Methods for estimating Structural Business Statistics variables not available from administrative sources

Ria Sanderson[1]

[1]on behalf of Work Package 3 of ESSnet on Admin Data,
email: ria.sanderson@ons.gov.uk

Abstract

We present results from Work Package 3 (WP3) of the ESSnet on the use of administrative and accounts data for business statistics. The aim of WP3 is to recommend estimation methods for variables not available from administrative data sources. This work is in part motivated by the common circumstance of statistical offices desiring to, or needing to, replace survey data with administrative sources. Suitable administrative data are not always available for all variables of interest, so the focus of WP3 is on recommending estimation methods in the case where administrative data sources cannot directly replace survey data. We report the results of the investigations carried out within WP3 on a number of Structural Business Statistics (SBS), and on the Short Term Statistic (STS) "New orders". We describe the evaluation criteria applied to the methods, and identify the key methods of use to statistical offices, including the administrative data requirements of the proposed methods. We present our recommendations of estimation methods for a first wave of variables, which includes the STS variable "New orders" and the SBS variables "Payments of agency workers", "Purchases of goods for resale in the same condition as received", "Number of employees in Full Time Equivalent" and "Changes in stocks of goods".

1. Introduction

The European Statistical System network (ESSnet) on the uses of administrative and accounts data for business statistics began in 2009 and is part of Eurostat's Modernisation of European Enterprise and Trade Statistics (MEETS) programme. The ESSnet is a collaboration between the National Statistical Institutes (NSIs) of European Member States to develop best practice recommendations for the uses of administrative data. The ESSnet consists of nine work packages (WP), which address the following:

WP0 – Management and administration
WP1 – Uses of administrative data
WP2 – Checklist for administrative data
WP3 – Methods of estimation for variables
WP4 – Timeliness of administrative data
WP6 – Development of quality indicators
WP7 – Statistics and accounting standards
WP8 – Creation of an Information Centre (http://essnet.admindata.eu/)
WP9 – Training and exchange of best practice

Work Package 3 (WP3) of the ESSnet on Admin Data focuses on estimation methods for variables that are not directly available from administrative (admin) data. Admin data sources do not include all the variables required by statistical regulations, and the aim of the WP is to develop estimation methods for these variables. WP3 began in Autumn 2009, and until June 2010 was a collaboration between the NSIs of the Netherlands, Lithuania, Italy, Germany and the United Kingdom. From July 2011 onwards, WP3 will be a collaboration between the NSIs of the Netherlands, Lithuania and the UK. This paper discusses the results of WP3 to date. We detail the methods of identification of the first set of variables that were studied until June 2010, and the estimation methods and results from this work. We also briefly introduce the second set of variables being considered, and discuss the future work.

2. Identification of the first set of variables

Variables of interest were identified from Structural Business Statistics (Commission Regulation (EC) No 2700/98) and Short Term Statistics regulations (Commission Regulation (EC) No 1165/98). The availability of admin data sources for these variables was determined

in each of the WP3 member countries, and variables were chosen for study where admin sources were not directly available. Four SBS variables and one STS variable were studied, and each of the five countries worked on a separate variable with support from a second country. The variables studied were: New Orders (Netherlands and Lithuania); Change in Stocks (Italy and UK); Purchases of goods and services for resale in the same condition as received (UK and Italy); Payments for Agency Workers (Lithuania and Germany); Employees in Full Time Equivalent (Germany and Lithuania). Generally, WP3 countries currently collect the variables of interest via a survey. The basic scenario investigated by WP3 was that these variables would no longer be collected via a survey either for all enterprises, or for a defined portion of enterprises in the target population. The estimation methods were assessed and evaluated for each variable, and a summary of the analysis undertaken is provided below.

3. Methods applied to the first set of variables

3.1 New Orders

The growth in new manufacturing goods (New Orders) is the only STS variable considered by WP3. The situation investigated is one where a survey still collects New Orders from the largest enterprises, but an estimate is made for the small and medium enterprises (SMEs). Four different scenarios were tested, and are described below:

1. The year-on-year growth in New Orders for SMEs was estimated using the growth seen in the large businesses in the same industry.

2. The year-on-year growth for SMEs was estimated using the year-on-year growth in VAT turnover in the same industry.

3. The year-on-year growth in New Orders was determined using qualitative information from a separate business cycle survey.

4. The year-on-year growth in New Orders was calculated using the year-on-year growth in modelled VAT turnover, determined by multiplying VAT turnover by the ratio between manufacturing turnover and total turnover calculated from an annual business survey.

The methods were tested using VAT and survey data from 2004 to 2007, where comparisons could be made between current survey-based estimates of New Orders for SMEs and the proposed methods listed above. Evaluation of the methods consisted of calculating the mean absolute difference and the correlation coefficients between the growths seen in the survey data, and the growths calculated using each method, at the 2 digit NACE level. Method 4 proved to be the most successful, with high correlations and low mean absolute differences (below 0.4% for all but four of the 2 digit NACE industries). We refer the reader to van der Holst (2011) for more information on this investigation.

3.2 Change in Stocks

Estimation methods were evaluated for the variable “Change in stocks” and its component variables “Changes in stocks of finished products and work in progress” (CSFP) and “Changes in stocks of raw materials and for resale” (CSRM). The following relationship holds between these variables: CS = CSFP – CSRM. Two scenarios were investigated for the estimation of changes in stocks. In the first scenario, administrative data is available for CS, but not for the components. In practice, CS and its components are available from financial statements for only part of the survey population. In the second, neither CS nor its components are available.

The first scenario was treated as partial non-response, and missing values were imputed using either Nearest Neighbour Donor Imputation (Statistics Canada, 1998), median

imputation or mean imputation. Additionally, robust linear regression modelling was used to model CSFP, with the relationship above being used to derive CSRM. Due to the large proportion of zeros in CSFP, a logistic regression model was used to model the probability of a business having a zero CSFP. The auxiliary variables used in the model were industry, number of employees, CS, turnover and purchases of goods and services for resale. The methods were evaluated by performing 200 iterations of a Monte Carlo simulation study, where non-response was randomly generated, and the methods were used to impute or predict a response. The relative difference between the estimated CSFP based on the original data, and the CSFP for each iteration was calculated, and its mean value (the mean relative estimation error) was used to compare the methods. Robust regression and median imputation were the best-performing methods, although median imputation works consistently well across all industries whereas some industries showed unsatisfactory results from the robust regression.

In the second scenario, no administrative data is available for any of the change in stocks variables. In this case, Nearest Neighbour Donor Imputation was used for the three variables using enterprises where all these variables were available as donors. Suitable donors were identified based on their industry (3 digit NACE), legal form, number of employees, turnover and purchases of goods and services for resale in the same condition as received. Robust regression was also used to predict CS, followed by an apportionment of the components using the methods tested under scenario 1. The methods were evaluated by artificially generating non-response for CS, CSFP and CSRM in data where these variables are available. The methods were then tested by calculating the relative difference between the original data and the predicted responses. Nearest Neighbour Donor Imputation performed well for most 2 and 3 digits NACE industries, but the robust regression modelling did not give satisfactory results. It is recommended that these methods be viewed as a starting point for the scenario where CS, CSFP and CSRM are all unavailable, as the methods are less well-developed than those used under scenario 1. We refer the reader to Luzi et al (2011) for more information on the estimation of the change in stocks variable.

3.3 Purchases of goods and services for resale in the same condition as received

The first scenario investigated was to assume the survey collecting purchases of goods and services for resale was discontinued. Unit level linear regression models were used to predict the value of purchases for those units where admin data were available. Imputation techniques (median and the trimmed mean) were then used in imputation classes to provide predicted values for the remainder of the survey population. Comparing the original survey-based estimate with the summed predictions for the years 2005-2006 showed this method to perform very poorly for the whole population, although for the matched population, the total from the predictions was much closer to the survey-based estimate (within 2-8%).

In the second scenario, the introduction of cut-off sampling to the SBS survey was investigated. The cut-off was set based on employment, meaning that the respondent burden on the smallest businesses was reduced. As admin data were not available for the whole survey population, comparisons were made for those units that could be matched to administrative sources. Three methods were tested to estimate the contribution to total purchases of goods and services for resale from the non-sampled businesses:

- A simple ratio adjustment was applied to the estimated total of purchases of goods and services for the employment size-band immediately above the cut-off. The ratio adjustment took the form of the total of the auxiliary variable below the cut-off, divided by an estimate of the total of the auxiliary variable above the cut-off. Both VAT turnover and company accounts variables were tested as possible auxiliary variables.
- A linear regression model was fitted at the unit level to predict purchases of goods and services for resale. Variables from the business register and VAT turnover were tested in this model.

- Generalised calibration estimation was investigated to try to reduce the bias introduced by cut-off sampling (Haziza et al 2010). This requires two auxiliary variables, one that is well-correlated with the variable of interest, and one that describes the probability of being above the cut-off.

The simple ratio method was found to perform the best over the years 2004-2008. The relative percentage difference between the current survey estimates and the estimates with the simple ratio adjustment was found to be small (approximately 2% or lower). We refer the reader to Sanderson et al (2011) for more information.

4.4 Payments for agency workers

Initial investigations found that there are no administrative data sources that show a strong correlation with payments for agency workers. Instead, it was decided that the income of temporary employment agencies may be able to serve as a proxy for payments for agency workers. Two methods were considered to see whether the income of temporary employment agencies could be used to estimate payments for agency workers:

- Using income of temporary employment agencies from their profit and loss accounts.
- Using income of temporary employment agencies from the annual services survey. This is an approach being investigated in Germany.

The first method aimed to estimate the income of temporary employment agencies using their profit and loss accounts. As payments for agency workers should only include domestic workers, it is crucial to remove the non-domestic part of the income. However, there was no administrative data describing the split into domestic and non-domestic income, so this method did not give reliable results.

The second method aimed to estimate the income of temporary employment agencies using the annual services survey. This survey asks separate questions for domestic and non-domestic income. The success of this method was tested by comparing current survey-based estimates of payments for agency workers with estimates of the domestic income of temporary employment agencies from the annual services survey in the years 2008 and 2009. The estimate of payments for agency workers exceeds the domestic income estimate by approximately 30%. This difference was attributed to sampling variation, measurement error and the fact that enterprises can employ workers from foreign employment agencies, which are not covered by the annual services survey. For more information, we refer the reader to Kavaliauskiene (2011).

4.5 Employees in Full Time Equivalent (FTE)

FTE converts the number of employees into full time equivalent, where one full time employee counts as one FTE. Any employee working for fewer than the standard number of working hours per week, or for fewer than the number of standard weeks in each year, needs to be converted to FTE. FTE is usually collected via the SBS survey, and generally there is no administrative source that contains the FTE variable although related variables are available. In Germany, the number of part time and full time employees is available for the financial services sector, and the number of hours worked is available from both survey and administrative sources. Data is also available on the number of hours paid. No one source on its own is sufficient to estimate FTE, due to differences in the survey periodicity, the quality of industrial classifications, and the level of aggregation at which the data are available. However, the estimation method proposed aims to bring these sources together.

The method estimates FTE using the number of full time and part time employees from the administrative data source, where the number of part time employees is multiplied by a conversion factor to convert it to FTE as follows:

$$\text{FTE} = \text{Full time employees} + \alpha \cdot \text{Part time employees}$$

$$\text{where } \alpha = \frac{\text{Hours of part time employees}}{\text{Hours of full time employees}}$$

The conversion factor was calculated separately in each industry. The method was evaluated by comparing estimates of FTE from the current SBS survey with estimates from this new method using combinations of the available administrative data sources to determine the conversion factor. Using hours paid data and separating the conversion factor for part-time employees into conversion factors for mini-jobbers, trainees and other part-time employees found a difference of 1.5%, suggesting this method produces estimates of sufficient quality to use these administrative data sources in the financial services sector. In conclusion, FTE can be estimated from data on the number of part time and the number of full time employees, with a conversion factor derived from the number of hours paid or the number of hours worked. Where these data are available from administrative sources or other surveys, it should be possible to produce estimates without the SBS survey. We refer the reader to Redling (2011) for more information.

5. Conclusions and future work

WP3 has identified useful estimation techniques for all the variables studied in the first stage of the analysis. It has proved important to understand the nature of the available data before using any of the estimation methods. In all cases, we recommend to Member States that the effects of the estimation methods are thoroughly tested, and the WP3 results should act as a guide in this regard. Our results have shown that it may be possible to stop collecting these variables for a subset of the population. During the first stage of our analysis, we have identified the types of data required for different methods, to ensure that the methods are applicable to other countries. From June 2011, WP3 has continued as a collaboration between the Netherlands, Lithuania and the UK. WP3 will explore further variables in the period 2011-2013, beginning with the SBS variable “gross investment in tangible goods” and its associated components, and the variable “sales of tangible investment goods”.

References

Commission Regulation (EC) No 2700/98, 17 December 1998, available from: http://epp.eurostat.ec.europa.eu/portal/page/portal/european_business/documents/REGULATION%20DEFINITIONS_CONSOLIDATED%20VERSION.PDF

Commission Regulation (EC) No 1165/98, 19 May 1998, available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1998:162:0001:0015:EN:PDF

Kavaliauskiene D., 2011, Report on Payments for agency workers, deliverable 3.4 for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics, http://essnet.admindata.eu/Document/GetFile?objectId=4819

Luzi O., Seri G., De Giorgi V., Siesto G., 2011, Report on Change in stocks, deliverable 3.2 for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics, http://essnet.admindata.eu/Document/GetFile?objectId=4817

Redling B., 2011, Report on Full Time Equivalent, deliverable 3.5 for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics, http://essnet.admindata.eu/Document/GetFile?objectId=4820

Sanderson R., Elliott D., Lewis D., Jones T., 2011, Report on Purchases of goods and services for resale in the same condition as received, deliverable 3.3 for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics, http://essnet.admindata.eu/Document/GetFile?objectId=4818

Statistics Canada, 1998, “Functional description of the Generalized Edit and Imputation System”, Statistics Canada technical report.

Van der Holst R., 2011, Report on Growth in New Manufacturing Goods, deliverable 3.1 for Work Package 3 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics, http://essnet.admindata.eu/Document/GetFile?objectId=4816