Combining multiple administrative data sources in estimating Structural Business Statistics variables

Guy Vekeman

Statistics Belgium, e-mail: guy.vekeman@economie.fgov.be

Abstract

Structural Business Statistics relies on both administrative and complementary survey data in order to compile statistical variables. Among the administrative data sources, the profit and loss accounts are most valuable. When making estimations VAT tax data may be used alongside profit and loss accounts. While VAT data also refer to concepts such as turnover, purchases or investments, their definition differs from that generally used in accounting. This is the major challenge in the combined use of these administrative sources to make estimates of a complete set of SBS variables for a given enterprise.

Keywords: Imputation  Robustness Proxy

1. The need for imputations

SBS variables are normally compiled using a number of profit and loss account codes with other items retrieved from the data entry fields of the XBRL on-line SBS survey forms. These forms include the administrative totals the enterprise has communicated in its profit & loss account and balance sheet. As such the respondent can quickly assess whether the data entered are coherent.

In an effort to reduce the response burden on small businesses, their survey frequency was reduced to one year out of three, whereby a rotation scheme is applied. Alternately small services businesses, small shops and gross trade businesses and small construction and manufacturing enterprises are surveyed. The missing information during two non-survey years is then imputed, using the complete data of the respondent during the previous survey and the two sets of former and new administrative data.

2. Current Status

In the on-going project on estimating variables for structural business statistics, we have been analysing the match between imputations made for respondents and their true data having come in. Three different situations may apply:
• Complete accounting scheme: then VAT data are normally not used as accounting data on turnover and purchases are available with definitions allowing to calculate a fairly decent proxy,

• Abbreviated accounting scheme: then VAT data may be used as accounting data on turnover and purchases are not always available and their lack of completeness doesn't allow calculating a decent proxy,

• Self-employed (or small employers) that don't file accounts: VAT data are then the only administrative source for turnover, purchases and investments. These rough proxies are but a guestimate on turnover and purchases.

3. Strength and weakness of a ratio imputation scheme

When making an estimate for a SBS variable, it needs to be emphasized that the major contribution of the main variables is found in the profit and loss account. Survey data fields generally constitute a correction on that administrative proxy. The ratio imputation scheme assumes these corrections to be enterprise specific. An enterprise reporting consistently on a survey data entry field will see the related SBS variable fluctuate proportionally to the administrative proxy. In such case, the administrative proxy consistently over- or underestimates the true SBS variable by a certain percentage. The ratio estimation scheme will correct at its best if the above assumption holds. Secondary variables within SBS generally derive from one main variable. All cost or input related characteristics have their code starting with ‘13’, while the codes for the main revenue related variables start with ‘12’. The current ratio imputation scheme assumes the cost structure to remain fairly constant. A retailer will consistently spend most of its total purchases on goods for resale. That observed ratio is then consequently applied to the total of purchases for the two years following the survey. Alternatives are being studied, as published by D. Lewis (2012) within the framework of the on-going ESSnet Admin Data (WP3) project, sponsored by Eurostat.

While ratio imputation gives very reliable results for a large majority of enterprises, the aggregates prove error prone due to some mismatches between administrative proxies and statistical variables. Moreover, whenever an enterprise in the year observed was in an atypical situation: e.g. just after starting up or activity suspended for whatever reason, it is not possible to calculate reliable estimates for the business running normally. Any mismatch between the statistical variable and its administrative proxy counterpart will in this case only aggravate the estimation error.

4. Improving the robustness of the ratio imputation scheme

A first benchmark exercise consists in comparing a complete set of SBS variables observed for any respondent to the set of prior estimations of SBS variables made for that same enterprise. The results of that exercise will be reported on in more detail.
The project of “Improving the robustness of a ratio imputation scheme” is aiming to target the mismatches between statistical variables and their administrative proxies of the same accounting period. The main goal is finding a procedure capable of avoiding flawed estimates. The current imputation program procures a set of Eurostat defined SBS variables. Therefore, the automated input editing program designed to checking input fields on existing XBRL electronic questionnaires is not fit for this purpose. A number of ad-hoc tests have been applied to check for potential outliers. As the number of potential flaws is considerable, there is a certain risk of misallocation of editing efforts. While on the one hand this leads to over-editing and a potential introduction of bias, some other anomalies have remained undetected.

Existing correlations, identities and inequalities among the output set of (historic) Eurostat variables and the ex-ante knowledge of any mismatches between these variables and their historic administrative proxy counterparts, can be used direct the imputation procedure along different programming branches where deemed necessary. The project aims at implementing these features in the imputation procedure.

### 4.1 Improving the robustness of VAT data based imputations

Definitions of VAT turnover and VAT purchases differ of those used in accounting. Moreover, though accounting definitions conceptually are closer to the SBS variables, some important differences exist. VAT based proxies being only a distant nephew of the SBS variables is in fact a strong rationale for using ratio imputation. The risk is that the evolution of the VAT proxies may differ from those of the SBS variables if enterprises were surveyed yearly.

A first observation is on *large enterprises* which rely on (generally complete) accounting schemes providing reliable proxies. We now check how the SBS variables for turnover and costs relate to the VAT proxies, using a distribution plot of the ratio. Firstly, it needs to be emphasized that 120 or 3% of the enterprises have at least one outlier for this ratio: this means that reported turnover and/or purchases are over three times larger than the VAT proxy. Using VAT data as proxies would in these cases yield an underestimate of the true values.

As shown in figure 1, the distribution luckily peaks at unity value, confirming that the reported turnover SBS variable is most often coinciding with the VAT proxy. A close match for 60% of the cases and an acceptable match (within -20% to + 25%) in 86% of the cases are more or less reassuring.

The situation is more cumbersome for VAT purchases which often seem to exceed the SBS purchases variable (13 110). The distribution peaks around 0.66 and averages 0.76, making the VAT purchases proxy a highly biased (almost consistently overestimate) of the purchases SBS variable. The ratio distribution also is a lot broader; proving that the VAT proxy is also is a much less reliable estimate for the SBS variable. An analysis of
1000 SME’s effectively surveyed in 2010 provides similar results, with the modal ratio between reported SBS purchases and its VAT proxy now standing at 0.85 (rather than 0.66). At the other side, the smaller number of observations makes that modal value less reliable.

Fig 1: Ratio of turnover as reported (SBS variable 12 110) to the VAT turnover proxy

Fig 2: Ratio of purchases as reported (SBS variable 13 110) to the VAT purchases proxy
By consequence, it is expected that VAT (turnover – purchases) as a VAT proxy for value added is most likely to be a highly biased underestimate of its statistical counterpart, the SBS variable 12 150. There is no clear proof for such discrepancy. A possible reason is that VAT records list purchases VAT inclusive. Businesses reclaim VAT for their purchases. Hence, in the profit and loss accounts purchases are likely to be listed VAT exclusive. Another possible discrepancy may be that VAT declarations break down enterprise total purchases differently into current expenses and investments (fixed capital formation).

4.2 Ratio imputation: its successes and failures

In previous section, we have studied how two main statistical variables relate to their VAT proxy. Even though VAT purchases proxies are biased relative to the statistical variable, ratio imputation still may work out well, provided that said bias is constant over time. A good indication is checking the evolution over time of the VAT proxies.

Table 1: Evolution over time of VAT declared turnover and VAT purchases.

<table>
<thead>
<tr>
<th>Evolution 2010-2011</th>
<th>Turnover</th>
<th>Purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median value</td>
<td>1.064</td>
<td>1.082</td>
</tr>
<tr>
<td>Average value</td>
<td>1.123</td>
<td>1.144</td>
</tr>
</tbody>
</table>

The evolution over time is similar for both proxies and purchases slightly outpacing turnover is consistent with the deterioration of the business climate. For the median enterprises or any enterprise in the bulk of the SBS variable to VAT proxy distribution, ratio estimation is providing decent estimation results. The above average values concern the distribution of year-to-year ratios cut off at a maximum of four (or a 300% increase).

Imputation failures that threaten the correctness of the aggregates concern several cases:

- Purchases ratio values near zero implying VAT proxies much larger than the statistical purchases variable as reported by the business. The number of such anomalies largely exceeds that of the equivalently defined VAT turnover ratio. Not only are VAT purchases proxies biased, a significant number of them have values far beyond what is reported by the businesses.
- VAT turnover or purchases year-to-year growth in excess of 4 (or over 300% higher). Whereas it is obvious that the VAT year-to-year growth of the VAT proxies sometimes is in excess of said 300%, few extreme values arise. Quite often they originate from a business just starting up or having suspended activities when surveyed. VAT proxies in the reference period then are extremely small and ratio imputation cannot possibly provide a reliable estimate for the business up and running normally.
- One of the elements missing or zero (the historic statistical variable or one of the VAT proxies). The ratio imputation then is either undefined or zero by definition.
As ratio imputation fails, the VAT proxies are a possible set of values for turnover and purchases. In this case, the bias of the purchases value needs being compensated for. Assuming that the bias originates from including VAT paid on purchases, we may compensate for this. VAT records don’t provide any detail on the VAT percentage paid; however the prevalent high VAT rate of 21% applies for most goods and services, except food and a few other products or services. As a rule of the thumb we divide VAT purchases by 1.2 to obtain an unbiased estimate of purchases as would be reported in the profit and loss account.

Imputing the proxy for turnover and the tuned purchases proxy also yields a proxy for value added. This set of values enables calculating the majority of the financial SBS variables, using a suitable donor enterprise for a hot-deck imputation of the ratios between the auxiliary variables and these three main variables.

At the start of this section, it was pointed out that median and average year-to-year growth of turnover and purchases are closely related. On the individual enterprise level, this unfortunately is not the case. The below graph shows the natural logarithm of growth of turnover and purchases plotted against one another. The origin indicates a company with stable turnover and purchases.

![Fig. 3 : Log-log representation of the year-on-year evolution of the VAT turnover proxy (LCA_Grow) and the VAT purchases proxy (LCST_grow). Natural logarithms are chosen, 1 meaning a multiplication by e=2.718.](image)

Most observations are predictably scattered on an increasing slope, with turnover and purchases fluctuating proportionally. Quite a few observations are atypical, with turnover
increasing while costs are reduced (upper left) or costs soaring but turnover declining (observations at the lower right). An atypical large or small value in the denominator (the previous observation) is most likely responsible for extreme variations. In the atypical cases, it needs to be checked whether the administrative proxies match the statistical variables at the previous observation. If so, the imputation process is yielding a very atypical observation, with value added completely distorted by the incoherent evolution of turnover and purchases. A mismatch between statistical variables and their proxies is a clear counter indication for the use of ratio estimation, especially when the administrative proxies seem to evolve out of lock. However, even with growth for both administrative proxies coherent, ratio estimation can be error prone for large growth observations (the upper right corner of the graph). It then is concluded that the business in the new statistical year is no longer in the same league as it was at the previous observation.

4.3 Improving the robustness of the imputation scheme for small enterprises using an abbreviated accounting scheme.

The vast majority of small employers are businesses using an abbreviated accounting scheme. In this case mentioning purchases and turnover in the profit and loss account is optional, while an operational margin (a good proxy for value added) is compulsory.

Having a triple input allows prior checking of the coherence between VAT proxies and value added as reported in the profit and loss (P&L) account. Whenever the P&L reports turnover and purchases, the VAT proxies will be disregarded, since P&L account values not only are coherent but also are thought to be the more reliable estimates, since their definition is closer to that of the SBS variables.

If turnover is absent from the P&L account, the first option is to use the VAT turnover and approximate purchases by subtracting value added. Until two years ago, this is how proxies were obtained. One advantage is that a coherent set of proxies is obtained, but we completely disregard VAT purchases, albeit that this is the less reliable proxy.

As with any “one tool fits all” approach, we learnt that quite a few anomalies were generated in this way. In few aberrant cases the value added reported exceeds VAT turnover, yielding a negative estimate for purchases. This may result from a mismatch between VAT turnover and accounting turnover. The former may include credit notes paid or revenues from the sale of assets. In both cases VAT turnover differs from accounting turnover. Apart from being defined differently, the value of VAT turnover generally is not consistently smaller or larger than accounting turnover, but the difference may also vary over time. This prior knowledge of potential sources of any mismatch between VAT turnover and accounting turnover may result in finding a better proxy, which is a calculation result of VAT fields (including those on credit notes) and accounting data on asset sales. Similarly purchases need to be compensated for credit notes received and for VAT paid. Comparing the VAT defined proxy for investments to
the one obtained in accounting also may contribute to finding a more reliable and less biased proxy for purchases as would be reported in the P&L account. After obtaining better proxies, their coherence: how well ‘purchases’ plus ‘value added’ equals ‘turnover’, is what is to be monitored. It may pay off to apply a calibration procedure. Observed values $T_o$ for turnover, $P_o$ for purchases and $V_o$ for value added do not satisfy the constraint in formula 2. We need to carefully choose the weights, taking into account how well the proxies are likely to fit the statistical variables. Purchases is the less reliable proxy and hence has the lower weight. The value added proxy obtained from the P&L account is most reliable and therefore gets the higher weight.

$$w_v(T - T_o)^2 + w_p(P - P_o)^2 + w_v(V - V_o)^2 = min$$ (1)

Whereby

$$T' = P' + V$$ (2)

### 4.4 Employment data

Data on employment and personnel costs are obtained directly from administrative sources without any imputation procedure using previous information on the enterprise. Both social security data (ONSS) and a few personnel related items in yearly accounts provide for reliable estimates on employment and personnel costs. The estimation procedure has been reported on before: Vekeman (2012).

Inserting those variables, it can be checked whether they are coherent with whatever estimates for the financial variables are obtained. Personnel costs generally should be smaller than the gross margin, less the business will have a negative cash flow and may head for bankruptcy. It is however clear that personnel data only provides some useful ratios, which can be benchmarked against aggregate sector data. No hard checking is possible.

### 4.5 The imputation procedure for enterprises using a complete accounting scheme

Large enterprises use a complete accounting scheme, whereby the profit and loss account includes much of the information necessary to obtain excellent proxies for the different statistical variables. The match between these variables and their proxies therefore is a lot better than in previous cases described above.

As most of these enterprises are surveyed yearly, the imputation procedure is used mainly as a back-up to provide timely estimates in case of non-response or a late response.
VAT data are not used in this case and proxies are coherent. However it is still necessary to carefully monitor the performance of the imputation procedure, since any errors can be very influential on the aggregates calculated.

5. Conclusion

Whereas a ratio estimation method provides reliable results for a majority of cases, the procedure may lead to few aberrant imputations, possibly making aggregates less reliable. Identify such cases is possible using some basic analysis techniques. Trying to enable some methods to obtain better estimates is still an on-going challenge.

In suspect cases, improved estimates are obtained when directly using administrative proxies in the estimation procedure. However it then is necessary to prevent introducing any bias due to definition differences, resulting from the different legal requirements for VAT declarations and for yearly accounts.

References
