# Short-term turnover statistics based on VAT and Monthly Business Survey data sources

*Li-Chun Zhang (L.Zhang@soton.ac.uk) and Alison Pritchard (alison.pritchard@ons.gov.uk)*

## 1 Introduction

There is currently a considerable drive at the National Statistical Offices to exploit the potential of administrative data in statistical production. For short-term business statistics, the VAT register is an important source and often the first one to be examined. A number of investigations have previously been carried out at ONS, such as forecasting VAT turnover at the unit-level, adjusting VAT register totals towards the existing MBS-based turnover estimates, *etc.* Some critical methodological questions, however, remain to be addressed, such as:

- The existing approach to target-frame construction may be practical and yielding acceptable results for MBS-based estimation. But is it the most suitable?
- Coherence between MBS and VAT-based estimates is important. But how should one take into account the errors in both when looking at the difference between them?

We propose in general to decompose the investigation in two parts: (I) prediction of VAT mature turnover total using VAT early and historic reports, and (II) harmonisation in concept and reconciliation in number between VAT mature totals and MBS estimates.

Like any observational data, the VAT reports are intrinsically associated with uncertainty. Prediction and design-based theoretical frameworks have long been established for statistical surveys based on sample data. New and different frameworks are needed for administrative data, because the nature of uncertainty (or sources of errors) is different from that arising in survey sampling. We shall refer to the VAT register as a *longitudinal progressive* data set based on two characteristics that are most relevant in the present context. It is longitudinal because various measurements such as turnover are recorded for different time points. It is progressive because the measurement of a given reference point is not fixed over time, due to delays in reporting and changes to the previously recorded values – this is a distinct feature that does not figure in traditional sample survey theory.

Here, due to limited space, we focus on two issues. First, in Section 2 we describe a general prediction framework for longitudinal progressive data. Next, in Section 3, we outline a two-part solution combining VAT and MBS as well as balancing both timeliness and precision requirements, and highlight some empirical evidences in support for this. More details can be found in "Towards VAT register-based monthly turnover statistics" (Zhang, 2013).

## 2 Prediction framework for longitudinal progressive data

It is convenient to phrase in terms of the VAT data, but it should be clear that the following exposition is equally valid for other longitudinal progressive data arising from administrative reporting that experience delay in reporting and changes of the previously recorded values.

The target population for *statistical period* $t$ consists of all the *VAT-active* units in period $t$, i.e. units with turnover that generates non-zero VAT, denoted by $U(t)$. The target total is the total VAT-generating turnover for units in $U(t)$, denoted by $Y(t) = Y_{U(t)}$.

Let $y_i(t;s)$ denote the $y$-value of unit $i$ according to the VAT-register at *measurement time point* $s$, for $s \geq t$. We put $y_i(t;s) = 0$ if no value is recorded for unit $i$ by time $s$. This can happen either because unit $i$ is inactive in period $t$, or because there is a delay in reporting such that the $y$-value is not yet recorded. Otherwise, $y_i(t;s) \neq 0$. However, it may happen that $y_i(t;s') \neq y_i(t;s)$ for $s' > s$ if changes are made to $y_i(t;s)$ for various reasons between time $s$ and $s'$. The VAT-active population for $t$ measured at $s$ is given by

$$U(t;s) = \{i; y_i(t;s) \neq 0\} \tag{1}$$

Two simplifying assumptions can be useful, at least for an initial development:

$$y_i(t) = \lim_{s\to\infty} y_i(t;s) \overset{def}{=} y_i(t;\infty) \tag{2}$$

$$y_i(t) = \min_{s:y_i(t;s)\neq 0} y_i(t;s) \tag{3}$$

By (2), over time the measurement of interest converges in the VAT register. Empirical evidences suggest this holds for the UK data: for practical purposes the convergence takes place within 6 - 12 months after the reference time point $t$, and we shall refer to $y_i(t;t+12)$ as the *mature* $y$-value and assume $y_i(t;t+12) = y_i(t;\infty)$. It follows that

$$\begin{aligned} U(t) &= \lim_{s\to\infty} U(t;s) = U(t;t+12) \\ Y(t) &= \lim_{s\to\infty} Y(t;s) = Y(t;t+12) \qquad \text{for} \quad Y(t;s) = Y_{U(t,s)} = \sum_{i\in U(t;s)} y_i(t;s) \end{aligned}$$

Next, in retrospect, the progressiveness of the VAT register consists overwhelmingly of delays in reporting rather than changes of previously reported values, leading to the second assumption (3). This simplifies the prediction form, which will be given in (5).

Denote by $t+d$ the *estimation time point* for $Y(t)$, where $d > 0$ is the *production lag*. Put $I_i(t;s) = 1$, if $y_i(t;s) \neq 0$, and 0 otherwise, indicating whether an *active* VAT report has arrived by time $s$. Then, at the production time point $t+d$, the target population $U(t)$ can be divided into 3 disjoint sub-sets, denoted by

$$U(t) = U_1(t;t+d) \cup U_2(t;t+d) \cup U_0(t;t+d) \tag{4}$$

where $U_1(t;t+d)$ consists of the *VAT-reports*, and is given by

$$U_1(t;t+d) = \{i; I_i(t;t+d) = 1\}$$

and $U_2(t;t+d)$ consists of the *VAT-delays*,and is given by

$$U_2(t;t+d) = \{i; I_i(t;t+d) = 0 \ \cap \ I_i(t;\infty) = 1 \ \cap \ \sum_{j=1}^{\infty} I_i(t-j;t+d) \geq 1\}$$

and $U_0(t; t+d)$ consists of the *VAT-birth delays*, and is given by

$$U_0(t; t+d) = \{i; I_i(t; \infty) = 1 \ \cap \ \sum_{j=0}^{\infty} I_i(t-j; t+d) = 0\}$$

It may be noted that we refer to a reported 0 turnover value as an *inactive* VAT report. Inactive reports do exist, and we have chosen formally to treat them as the unreported inactive units since $I_i(t; s) = 0$ if $y_i(t; s) = 0$, and to exclude them from the target population (4).

The general prediction form of $Y(t)$ can now be given as

$$\hat{Y} = Y_{U_1(t;t+d)} + \hat{Y}_{U_2(t;t+d)} + \hat{Y}_{U_0(t;t+d)} \tag{5}$$

where $Y_{U_1(t;t+d)}$ is directly given as the (active) VAT-report total by virtue of (3), and $\hat{Y}_{U_2(t;t+d)}$ is the predicted VAT-delay total and $\hat{Y}_{U_0(t;t+d)}$ the predicted VAT-birth delay total. Notice that, without the assumption (3), one must replace $Y_{U_1(t;t+d)}$ in (5) with a predictor $\hat{Y}_{U_1(t;t+d)}$.

A prediction approach to the VAT total requires modelling of $\{(I_i, y_i); i \in U(t)\}$, with or without further conditioning variables such as historic $y$- and $I$-values and other auxiliaries. Several simple alternatives are explored in Zhang (2013), which will not be related here. The modelling of longitudinal progressive data should provide interesting topics for future research.

# 3 UK VAT data

## 3.1 Outline of a two-part solution

The VAT mature total being available in 6 - 12 months after the statistical reference time point, the key objective is to improve the timeliness by predicting the VAT mature total at an earliest possible time point with acceptable accuracy. A two-part solution emerges as the following:

- Monthly turnover report of the largest business units.

  - These can be considered as an MBS sample of only self-representing units.

  - Cut-in threshold based on previous VAT returns should be used both for the initiation and maintenance of the group over time. Emerging in-scope units are admitted on a running basis, and outdated units gradually released.

- The turnover total of the remaining units may be produced as a combination of prediction based on early VAT turnover reports and forecast based on historic VAT turnover reports.

  - The VAT-birth delay total may be produced by projection (Zhang, 2013).

  - For more than 80% of the VAT existent units that are below a cut-off turnover threshold, the total can be produced band-wise by combining the VAT-report total and substitution of the VAT-delay total by the corresponding historic values – see Section 3.2 below.

  - For the turnover total of VAT existent units that fall between the cut-off and cut-in thresholds, two data scenarios for prediction may be considered depending on the timeliness of the VAT-reports and the stability of the reporting process: (1) VAT-reports only: this will be acceptable provided enough VAT-reports can be delivered to ONS before the end of $t+1$; (2) VAT-reports supplemented with a small sample: the current MBS sample size will be reduced due to no sampling of cut-off units and certain NACE groups.

## 3.2 Turnover total of active existent non-large units

Figure 1 gives the results of a substitution exercise. The active VAT existent universe is divided into 7 groups at $c_{(g)} = 0, 4000, 8000, 12000, 20000, 40000, 80000, \infty$. For each reference time point $t$ in 2011, we first substitute the within-group turnover total $Y_{(g)}(t)$ by the corresponding $Y_{(g)}(t-12)$ of one-year old, for all the groups except the one with the largest units (i.e. turnover above 80000). Just below 80% of the units are substituted in SIC-13 and about 85% in SIC-45. The resulting turnover total (square) in SIC-13 is given in the upper-left panel together with the true target total (circle) for the 12 months in 2011. The maximum ARE is 0.2% over this period. Similar results are obtained for SIC-45 in the lower-left panel.

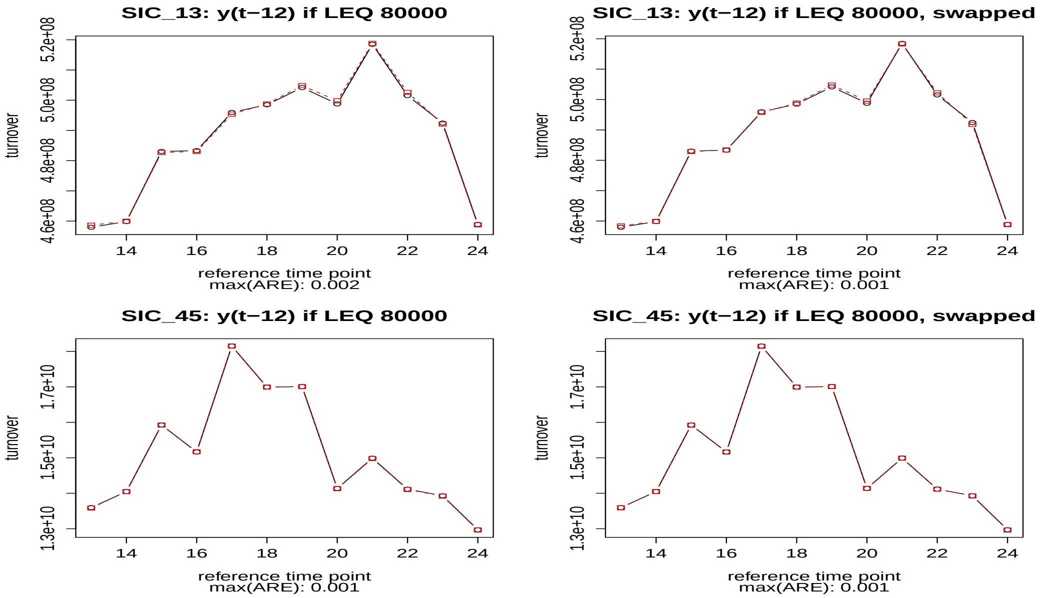

**Figure 1. VAT turnover total of existent units (circle) and by substitution of $t-12$ values for non-large units (square) for SIC-13 and 45, with and without swapping.**

Next, we experiment substitution with SIC-swapping in addition. For SIC-13, we use $\hat{Y}_{(g)}(t|\text{SIC-13}) = N_{(g)}(t-12; t|\text{SIC-13})\ \bar{Y}_{(g)}(t-12; t|\text{SIC-45})$. That is, we use the size of the VAT existent universe in SIC-13 at $t-12$, but the corresponding group-mean of the VAT existent universe in SIC-45 at $t-12$. The results for 2011 are given in the upper-right panel of Figure 1. Similarly for SIC-45 in the lower-right panel.

Within-group forecast based on substitution for non-large existent units using mature historic data is thus shown to be a viable approach in practice. Notice that the threshold value 80000 for non-large units here was chosen arbitrarily before we had examined the data. Further exploration in the same manner will reveal whether the threshold can be further raised.

## 3.3 Population size of active existent units

We apply projection (Zhang, 2013) to predict the population size of active VAT existent units. Given the feasibility of forecasting for the non-large units, we focus on the group of the largest units with turnover above 80000. Figure 2 provides some details. It can be seen from the left panels that, while the reporting rate $\pi^c_{(g)}$ at $t+3$ varies greatly, the proportions of inactive units ($\pi_0$) and the group of largest units ($\theta^c_{(g)}$) are quite stable over time. This provides the basis for successful projection shown in the right panels, where the median ARE of the predicted population size $\hat{N}_{(g)}$ of this group of units is about 1% over the period of 2011.

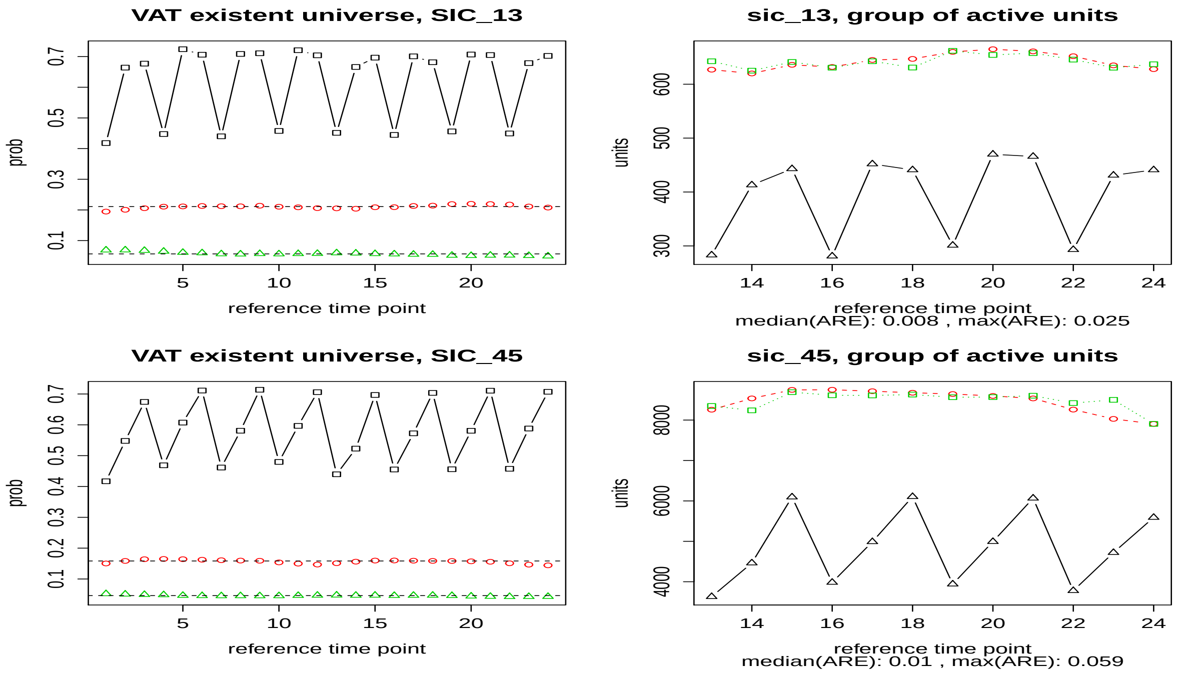

**Figure 2. Left: reporting rate $\pi^c_{(g)}$ at $t+3$ (square) and propotion $\theta^c_{(g)}$ (circle) of units with turnover $\geq$80000, proportion $\pi_0$ (triangle) of inactive units; reference time points in 2010 and 2011. Right: number of reporting existent units $N_{1(g)}$ at $t+3$ (triangle), population size $N_{(g)}$ (circle) and predicted population size $\hat{N}_{(g)}$ (square) of units with turnover $\geq$80000; reference time points in 2011.**

The results suggest to group the units that are between the cut-off and cut-in thresholds, and to separate the prediction of group total in two parts, i.e. that of the group size and that of the group mean, in order to curtail the effect of potential outliers.

## 3.4 Summary

Better uses of the VAT register can be made for (i) construction of the target population, (ii) selection and maintenance of the self-representing units or monthly staggers, (iii) exemption of survey compliance for the more than 80% units that are below the cut-off threshold, and (iv) prediction of the remaining non-self-representing units for which substitution of historic totals is misleading. Some issues remain to be clarified for the implementation to become feasible:

- Harmonisation between VAT register and IDBR regarding units and SIC code.
- VAT data delivery scheme both in terms of time point and frequency.
- Constitution of the monthly sample/staggers and its maintenance over time.
- Evaluation of results in all SIC-groups and at all levels of dissemination.