WP. 8
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (i): Editing of administrative and Census data

Evaluating the benefits of using VAT data to improve the efficiency of editing in a multivariate annual business survey

Invited Paper

Prepared by Daniel Lewis, Office for National Statistics (UK)[1]

I. Introduction

1. In common with many other National Statistical Institutes, the UK Office for National Statistics (ONS) is currently investigating ways to improve the efficiency of statistical outputs by making greater use of administrative data. Whilst this is often achieved by replacing survey data with administrative data, efficiency gains can also be made through innovative use of administrative data in editing. This paper describes work to investigate the possibility of using Value Added Tax (VAT) data to improve the efficiency of data editing for the multivariate UK Annual Business Survey (ABS).

2. The work involved creating predictors for key ABS variables based on VAT Turnover and Expenditure data, modelled with ABS survey data from previous periods. These predictors were then utilised in ABS editing in two ways. The first method involved using the predictors to try to improve the current traditional edit rules in the survey. The second method used the predictors as expected values in a selective editing application currently being investigated by ONS, based on the Swedish program Selekt. This work built on previous ONS studies described in Elliott (2009) and Lewis (2009). This paper briefly describes the modelling work that created the predictors, then describes the editing approaches tested, the methods used to evaluate them and a summary of the results.

II. Creation of predicted values

3. Predicted values were created for each of eight key ABS variables: Total purchases, Total taxes, Opening stocks, Closing stocks, Net capital expenditure (Net Capex), Total turnover, Employment costs and Gross value added (GVA). The predicted values were created by modelling Value Added Tax (VAT) data with previous ABS survey data for 2005 and 2006 data. The idea was to develop models explaining the relationship between the VAT and key ABS variables based on previous period data. The predicted values were then created by applying the coefficients resulting from those models to current period VAT data. The majority of the work was carried out using VAT Turnover, but the VAT Expenditure variable was also incorporated into the predictors towards the end of the modelling part of the work, after gaining temporary access to the data from the UK tax office.

4. The main modelling technique was to fit General Linear Models (GLM) to the VAT and ABS data. As well as VAT Turnover and Expenditure other potentially useful covariates, such as the industry, size and region of the business, were considered in the models. Because some of the key variables have a

[1] The author acknowledges the valuable work of Sarah Skinner, Ria Sanderson, Duncan Elliott and Allyson Powell of the Office for National Statistics in modelling the predicted values described in this paper.

large number of zeros, which can be problematic in the type of models tested, this study also investigated combining the GLM modelling with logistic regression. The logistic regression was used to model the probability of a business returning a zero value for a particular variable. These probabilities were combined with the predictions from the GLM model to provide an alternative set of predicted values.

5. The models developed were all tested using the usual range of model statistics and diagnostic checks. For the most promising models, the error of the resulting predicted values was also estimated, by comparison with the true ABS survey returns. The final model in each case was the one with the smallest prediction error. Data splitting was used to check that individual models were robust. This approach involved randomly selecting half the data, fitting the model, and then applying the model to the remaining half of the data. For more details on the modelling and evaluation see Lewis (2010).

6. The modelling work resulted in six predicted values for 2005 and 2006 for each of the eight key ABS survey variables. Two of the predicted values used VAT Turnover, two used VAT Expenditure and the final two used both VAT Turnover and VAT Expenditure. In each case, there was one predicted value derived only from the GLM modelling and another which also incorporated logistic regression to model the probability of a zero value in the variable being predicted. One further predicted value was created for each key variable, using the previous response for the business when available and the best performing of the VAT predictions otherwise. These seven predicted values were then used to try to improve the efficiency of ABS editing. Section III describes methods and results for using the predicted values to improve the current traditional micro editing rules used by the survey. Section IV describes use of the predicted values in a selective editing approach currently being developed for the survey.

III. Use of predicted values to improve traditional editing

A. Current ABS micro editing rules

7. The ABS survey is split into seven industrial sectors – Catering, Motor Trades, Production and Construction, Property, Retail, Services, and Wholesale. Each of these surveys uses a different set of edit rules (and some edit rules are only applied to certain types of business). Many of the rules focus on inconsistency between variables (for example where components do not sum to totals or where responses to two related variable are impossible). There are some rules which compare responses with previous or register values, but these only exist for a small number of the survey variables. Some of the eight key variables being investigated by this study currently have no edit rules of this type and one variable (Net Capex) has no edit rules at all.

8. In order to reduce the cost of ABS editing, a change was implemented in 2005 and 2006 to exclude some of the less important smallest responders from micro editing. In order to accurately reflect the effect of any changes to micro editing, the current rules tested in this study included the most recent threshold of this cut-off.

9. A separate project in ONS is currently investigating methods to improve ABS micro editing using a selective editing method. There are more details of this approach in section IV. The study initially focused on two of the industrial sectors covered by the survey – Catering, and Production and Construction. Bearing this in mind, it was decided to focus the VAT predicted value work on improving edit rules in the same two sectors.

B. Use of predicted values in traditional micro editing

10. The predicted values were used to create micro editing rules to be used in place of all of the rules currently used in the Catering, and Production and Construction sectors of the survey. As there were seven final predicted values generated for each key variable, this resulted in seven sets of alternative edit rules. In each case, the alternative edit rules involved comparing the returned value for the key variable with the predicted value. More formally, rules of the following type were implemented for each variable:

$$\text{If } \frac{|\text{returned value - predicted value}|}{\text{predicted value}} > x \text{ then fail}$$

The parameter *x* varied for each variable and in each of the two sectors. The value of *x* was fine tuned in each case so that a similar number of businesses failed micro editing as in the current ABS edit rules. This allowed for comparison of the accuracy of resulting estimates with the current rules. It would also be possible to test the efficiency of the new rules by fine tuning the value of *x* to give a similar level of accuracy and comparing the number of failures.

C. Evaluating the efficiency of micro editing rules based on VAT predictors

11. Having set the edit rules based on VAT predictors to fail a similar number of businesses to the current ABS micro editing rules, the accuracy of resulting estimates was estimated by calculating two bias measures. The bias measures are based on concepts described in Silva et al (2008). Each measure estimates the relative difference between the final version (post macro editing) of the survey estimate for each key variable and the estimate which would have resulted by only applying the micro editing rules. In effect, this is an estimate of the amount of error left after the micro editing stage. This error is easily estimated since ABS data are available in both unedited and edited form. The unedited data are, roughly speaking, the response values before any micro editing has been applied, and the edited data are the values after macro editing.

12. The first bias measure considers the edits of each key variable in isolation. For this local bias measure, a business is given the edited value for the variable if it fails a micro editing rule relating to that variable. If the business does not fail an edit rule relating to the variable, it is given the unedited value. The local bias is calculated as follows:

$$\text{Local bias for variable } j = \frac{\sum_s I_{ji} \left| y_{ji,unedit} - y_{ji,edit} \right|}{\sum_s y_{ji,edit}} \times 100$$

where *s* indicates that the sums are over the business sampled in the ABS sector, I_{ji} is an indicator variable equal to 0 if business *i* fails an edit rule relating to variable *j* and equal to 1 otherwise, $y_{ji,unedit}$ is the unedited value for business *i* and variable *j*, and $y_{ji,edit}$ is the edited value for business *i* and variable *j*.

13. The second bias measure considers the edits of all key variables together. For this global bias measure, a business is given the edited value for the variable if it has failed a micro editing rule relating to any of the key variables. The business is only given the unedited value if it passes all edit rules. The global bias is calculated as follows:

$$\text{Global bias for variable } j = \frac{\sum_s I_i \left| y_{ji,unedit} - y_{ji,edit} \right|}{\sum_s y_{ji,edit}} \times 100$$

Where I_i is an indicator variable equal to 0 if business *i* fails an edit rule for any of the key variables and equal to 1 otherwise. The global bias is particularly appropriate when it can be assumed that any failure in the ABS survey leads to all questions being confirmed with the respondent.

14. The key focus of this study is to investigate the possibility of using VAT data to improve the efficiency of ABS micro editing. Because the current ABS micro editing rules may not be optimal, it was decided to also compare with alternative edit rules based only on previous survey data. These rules took the same form as those implemented for the VAT predicted values. The only difference was that the predicted values were created using the previous survey response, where available, and the median of the

previous responses for similar types of business otherwise. The similar types of business were defined by the cross-classification of 2 digit Standard Industrial Classification (SIC) and employment size band.

D. Results of using predicted values in traditional micro editing

15. This section gives the local and global bias for each of the traditional micro editing approaches in the Catering, and Production and Construction sectors, averaged over the two years 2005 and 2006. Because of the large number of different editing strategies tested, only the best performing VAT prediction method is shown in each case. In the tables below, 'Current' refers to the current ABS micro editing rules, 'VAT' refers to edit rules based on the best VAT prediction for the variable, 'Previous' refers to the alternative edit rules based only on previous survey data, and 'VAT Previous' refers to edit rules based on the previous survey response when available and the best VAT prediction otherwise.

16. Previous editing studies in ONS have often attempted to reduce edit failures whilst keeping the relative global bias below 1%. This was difficult to achieve for most sectors, so instead the focus was on keeping the number of failures at the same level or slightly below that of the current micro editing rules and analysing the effect on the bias. Reduced bias with a similar number of failures indicates that efficiency improvements are possible. The magnitude of savings would be dependent on how much bias is acceptable to leave in estimates after the micro editing stage (bearing in mind that the estimated bias reported for the current editing rules is entirely removed by macro editing).

17. The tables below show the local and global bias in the Catering sector for each variable, keeping the number of edit failures around the same level as the current micro editing rules.

Table 1. Local bias (%) for the Catering sector

	Current	VAT	Previous	VAT Previous
Total purchases	4060	269	180	269
Total taxes	3964	4	4	4
Opening stocks	1014	37	49	48
Closing stocks	1547	977	882	977
Total turnover	3	3	3	3
Net Capex	3253	204	942	282
Employment costs	573	2	2	2
GVA	1279	79	104	80

Table 2. Global bias (%) for the Catering sector

	Current	VAT	Previous	VAT Previous
Total purchases	369	7	6	7
Total taxes	512	2	1	2
Opening stocks	257	6	4	5
Closing stocks	283	5	4	5
Total turnover	2	1	1	1
Net Capex	2	49	50	49
Employment costs	538	1	1	1
GVA	659	7	7	7

18. In the Catering sector it is possible to greatly improve on the current edit rules for every variable apart from Net Capex in terms of the accuracy of the resulting data (although the improvement is marginal for Total turnover). However, in this sector there is no advantage in using the VAT predictions, since using predictions based on only previous ABS data perform at least as well, and sometimes better than the VAT predictions.

19. Note that there are no current edit rules for Net Capex, but errors are detected from failures in related variables. This means it is impossible to produce new edit rules with a similar number of failures for this variable. The results for Net Capex reflect an attempt to introduce minimal editing and keep the bias low. The local bias for the current situation is very high. This is not surprising as none of the error is detected at the micro editing stage. However, it transpires that a large amount of the error in Net Capex is corrected when respondents are re-contacted after failing current edit rules for other variables. The result of this is that the global bias for the current micro editing is very low so that it was impossible to improve on the global bias using predictions only.

20. The tables below show the local and global bias in the Production and Construction sector for each variable, keeping the number of edit failures around the same level as the current micro editing rules.

Table 3. Local bias (%) for the Production and Construction sector

	Current	VAT	Previous	VAT Previous
Total purchases	867	12	16	13
Total taxes	659	661	222	661
Opening stocks	819	423	690	423
Closing stocks	828	467	687	466
Total turnover	3	5	17	5
Net Capex	1395	205	772	217
Employment costs	204	10	25	10
GVA	2054	14	13	16

Table 4. Global bias (%) for the Production and Construction sector

	Current	VAT	Previous	VAT Previous
Total purchases	53	3	7	3
Total taxes	149	2	6	2
Opening stocks	117	3	6	3
Closing stocks	132	3	7	3
Total turnover	0	1	10	1
Net Capex	6	39	44	45
Employment costs	148	2	20	2
GVA	86	9	8	8

21. It is also possible to greatly improve on the editing in the Production and Construction sector for all variables apart from Net Capex and Total turnover. The comments on Net Capex above apply equally for Production and Construction. In this sector, the VAT predictions offer a genuine advantage over previous ABS data. It is necessary to use the VAT predictions to get the greatest benefits for Total purchases, Total taxes, Opening stocks, Closing stocks and Employment costs.

22. In general the results above show that it is possible to greatly improve the efficiency of ABS micro editing by using the alternative edit rules based on comparing survey returns with VAT predicted values. For the Catering sector, an equivalent improvement is generally possible by using similar style edit rules based only on previous ABS responses. For the Production and Construction sectors it is necessary to use the VAT predicted values to get the best savings for most of the key variables.

23. The one key variable where the alternative edit rules perform poorly is Net Capex. There are currently no ABS micro editing rules in place for this variable, but failures to other existing micro editing rules appear to be quite effective at identifying errors in Net Capex. If the alternative edits were implemented in place of the existing ones, it would be necessary to explore which of the existing rules are effective at identifying errors in Net Capex and to make use of them in the new system.

IV. Use of predicted values in selective editing

A. Selective editing for ABS

24. The most common application of selective editing involves calculating a score for each business, calculated as the weighted relative absolute difference between the returned value for a particular variable and an expected value for that variable. The standardising factor is usually an estimate of the total or an estimate of the standard error in the domain that the business belongs to. These estimates generally come from the previous period of the survey. When a survey has more than one important variable, the scores are often combined in some way to produce a unit level score. Scores for each business are then compared with a pre-defined threshold. Any business with a score larger than the threshold fails selective editing and is sent for manual follow up. The thresholds are set using previous data and attempt to minimise the amount of failures whilst maintaining the accuracy of survey outputs.

25. This simple selective editing approach works well for surveys with small numbers of key variables and with reasonably good expected values available. The response from the same business in the previous period is often used as an effective expected value. ABS is a multivariate survey with many important variables. The design of the survey means that less than half the businesses in the sample for a particular year were also in the sample in the previous year. For these reasons, the selective editing method described above is not suitable for the survey. After investigating a number of options, it was decided that the method most likely to work for ABS is a selective editing method implemented in the Statistics Sweden Selekt program.

26. The Selekt method calculates a generic score function for each business, variable and domain grouping. These score functions are then successively aggregated to create a unit level score for each business. Thresholds are included for each level of aggregation (so that only scores above the threshold are included when summing scores). The unit level score is then compared to a final threshold, and any businesses with a score over the threshold fail the selective editing. The scores are made up of three main terms: $\text{Score} = \text{Suspicion} \times \text{Impact} \times \text{Importance}$.

27. The suspicion of a unit (for a particular variable) is calculated in two ways, using traditional edit rules and ‘test variables’. The suspicion is always between 0 and 1. Any unit failing a traditional edit rule is given a constant suspicion value, usually equal to 1. The test variables compare returned values with other, related variables. Selekt then calculates a suspicion between 0 and 1 based on a relative difference between the returned value and the other variables defined in the test variables.

28. The impact measures the potential impact on the domain estimate for the particular variable if the returned value for the business was an error. This is similar to the selective editing score described above and is calculated as the weighted absolute difference between the value reported by the business and an expected value for the variable.

29. The importance has two roles. Firstly it allows the possibility of giving extra weight to particular variables or domain groupings in the score. The second role is in standardising the score – giving the option to divide by an estimated total or standard error (or some function of those estimates) for the domain of interest. The importance is calculated as:

$$\text{Importance} = \frac{\text{weight for variable } j \times \text{weight for domain } d}{\text{estimated domain total or domain standard error}}$$

For more details of the Selekt method, see Norberg and Arvidson, 2008.

30. Work is currently underway to implement the Selekt method for ABS. At the time of this study, the method had been tested for two sectors of ABS – Catering, and Production and Construction. For both sectors, the testing showed that savings can be made using the Selekt method. For more details on the ABS Selekt work, see Skentelbery et al (2011).

B. Use of predicted values in Selekt

31. The aim of this part of the study was to see if the VAT predicted values can improve the efficiency of Selekt when used for selectively editing ABS. There are two parts of the Selekt score where these predicted values could be of use. In the impact part of the score, the predicted values could be used as expected values. For the suspicion part of the score, the predicted values could be used as test variables. In the current (test) implementation of Selekt, both of these parts of the score currently make use of the previous survey response when available and a median of previous responses for similar businesses otherwise. Similar businesses are defined in this case by 4 digit SIC.

32. The Selekt program requires the setting of a fairly large number of parameters, controlling different aspects of the score calculation. In order to make the comparison fair, it was decided to keep all parameters the same apart from the score thresholds, which were fine-tuned to give a similar number of failures for each implementation.

33. The results of the current implementation were compared with those using the VAT predicted values both in place of the previous responses, and as an alternative to the median of previous responses for those businesses that were not in the sample in the previous period.

34. The results showed no significant difference between whether the VAT predicted values were used for just the expected value in the impact part of the score, or if they were also used as test variables in the suspicion part of the score. In the results produced below, they were used for both of these parts of the score.

C. Results of using predicted values in Selekt

35. Selekt uses two years of data to fine tune parameters and thresholds and then implements the results on a third year. Due to data availability, this meant it was only possible to calculate Selekt results for one year – 2006.

36. ONS have been evaluating Selekt by calculating savings (percentage reduction in number of businesses failing editing) and the average of the relative global biases of five key variables (GVA, Total turnover, Net Capex, Employment costs and Total purchases). Tables 5, 6 and 7 show respectively the results for the original test implementation of Selekt (not using any VAT data), an implementation where VAT predicted values completely replace previous data, and an implementation where VAT predicted values are only used for businesses that did not have a previous value.

Table 5. Selekt results (all in %) in original test implementation

Sector	Savings	Average global bias
Catering	19	0.4
Production and Construction	12	0.4

Table 6. Selekt results (all in %) using VAT predicted values in place of previous data

Sector	Savings	Average global bias
Catering	19	1.0
Production and Construction	12	0.4

Table 7. Selekt results (all in %) using previous returned values when available and VAT predicted values otherwise

Sector	Savings	Average global bias
Catering	19	0.6
Production and Construction	12	0.4

37. The results above show that the VAT predicted values do not give any improvement for the implementation of Selekt that has been developed so far. For Catering, it is clearly better to use previous data in the expected values and test variables. For Production and Construction, the VAT predicted values perform equally well as previous values, but do not offer any improvement.

38. One of the key issues when using selective editing is finding appropriate expected values to be able to calculate an accurate score for each business. For ABS, less than half of the businesses have previous survey responses, which is a common and generally accurate expected value. However, it appears that in the context of Selekt, this problem can be more accurately solved by using the median of previous responses in the same 4 digit SIC, rather than by employing VAT predicted values.

39. These results only apply to two sectors of ABS and the results are marginal for one of these sectors. The VAT predicted values may prove more useful in other sectors. Furthermore, Selekt is a very complex selective editing method with many parameters. ONS are only just beginning to fully understand the interaction of these parameters. It is possible that as that understanding develops, other uses can be found for the VAT predicted values to enhance the efficiency of editing using Selekt.

V. Conclusion

40. This paper has described work to evaluate the benefits of using VAT data to improve the efficiency of editing in the UK Annual Business Survey. The VAT data were modelled with past survey data to produce a variety of predicted values for each of eight key ABS variables. These predicted values were first used in an attempt to improve the efficiency of the current ABS traditional edit rules. The results showed that the alternative edit rules were clearly better than those currently in use, giving much more accurate data for a similar number of failures. For the Catering sector a similar improvement was possible by defining similar alternative edit rules based on previous ABS data. However, for the Production and Construction sector the greatest benefits are only possible if VAT data are used.

41. The second use of the VAT predicted values was with the application of the Selekt selective editing program currently being tested for ABS. In this instance it proved, at least for the two sectors currently under analysis, that the VAT data did not give an improvement in the efficiency of editing. However, ONS experience of using Selekt is still at an early stage and it is still possible that the VAT data can have a role in developing the best selective editing approach for ABS.

42. The key conclusion from this study is that incorporating VAT data is necessary to produce the most efficient traditional micro editing rules for ABS. Any future review of editing for ABS should at least consider the benefits of using VAT data.

VI. References

Elliott D., 2009, *The potential use of additional VAT data in ONS business surveys*, UK Government Statistical Service Methodology Conference, London, June 2009.

Lewis D., 2009, *Assessing the potential of VAT turnover and expenditure data for use in ABI/2 editing and imputation*, ONS internal report.

Lewis D., 2010, *Developing predicted values for use in editing of structural business surveys*, final report for Eurostat Grant Agreement Number 30121.2009.004-2009.8.

Norberg A. and Arvidson G., 2008, *New Tools for Statistical Data Editing*, UNECE Work Session on Statistical Data Editing, Vienna, 21-23 April 2008.

Silva P.L.N., Bucknall R., Zong P. and Al-Hamad A., 2008, *A generic tool to assess impact of changing edit rules in a business survey – an application to the UK Annual Business Inquiry Part 2*, UNECE Work Session on Statistical Data Editing, Vienna, 21-23 April 2008.

Skentelbery R., Finselbach H. and Dobbins C., *Improving the efficiency of editing for ONS business surveys*, UNECE Work Session on Statistical Data Editing, Ljubljana, 9-11 May 2011.