An assessment of the feasibility of editing and imputing administrative tax return data to provide a substitute for survey data

Prepared by Charlotte Gaughan, Office for National Statistics, UK

I. Introduction

The Office for National Statistics (ONS) and the National Tax Office (HMRC) currently duplicate the collection of key economic variables such as companies’ turnover and sales revenue. The aim of the feasibility study is to assess whether the editing and imputation of tax return data results in sufficient quality data to serve as a substitute for survey data in national statistics.

Previous work has been conducted by the ONS on the methods of cleaning administrative Value Added Tax (VAT) data (Lewis and Finselbach, 2012). This paper discusses the limitations of using tax data, the methods employed to enable the data to be compared with current survey estimates and the techniques utilised to edit and impute the data. Imperative to the success of this work is an accurate comparison between the final HMRC data and our current survey estimates to assess quality. The initial study focuses on turnover for the calendar year 2012. Turnover was chosen because of its importance to ONS, and 2012 for quality reasons (e.g. the data can no longer be amended).

Other variables of interest such as sales revenue should be investigated before conclusions are drawn across the dataset.

II. Limitations

A number of limitations to this study reduce the reliability of the analysis and our confidence in the results; tax data whilst compulsory to submit is not compulsory to tag via an online system. It is only data which are tagged that are available to use, however data are neither consistently nor reliably tagged. Variables which would otherwise be useful for our analyses are poorly populated; for example ‘number of employees’ has a tag rate of around 1%. This has meant we have utilised current ONS data to substitute employment where necessary. Until legislation binds a company in to tagging (and not purely submitting) all variables the full benefits of the administrative data will not be realised. Legal issues surrounding the tax data also limits our access to the data which is stored in London, this in turn limits the scope of our study. Furthermore we are unable to query suspicious data with businesses resulting in an overreliance on our estimations. The nature of the data i.e. data submitted for tax purposes may also impact on the accuracy of the data compared to survey data, however we envisage that our analysis will shed sufficient light on this
potential drawback. Whilst the limitations discussed above potentially reduce the benefit of utilising administrative data this must be weighed against the cost savings, the reduction in burden on businesses and the increased coverage.

III. Data linking

In order to compare the tax data with survey data, the data needed to be transformed into a comparable format. Firstly, companies submit tax data according to their own accounting period, to enable comparisons the company returns were assigned to a calendar year based on the mid-point of their period of accounts. The data were then summed by company reference number (CRN) and by calendar year. This ensures that each row uniquely represents a particular company-year combination and ensures the data is a suitable format to link.

The data are submitted using a company reference number which only in the case of a single entity business correlates to the reporting unit (RU) data which ONS produces. In order to compare the data to ONS survey data, the HMRC data needed to be translated to RU level. There was no direct linking system between CRN and RU, however we were able to combat this by summing at CRN at enterprise (ENT) level and then apportioning out to RU level based on employment (taken from ONS RU level data) as shown in the diagram below:

Diagram 1

IV. Data Cleaning

For the year 2012 around 0.66% of returns had all variables missing. For turnover this figure was around 8% which meant we could confidently proceed with our analysis. Currently very large responses are going on to the system unchecked, in 2012, 99.74% of the total turnover in the HMRC data can be attributed to the company with the largest reported turnover. A variety of thousand pound style errors appear on the data set; in June 2015, HMRC identified around 350,000 scaling errors (i.e. £1000 style errors) for turnover, which we estimate is similar for 2012. Our analysis was restricted by the fact that data from previous quarters have not been edited thus we were unable to perform
previous period validation. However, ONS data could be utilised at RU level which enabled scaling error checks to be conducted. The following formula was used and adjusted to account for other scaling errors.

\[ 650 < \frac{\text{TaxTurnover}}{\text{ONS Turnover}} < 1350 \]

The thresholds of 650 and 1350 were based on those used with ONS survey data. In the past these values have proved to be most effective at identifying thousand pound errors without risking too many false hits.

A number of issues presented themselves unique to the tax data; the tagging system utilised meant that some data are submitted by companies that is not tagged. The untagged data are inaccessible to us. Incorrect tags have also been used (e.g. sales tagged as purchases). This problem was difficult to circumvent, and requires a change in legislation to force companies to tag all variables. Companies are permitted to return figures in currencies other than GBP, in 2012 0.4% of the companies did so. As we were unable to infer the exchange rate these were removed from the dataset. Negative values were returned which were invalid for example 0.02% returned negative ‘employees’ and ‘other income’ was 6.57% negative. Negative values were also removed from the dataset.

For data returned in currencies other than GBP and negative values, we imputed reasonable values in their place. In the next phase of the project we plan to investigate methods for identifying incorrect tags.

V. Imputation methods

An imputation study was required i) to provide a complete dataset which would allow comparisons to ONS survey and ii) to ascertain whether the missing data in the HMRC dataset could be accurately imputed. The dataset needed to be complete at the RU level; missingness at the CRN level did not in itself present a problem.

Referring to diagram 1 above, three scenarios presented themselves:

1) **All** CRN Turnovers are present at the ENT level; the CRN Turnover values are summed to create ENT turnover. The ENT value is then apportioned by RU employment to create RU Turnover.

2) **Some** CRN Turnovers are present at the ENT level; the missing CRN Turnover is imputed using a median imputation and then summed with the CRN Turnovers which are present to create ENT turnover. The ENT value is then apportioned by RU employment to create RU Turnover.

3) **None** of the CRN Turnovers are present; Turnover is imputed directly at the RU level.

In scenario 1 we were able to rely purely on turnover data from HMRC however for scenarios 2 and 3 data needed to be imputed.

*Creating strata*

In scenario 2, CRN level strata needed to be created. Given the lack of populated variables on the HMRC data set the strata were created by summing the number of CRN’s per ENT level to give an estimation of the size of the business. This was combined with a 2 digit SIC at ENT level to account for industry.
The frequency of CRN’s per ENT level was analysed and the size groups were defined as follows:

A = 1657049      1 CRN per ENT level
B = 56339          2 and 3 CRNs per ENT level
C = 19215        Between 4 and 15 inclusive CRNs per ENT level
D = 19624        16 or more CRNs per ENT level

In scenario 3, imputation could occur by creating strata based on employment and SIC; employment level data were available via the ONS business register data at RU level. The following methods were tested:

**Median imputation**
Within each strata (class) the median value was imputed

**Trim Mean imputation**
Within each strata (class) the mean value was imputed, with extreme values being trimmed before taking the mean to reduce the influence of outlying values. Two types of trimming were tested, removing the top and bottom 5%, and top and bottom 10% of values.

**Ratio of means imputation**
Auxiliary data taken from ONS business register data

\[ y_i^* = R \times x_i \]

\[ R = \frac{\sum_{i}^{n} y_i}{\sum_{i}^{n} x_i} = \frac{\sum_{i}^{n} y_i}{\sum_{i}^{n} x_i} \]

\[ R \] is the ratio
\[ x_i \] is the ONS turnover value
\[ y_i \] is the HMRC turnover value
\[ y_i^* \] is the imputed turnover
Class is defined as strata above

The methods tested for the 3 scenarios above were compared to test for the accuracy of the imputation. In each of the above scenarios, a simulation study was conducted, whereby a simple random sample of 10% was taken from the complete data set and imputed with each of the methods above. The methods above were tested to estimate which method introduces the least bias using the following formula:

\[ \frac{\sum_{i}^{n} (y_i^* - y_i)}{\sum_{i}^{n} y_i} \]
Secondly a measure for accuracy was tested to establish whether the methods were accurate in predicting turnover:

\[ \sum_{\text{class}} (y_i - y_i)^2 \sum_{\text{class}} y_i \]

The results are currently being finalised and will be published in a subsequent paper.

VI. Summary

The first stage of this study involved understanding the tax data, the systems used to compute the data and the caveats of the data. It was decided to conduct the study based on turnover for 2012. Various limitations reduced our confidence in the data, for example the tagging system employed engendered errors in tagging and poorly populated variables. In order to proceed with the analysis the data were linked with existing ONS data; the turnover data were summed by CRN at enterprise (ENT) level and then apportioned out to RU level based on employment taken from ONS data. The data were cleaned and a variety of errors such as scaling and typing errors were either edited, removed or imputed. Three scenarios occurred depending on the missingness of CRN level turnover - each of these was dealt with via a different imputation method. An imputation study was conducted at RU level imputed data to test for the bias introduced by different imputation methods and to test for the accuracy of the imputation method. The results are currently being finalised, and should be published by winter 2015. The study also has the scope to evaluate other variables intended for use such as sales revenue and purchases. This is highly recommended before the data are used for national statistics.

VI. References

Lewis D and Finselbach H (2012), Improving the efficiency of editing for ONS business surveys, Conference of European Statisticians, work session on statistical data editing.