WP. 16
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (iii): Editing and imputation of administrative and census data

USING TAX DATA TO ASSIST WITH EDITING AND IMPUTATION

Supporting Paper

Submitted by the Office for National Statistics, UK[1]

I. INTRODUCTION

1. The UK Office for National Statistics (ONS) currently receives annualised turnover data from Value Added Tax (VAT) returns held by Her Majesty's Revenue and Customs (HMRC). The data arrive regularly - daily, monthly and quarterly, and form a large part of the Inter-Departmental Business Register (IDBR) annual turnover variable. This register turnover is widely used in ONS business surveys as an auxiliary variable for ratio estimation, but is also used to assist with editing and imputation.

2. Around 90% of businesses in the VAT data set report to HMRC on a quarterly basis, with around 10% reporting monthly and a very small number annually. Since the majority of the businesses report quarterly, it may be possible to make greater use of the data than the annualised register turnover value which is currently used. This paper reports on a study investigating whether the VAT turnover data could be used more effectively to help with editing and imputation on short term business surveys. This is part of a wider ONS project looking at ways to make better use of VAT data. The work involved modelling the data in an attempt to make it as similar to monthly Retail Sales Inquiry (RSI) turnover as possible and testing the potential of the data as a predictor for editing or imputation on this survey in comparison to existing predictors.

II. DIFFERENCES BETWEEN VAT AND RSI TURNOVER

3. There are two key differences between VAT and RSI turnover - definition and periodicity. When businesses return VAT turnover data to HMRC, they are asked to include capital sales. The definition of RSI turnover excludes capital sales. Most businesses reporting turnover for VAT returns send their data quarterly. The RSI is a monthly survey, which collects retail turnover on an approximately monthly basis and converts this to an average weekly turnover figure for the month. In order to effectively use the VAT turnover data to assist with editing or imputation in the RSI it is necessary to model the data. There were two stages of modelling to convert the VAT turnover to be similar to RSI turnover. Firstly, the VAT data were modelled against RSI data in an attempt to reduce the effect of the difference in definition. Secondly, the resulting modelled VAT data were transformed into monthly data using time series techniques. The two processes are described below.

[1] Prepared by Daniel Lewis & Alaa Al-Hamad (daniel.lewis@ons.gov.uk , alaa.al-hamad@ons.gov.uk)

III. MODELLING TO TAKE ACCOUNT OF DIFFERENCE IN DEFINITION

4. The VAT turnover data were matched with RSI data in order to model the difference between the two sources. Data were used from both sources for the period January 2005 to June 2007. The VAT datasets from HMRC contained returns relating to different reference periods and on different reporting cycles. Most returns are quarterly, but there are three separate quarterly reporting periods, starting at January, February and March respectively (and continuing through the year). Each business responding will use one of these quarterly reporting periods. In addition to this, it is not uncommon for businesses to return late, so that returns in any particular dataset can refer to a variety of reference periods.

5. The first step was to separate and re-combine the VAT data to create a file for each month containing just the quarterly data relating to that reporting period. The data were then matched to RSI returns for the corresponding periods. This was not straightforward, since the reporting unit reference on the RSI datasets does not have a one to one match with the VAT reference. However, it was possible to match the majority of data using the reference for the enterprise (which is made up of one or more reporting units), which is already matched to both identifiers. Where there were many VAT references to a single enterprise reference, it was possible to sum the turnover across the VAT references. However, where there were many enterprise references to a single VAT reference (or many to many) there was no way to accurately share the VAT turnover. The same was true when matching the newly created enterprise level VAT dataset to the reporting unit level RSI dataset. Where it was not possible to match references the businesses were eliminated from the analysis. On average there were around 1000 businesses each month with matched RSI and VAT turnover and these were used to model the data. Note that around 4000 businesses respond to RSI each month, so only a quarter could be properly matched to the VAT data.

6. After matching VAT data to RSI data, the businesses were separated according to reporting frequency. Monthly and quarterly VAT responders were modelled separately, with the quarterly VAT data being modelled with quarterly RSI data (derived by adding up the three relevant months of RSI data) and the monthly VAT data being modelled with monthly RSI data. The small number of businesses who return VAT data annually were not included in the study.

7. Note that the RSI turnover is an average weekly figure, calculated by assuming a 4, 4, 5 pattern for weeks in each month. That is, January and February have 4 weeks and March has 5 weeks. The pattern then continues throughout the year. Occasionally there is a 5 week January to correct for the fact that a year is not made up of an exact number of weeks. From 2005 to 2007, all of the Januarys had 4 weeks. The RSI average weekly turnover data were converted to monthly figures by multiplying by 4 or 5, depending on the month.

8. A relationship between VAT turnover and RSI turnover was established using regression modelling. Initially models were tested including capital and non-capital expenditure data from the Annual Business Inquiry (ABI) in the hope that this could help explain the difference between VAT turnover, which includes capital sales, and RSI turnover, which does not. However, this turned out not to be helpful for the model and also limited the size of the dataset that could be used because of the additional need to match to ABI returns. A simple linear regression between VAT turnover and RSI turnover produced good results. This was improved by splitting the data into VAT turnover bands as follows:

Band 1: VAT turnover <= 99,999
Band 2: 100,000 <= VAT turnover <= 249,999
Band 3: 250,000 <= VAT turnover <= 999,999
Band 4: VAT turnover >= 1,000,000

9. Separate models were produced within each turnover band (and for monthly and quarterly data separately). The two variables were highly correlated and the models were all well fitted, with R^2 values around 0.98, highly significant P-values and well behaved residuals. In each case a model equation was produced of the form:

RSI turnover = β VAT turnover

There was no significant intercept term in any of the models. The estimated β parameters from each regression were then multiplied by the turnover responses in the VAT dataset (matching with the corresponding turnover size and periodicity) to give a modelled VAT turnover variable hopefully more similar to RSI turnover than the original VAT data.

IV. TIME SERIES MODELLING TO TAKE ACCOUNT OF DIFFERENCE IN PERIODICITY

10. The regression modelling described above should help correct for some of the difference in definition between VAT and RSI turnover. However, to be useful for editing or imputing RSI data, the VAT turnover needs to be monthly rather than quarterly. The VAT data are also less timely than the RSI survey data. In order for the VAT data to be available quickly enough to use for editing and imputation they would need to be forecast forward two months. To attempt to solve both of these problems, we used time series modelling. For this part of the project, we had the help of Gary Brown, Kevin Moore and Tullio Buccellato of the ONS Time Series Analysis Branch.

11. Approximately 90% of the modelled VAT turnover data were quarterly. These were transformed into monthly data by using the cubic spline method. This involves fitting a third degree polynomial through the quarterly data points and reading off the value of the polynomial at the required monthly points. This is an approximate method and in practice there were a small number of businesses who were given a negative estimate of monthly turnover. There is nothing in the cubic spline method which prevents the fitted polynomial from going below the x-axis. The turnover for these businesses was set to missing.

12. Having created estimated monthly VAT turnover, the data were then forecast forward two months using the Holt Winters method. This method uses an estimate of the trend and the seasonal pattern of the data to predict forward the required two months for the VAT data to be useful. For comparison with the RSI average weekly turnover, these monthly figures were then divided by 4 or 5, to fit in with the 4 and 5 week months of RSI.

V. USING MODELLED VAT TURNOVER DATA TO ASSIST WITH EDITING AND IMPUTATION

13. If the modelling worked well, the resulting VAT turnover data could be a good predictor of RSI turnover. If this predictor is accurate, the modelled average weekly VAT turnover data could be used in editing and imputation. For editing, it is useful to have an expected value for a question answered by a business in order to compare the two figures and decide whether the returned value is suspicious. One particular application is in selective editing, where expected values are required to score returned values and decide whether they have enough impact on estimates to be worth re-contacting. For imputation, the VAT turnover data could be used directly to predict turnover for non-responders, or to assist in imputing RSI turnover based on the relationship between the two turnover variables.

14. In both cases we would effectively be using the VAT data to predict RSI turnover. To ascertain whether the VAT data is useful for editing and imputation, we need to compare the accuracy of these predictions against currently existing predictors of RSI turnover. At present, the main predictor is the returned turnover from the previous period. For businesses which didn't return turnover in the previous period, the annual IDBR turnover is used instead. Note that in this instance, the IDBR turnover is divided by 52 to give an approximate weekly figure and multiplied by 1000, since the IDBR figure is stored in thousands of pounds and the RSI collects turnover in pounds.

15. Four predictors of RSI turnover were examined. The VAT turnover was tested as a predictor in its own right and as an alternative to IDBR turnover for businesses which did not return turnover in the previous period. The predictors tested were:

1. VAT turnover if available, otherwise use IDBR turnover
2. Previous RSI turnover if available, otherwise prefer VAT turnover over IDBR turnover
3. Previous RSI turnover if available, otherwise use IDBR turnover
4. IDBR turnover

The accuracy of these predictors were analysed by estimating their relative error and relative absolute error in predicting the true RSI turnover responses.

VI. RESULTS OF TESTING THE ACCURACY OF VAT TURNOVER FOR PREDICTING RSI TURNOVER (AVERAGED OVER 30 MONTHS)

16. For the approximately 1000 businesses in each of the periods January 2005 to June 2007 which could be matched to VAT data, each of the predictors 1 to 4 was calculated and compared with the returned RSI turnover value. The accuracy of the predictors was estimated using the following indicators:

$$\text{Estimated relative error} = \frac{\sum_{i \in s} w_i (y_i - \hat{y}_i)}{\sum_{i \in s} w_i y} \times 100$$

$$\text{Estimated relative absolute error} = \frac{\sum_{i \in s} w_i \left| y_i - \hat{y}_i \right|}{\sum_{i \in s} w_i y} \times 100$$

where w_i is the estimation weight for unit i,

y_i is the edited average weekly turnover value from RSI,

$\hat{y}_i$ is the predicted average weekly turnover value.

All sums are over the full RSI sample, s.

17. Tables 1 and 2 show the mean, median, lower and upper quartiles of the prediction errors averaged over the 30 months of RSI survey data.

Table 1: Estimated relative errors of predictors (expressed as percentages)

Predictor	Mean	Q1	Median	Q3
1	-59.9	-68.6	-61.0	-47.1
2	-8.9	-8.9	-4.8	-1.5
3	-1.1	-0.5	0.7	3.0
4	4.6	0.8	3.0	5.0

Table 2: Estimated relative absolute errors of predictors (expressed as percentages)

Predictor	Mean	Q1	Median	Q3
1	70.7	58.7	68.9	81.4
2	22.3	14.7	18.2	14.0
3	15.4	10.4	11.2	14.0
4	32.4	28.5	30.6	35.8

18. The results clearly show that VAT turnover does not perform better than currently existing predictors. The best predictor is the RSI turnover value from the previous month when this is available (predictors 2 and 3). When there is no previous RSI value, it is more accurate to use the IDBR turnover (predictor 3) than to use the modelled VAT turnover (predictor 2). Directly using modelled VAT turnover (predictor 1) is also less accurate than directly using IDBR turnover (predictor 4). From the figures above it appears that the modelled VAT turnover generally overestimates RSI turnover. This may be due to the fact that VAT turnover includes capital sales, whereas RSI turnover does not. The modelling may not have entirely removed this difference.

19. Examining individual businesses (on average across the 30 months) there were 719 times when predictor 1 did better than predictor 4, compared to 3441 times when predictor 4 did better than predictor 1. That is, IDBR turnover was a better predictor than VAT turnover for over 82% of businesses. There were around 33 times each month when predictor 2 did better than predictor 3, compared to 4127 times when predictor 3 did better than predictor 2. That is, using IDBR turnover as a back up to the previous RSI value was better than using VAT turnover as the back up for 99% of businesses.

20. The annual IDBR turnover is mostly made up of annualised VAT data. This suggests that the advantages of using more timely VAT data are outweighed by the errors introduced by the modelling required to produce monthly VAT data quickly enough to be useful for editing and imputation.

VII. MONTHLY RESULTS TESTING THE ACCURACY OF VAT TURNOVER FOR PREDICTING RSI TURNOVER

21. The results above are averaged over 30 months. RSI turnover is very seasonal and so it might be expected that the accuracy of the different predictors varies depending on the month. To test whether there are particular months where VAT turnover performs better than other predictors, table 3 shows the estimated relative absolute errors by month (errors are averaged over years).

Table 3: Estimated relative absolute errors of predictors by month (expressed as percentages)

Month	Predictor 1	Predictor 2	Predictor 3	Predictor 4
January	122.4	75.0	57.2	31.2
February	76.8	20.3	14.0	27.6
March	46.0	14.6	11.4	29.4
April	62.0	18.0	12.3	30.6
May	69.1	20.0	8.9	31.1
June	54.5	19.1	9.8	30.8
July	73.5	17.0	11.1	31.6
August	70.4	18.2	11.6	32.1
September	57.6	13.9	10.1	32.2
October	82.9	18.1	11.8	32.4
November	87.1	19.5	15.2	36.3
December	70.1	27.6	25.2	47.3

22. For the 11 months February to December, the relative accuracy of the four predictors is the same. That is, predictor 3 is most accurate, followed by predictor 2, predictor 4 and predictor 1. In January, predictor 4 is most accurate, followed by predictor 3, predictor 2 and predictor 1. In general, the predictors that make use of the previous month's data (predictors 2 and 3) are most accurate. However, in January it is more accurate to use the IDBR turnover (predictor 4) to predict. This is presumably due to the larger turnover figures experienced every December. However, using VAT turnover in any way is always less accurate, no matter what the month.

VIII. CONCLUSION AND FUTURE WORK

23. This paper has described work analysing the potential of modelled VAT turnover to assist with editing and imputation in a monthly business survey, the Retail Sales Inquiry. The VAT turnover is defined differently from RSI turnover, is mostly quarterly, and is generally not timely enough to be used for editing and imputation in the RSI. Various modelling was undertaken in an attempt to overcome all of these problems. The resulting modelled VAT turnover variable was compared against currently existing predictors of RSI turnover to assess whether it could be useful for editing and imputation. VAT turnover proved to be a less accurate predictor of RSI turnover than the current predictors - previous period RSI turnover and IDBR turnover. It appears that the errors introduced by the various modelling required to be able to use VAT turnover for RSI editing and imputation outweigh the advantages of having a relatively timely administrative source.

24. Whilst the results of this particular study were not encouraging for using VAT data to assist with editing and imputation, it should be borne in mind that this was partly because the VAT data were mostly of a different periodicity to the RSI data and partly because RSI already makes some use of annualised VAT turnover data in editing and imputation. Work due to be completed later in 2009 will look at the potential benefits of using two VAT variables, turnover and expenditure, to assist with editing and imputation in the Annual Business Inquiry. Because it is easier to convert the VAT data to be annual and because at present no use is made of the VAT expenditure variable (which is correlated with some of the ABI variables), it is hoped that the results of this study will be more positive.