**Modelling enterprise level statistics using administrative data**

Gary Brown, Jonathan Digby-North, Megan Pope[[1]](#footnote-1)

**Abstract**

The UK is required to provide estimates of Structural Business Statistics (SBS) to Eurostat on an enterprise reporting basis for the 2016 reference year. This represents a new output, as currently the Office for National Statistics (ONS) produces estimates on a Reporting Unit (RU) basis. For those enterprises with more than one RU, any intra-flows within the enterprise must be removed to produce consolidated accounts. A feasibility study was conducted to investigate whether it would be viable to estimate enterprise level statistics (e.g. turnover) using existing data and two external/administrative sources.

Two external/administrative sources were identified as able to provide enterprise level data: FAME and VAT. FAME is produced by Bureau Van Dijk and VAT data is provided to ONS by HMRC. The data were matched to the Inter-Departmental Business Register (IDBR) and multiple regression modelling was used to derive relationships between RU level ABS/IDBR variables and the external enterprise level data.

In general, total RU turnover was a strong predictor of consolidated turnover. When applying the models to ABS data, consolidated turnover was found to be approximately 10-20% lower than the simple sum of RU turnover. Although the findings have been more or less consistent across data sources, there are data quality and implementation issues which require further work to resolve.

Keywords: Administrative data, Regression modelling, Enterprise level statistics

**1. Introduction**

To comply with Eurostat regulation 696/1993, the UK Annual Business Survey (ABS) is required to provide Eurostat with estimates of Structural Business Statistics (SBS) on an Enterprise reporting basis for the 2016 reference year. This represents a new output, as current SBS estimates from ABS are on a Reporting Unit (RU) basis – as this meets UK National Accounts requirements.

A feasibility study was conducted to investigate whether it would be viable to estimate enterprise level statistics (e.g. turnover) using existing data and two external/administrative sources identified as being able to provide enterprise level data. Enterprise level statistics are required for a wide range of variables collected on the ABS, however this study looked at turnover only.

This short report documents the findings of this study. Using data existing from a number of sources, regression modelling was used to create a model to estimate consolidated accounts at enterprise level.

**2. Current Data Collection Method**

For the majority of enterprises held on the Inter-Departmental Business Register (IDBR) each enterprise consists of just one RU. RUs are assigned when a business is birthed onto the IDBR and hold general and survey specific contact details. For a small proportion of enterprises on the IDBR (around 2%), enterprises consist of more than one RU. For these 1,200 enterprises, Local Unit (LU) list reporters have been set up, either because this has been requested or because there is a legal requirement for separate GB and Northern Ireland reporting. Figure 2-1 presents example simple and complex enterprise structures. The complex enterprise structure has three RUs: Retail, Manufacturing and Wholesale.

**Figure 2-1
Simple and complex enterprise structures**



For complex enterprises, providing enterprise reporting estimates for additive variables such as employment is very simple, as the figures for each RU can be summed to provide the enterprise level total. For non-additive variables however, further work is required due to transactions between the RUs, which need to be removed from the accounts. These intra-flows within an enterprise are currently captured on the ABS but cannot be identified. In order to model these 'consolidated' accounts (sum of individual RUs minus transactions between RUs in the same enterprise) for the ABS, two external sources of information have been identified as providing the requisite detail: FAME and VAT.

**3. Data Sources**

Data exist from at least four sources – see Figure 3-1 – all of which suffer from missing data to some degree, and can be linked together by enterprise reference.

**Figure 3-1
Four data sources available to model consolidated accounts**

**3.1 IDBR**

The Inter-Departmental Business Register (IDBR) contains information on around 2.5 million businesses in the UK. It provides the main sampling frame for surveys of businesses carried out by the ONS and by other government departments. It is also a key data source for analyses of business activity

Data are available for all RUs in the UK on the IDBR. The IDBR contains a wide range of variables defining the characteristics of each business. Some variables available include:

* General business structure, with enterprise, RU and LU identifiers.
* Legal information within an enterprise such as PAYE and VAT units
* Turnover for RUs
* Employment for RUs
* Standard Industrial Classification (SIC) for each RU and LU
* Region that RU or LU operates in
* Company Registration Number (from Companies House)
* Birth and Death dates for each RU

**3.2 ABS**

The Annual Business Survey (ABS) is the largest business survey run at ONS and samples approximately 62,000 businesses in Great Britain. The ABS provides information on turnover, purchases, employment costs, capital expenditure and stocks. Data are used to calculate each industry’s approximate Gross Value Added (aGVA) – its contribution to the UK economy. ABS is also the main source of figures for the European Structural Business Statistics Regulation (SBS).

ABS data are only available for a small proportion of RUs on the IDBR.

**3.3 FAME**

FAME is a source of comprehensive information on companies in the UK and Northern Ireland produced by Bureau Van Dijk, an external company. Data are available for over 9 million companies, covering 2 million companies in detail. Data are held for:

* Financial information
* Information on managers and directors
* Credit scores and ratings
* Corporate structures
* Shareholders and subsidiaries

Data are identified by company registration number (CRN) on FAME, however CRNs are not directly comparable to RUs on the IDBR. Consolidated and unconsolidated accounts are available from FAME for enterprises only if they are part of a simple enterprise group on the IDBR, i.e. an enterprise group with just one enterprise.

**3.4 VAT**

VAT is tax on consumer expenditure and is collected by HM Revenue and Customs (HMRC). It is one of the main administration sources that the IDBR replies on to cover all businesses in the UK. If a business exceeds the VAT threshold (~£80k) then they must register for VAT. A business can also voluntarily register for VAT.

HMRC supply ONS with information in various frequencies (depending on the trader returns), however all traders are obliged to submit annual returns at the end of each financial year. Monthly VAT data is submitted to ONS each month; however this turnover is not updated for RUs or enterprises on the IDBR.

ONS receives different types of VAT registrations:

* VAT standards
* VAT Groups
* VAT divisions

**4 Data Mining and Linking**

Both external sources of administrative data have different reporting structures. FAME data is produced via CRN and VAT data is reported via VAT unit.

**Figure 4-1
Example enterprise structure using FAME and VAT**



For the majority of enterprises on the FAME database, one CRN per enterprise reports a consolidated figure on behalf of all other CRNs. For VAT however, it is possible that more than one VAT unit reports a consolidated figure. For these reasons, CRN and VAT unit are not directly comparable, but when aggregated at the overall enterprise level, they should in theory be equal.

In total, there are approximately 1,200 enterprises on the IDBR consisting of more than one RU. Data for around 350 of these enterprises was obtained using FAME data and around 650 enterprises using VAT data.

Having produced these two datasets, enterprises were linked to the ABS to obtain data for each RU. Currently the ABS is not designed to sample all of these RUs and so data was only available for a small number. Of the enterprises identified with ABS turnover for each RU, enterprises were only used in the analysis if the sum of RU turnover was larger than the enterprise turnover reported on FAME or VAT. This would indicate that intra-flows are present within the enterprise and should be removed.

The number of enterprises dramatically reduced, leaving a much smaller number of enterprises to use when modelling.

**5 Regression Modelling**

Multiple linear regression was used to predict the observed values of the dependent variable (consolidated turnover obtained using FAME and VAT data) using the linear function of two or more variables (explanatory variables from IDBR and ABS). The multiple regression model is as follows:

$$Y\_{i}=β\_{0}+ β\_{1}X\_{1i}+β\_{2}X\_{2i}+ ………+β\_{P}X\_{Pi}+ε\_{i}$$

where P is the number of explanatory variables

$β\_{0}$ is a regression coefficient
$β\_{P}$ is the regression coefficient for variable p
$ε\_{i}$ is the error term for trial i
y is the dependent variable
$X\_{1}… X\_{P}$ are the explanatory variables

The explanatory variables used in the modelling were obtained from the ABS and the IDBR:

* Enterprise turnover from ABS (sum of RU turnover within an enterprise)
* Enterprise employment (sum of RU employment within an enterprise)
* Number of RUs within an enterprise
* Number of LUs within an enterprise
* Number of VAT units within an enterprise
* Number of PAYE units within an enterprise

The most successful model was produced used FAME turnover as the dependent variable. All model assumptions were met and the model had good predictive power. Similar results were obtained using VAT data; however some of the assumptions underpinning the linear regression model were violated.

Three variables were statistically significant in the regression model: the sum of RU turnover within an enterprise (obtained from the ABS), the number of PAYE units within an enterprise and the number of VAT units within an enterprise.

**Table 5-1
Multiple regression coefficients obtained using FAME data**



The sum of RU turnover within an enterprise has a coefficient of 0.7, indicating that approximately 30% of RU level turnover is due to intra-flows within the enterprise. In addition to this, for businesses with a large number of PAYE or VAT units, the consolidated figure is much smaller. Figure 5-1 shows a scatter plot of actual FAME turnover and predicted turnover using the model above.

**Figure 5-1
Scatter plot of predicted turnover and actual FAME turnover**



The scatter plot shows that for the majority of enterprises, turnover has been predicted successfully using the model. Although this seems like a plausible model, it has however been based on a very small number of enterprises with respect to the total amount of enterprises required.

**5 Conclusions and Recommendations**

The findings from this feasibility study indicate that modelling enterprise level turnover using administrative data is a viable approach, however all models have been based on a small number of enterprises. The analysis has also left unresolved potential problems that could be faced when putting these models into practice.

**Recommendation 1** - All RUs in a multiple-RU enterprise should be sampled in the ABS.
As mentioned in Section 4, not all RUs are currently sampled in the ABS. Some RUs with large employment are picked up automatically but some smaller RUs are currently only picked up by chance. Changes have been made to the ABS sample to ensure that these RUs are now included.

**Recommendation 2** – Further work to decide which enterprises to model.
As mentioned in Section 4, enterprises were only modelled if the sum of the RU turnover was larger than the consolidated turnover reported in the FAME or VAT data. Some enterprises do therefore not require modelling because their turnover does not include any intra-flows. It is important that these enterprises can be identified and their turnover not reduced.

**Recommendation 3** – Investigations into VAT data.
The models produced using VAT data gave very similar results to FAME data, however all models violated the assumptions underpinning the linear regression model. Large differences were also found between ABS turnover and VAT turnover, indicating that more work should be carried out to investigate the elements and construction of VAT turnover.

1. Gary Brown, ONS, gary.brown@ons.gov.uk; Jonathan Digby-North, ONS, jonathan.digby-north@ons.gov.uk; Megan Pope, ONS, megan.pope@ons.gov.uk [↑](#footnote-ref-1)