayesian Information Criteria (AIC and BIC) and the so-called ICL-BIC (McLachlan and Peel, 2000). The ICL-BIC adds a term to the BIC based on the entropy of the group-assignment probabilities τ_{gi}, to measure how well the model is able to assign individual units to a single group.

4.2 Results

The best performing model depended on economic sector and year, but overall a model with six groups was performing best. This model is denoted by M6. Furthermore, the banded and free structures of the variance-covariance matrix always outperformed the diagonal structure, and the differences in estimated slopes between banded and free were usually small.

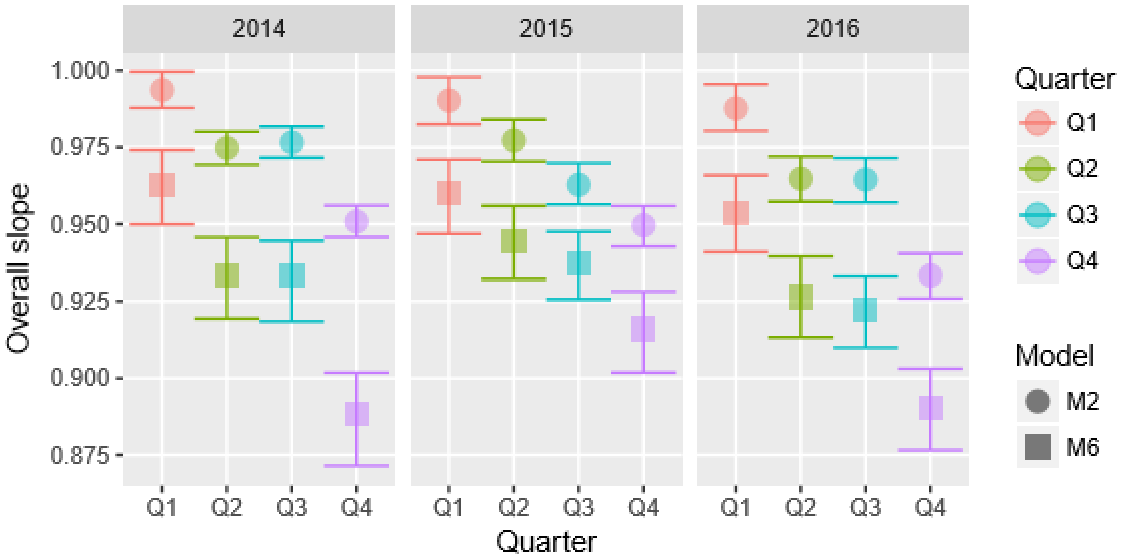
For Job placement, the best performing M6 model had a free variance-covariance structure. The relation between quarterly survey and VAT turnover in 2016 for the six groups is plotted in Figure 4.2-1. The first group (4.8% of the units) are units that report nearly the same turnover values in both sources. The second group (0.0% of the units) are units that erroneously included a VAT rate into their reported turnover. The third group (14.7% of the units) concerns a group that has a larger variance than group 1 but the same slope. Group 4 (6.9% of the units) are units with very large outliers. Group 5 (34.0% of the units) and 6 (39.6% of the units) are units with seasonal effects. The quarterly effects in group 5 are larger than in group 6, their quarterly slopes are smaller and the variance is larger. The black line in Figure 4.2-1 indicates the common estimated yearly slope in groups 1 to 4, the coloured lines indicate estimated quarterly slopes in groups 5 and 6.

Figure 4.2-1
Plot of the relation between survey and VAT turnover for the six groups within the M6 model for 2016



For the M6 model, we computed weighted average quarterly slopes, using the group membership probabilities τ_{gi}; see Figure 4.2-2. We found that found the slopes according to the M6 model were smaller than for the M2 model, but the relative differences between quarters were similar.

Figure 4.2-2
Weighted average quarterly slopes for the two-group mixture model (M2) and the six-group mixture model (M6). Model error bars give the 95 confidence intervals, computed by a bootstrap procedure



5. Conclusion and discussion

Using a simple two-group mixture model, we found seasonal effects in all four economic sectors and throughout three subsequent years. This strongly suggests that there are indeed systematic seasonal reporting differences between survey and VAT turnover. Results of the extended mixture model indicated that a large group of about 75% of the units may contribute to these seasonal effects. According to the M6 model, which fitted the data well, these quarterly reporting differences can be accounted for by two different groups of units in the population: one group with rather large quarterly effects and slopes well below 1 and a large variance, and a group of units with smaller quarterly effects, with slopes closer to 1 and with a smaller variance.

As a next step, we will try to understand the causes of those quarterly patterns in terms of the actual administrative reporting behaviour of businesses. To that end, we wish to interview employees of administration offices and a selection of businesses with specific reporting patterns. Using that information we want to learn for which of the two series and for which of the units the reported seasonal patterns have the smallest measurement errors. We will use this information to derive an approach to correct the seasonal effects in either the survey or VAT data, or both, in order to facilitate future benchmarking.

The use of mixture models for detection of measurement errors has been proposed before in official statistics, for instance by Di Zio and Guarnera (2013) and Guarnera and Varriale (2016). We have extended that approach to the detection of seasonal reporting effects in two sources. The application to VAT and survey data may also be relevant for other countries that use, or plan to use, VAT as a source for turnover. More generally, there may be also other infra-annual administrative data sources that suffer from such effects and which could benefit from this approach.

Acknowledgement

We thank Jeroen Pannekoek for critically reviewing a preliminary version of this paper. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

References

Bikker, R.P., J. Daalmans, and N. Mushkudiani (2013), “Benchmarking Large Accounting Frameworks: a Generalised Multivariate Model”, *Economic Systems Research*, 25, pp. 390-408.

Delden, A. van, and S. Scholtus (2017), “Correspondence between survey and administrative data on quarterly turnover”, *CBS Discussion paper 2017-3*.

Denton, F.T. (1971), “Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization”, *Journal of the American Statistical Association*, 66, pp. 99–102.

Di Zio, M., and U. Guarnera (2013), “A Contamination Model for Selective Editing”, *Journal of Official Statistics*, 29, pp. 539–555.

Guarnera, U., and R. Varriale (2016), “Estimation from Contaminated Multi-Source Data Based on Latent Class Models”, *Statistical Journal of the IAOS*, 32, pp. 537–544.

McLachlan, G.J., and D. Peel (2000), *Finite Mixture Models*, New York: John Wiley & Sons.

Scholtus, S., and P.P. de Wolf (2011), “Sampling design for the short-term statistics”, unpublished report, Den Haag, The Netherlands: Statistics Netherlands.