

DEVELOPING SELECTIVE EDITING METHODOLOGY FOR SURVEYS WITH VARYING CHARACTERISTICS

Pam Tate
Office for National Statistics

1 Introduction of selective editing

Data validation and editing is one of the most time consuming processes in the production of official statistics; for business surveys it is generally reckoned to account for up to 40% of survey costs. Yet many validation failures result in no change; many editing changes have negligible effect on the survey estimates; and some are thought to introduce further errors.

In view of this, recent research in the Office for National Statistics (ONS) on data editing processes for business surveys has focused on developing new methods that would improve efficiency, without impacting adversely on data quality. A major element in this has been the development of suitable methodology for selective (or significance) editing, an approach which selects for follow-up of validation failures only those contributors where the likely size of the correction is expected to make a material difference to the results. (See for example Lawrence and McKenzie (2000) for a detailed description.)

In ONS, the approach was first applied to the Monthly Inquiry for the Distribution and Services Sector (MIDSS). The methodology developed was, for each case which failed validation, to compute a score which was designed to reflect the importance of editing that record. This was calculated as the absolute difference between an 'expected' value, taken to be the value from the previous survey period, and the returned value, weighted approximately as in the estimation procedure. The score was then standardised across domains by dividing by the domain estimate from the previous period.

Cases with scores above a pre-determined threshold were selected for scrutiny and editing, and those below the threshold were accepted without further scrutiny. Using data from the development phase of the methodology, the threshold was set at a level where the difference made to the estimate of the key variable was less than 10 per cent of the standard error of the estimate, thereby ensuring that the coverage probability of the estimates was not materially changed, (Sarndal *et al* (1992, p. 165)). Since MIDSS has two key variables, a score was calculated for each variable, and the case selected if either score was above its threshold.

During three months of piloting, the difference made by selective editing to the survey estimates, both overall and for the 101 key domains, was found to be less than 10 per cent of the standard error of the estimate in almost all domains. Between 3 and 5 domains had greater differences, but all were less than 30 per cent of the standard error of the estimate, so that the coverage probability of the 95% confidence interval was still greater than 94%. It was concluded that the difference to the survey outputs was not material.

The improvement in efficiency was marked – between 34% and 37% of validation failures did not need to be followed up. Additionally, a further 10% to 19% only needed to have contributors' comments scrutinised¹, a much faster process than full editing.

¹ In many ONS business surveys, provision is made for the contributor to write in a comment, for example to explain the reasons for large changes or to give notification of structural change. All such comments need to be scrutinised, even where nothing else on the form needs attention.

2 Considerations in extending selective editing to other surveys

The success of selective editing in MIDSS, in substantially improving efficiency without adverse impact on data quality, led to an objective of introducing the approach to other business surveys in ONS. This was not however a straightforward matter.

It is critical to the success of selective editing that the scoring mechanism reflects as closely as possible the effect of editing on the key survey estimates. Thus a suitable mechanism needs to be developed for each survey, based on its key variables and key domains, and reflecting the estimation process for the survey outputs. If there are many key variables, some method for combining or reducing them may be needed.

The mechanism then has to be evaluated on that survey, and the appropriate threshold found in the light of the variability of the data and the precision requirements of the estimates, both overall and for the key domains. Pre-edited data have to be available in order to be able to do this evaluation.

Finally, a source of the 'expected value' of the key variables needs to be available. For most short period business surveys, data from a previous period are available for most of the sample. Where this is not the case, an alternative source is needed.

Each survey that was being considered for selective editing thus needed a methodology appropriate for that survey to be developed and evaluated. This process is described below for four surveys with differing characteristics.

3 Extending selective editing to other surveys

3.1 Monthly Production Inquiry

The Monthly Production Inquiry (MPI) has more variables than MIDSS, and five of them were identified as key, (although two of the five were only applicable to a small proportion of the sample). It also has a large number – 216 - of key domains. It was not clear beforehand whether it would be possible for selective editing to work efficiently with as many individual key variables as five, especially with the additional volatility that could arise with a large number of domains.

However, the same scoring mechanism was used as for MIDSS, for each of the five key variables, and a contributor was selected for editing if any of the five scores was above the threshold. Again, the thresholds were set so that differences to the estimates, both overall and for the key domains, were not material, being small by comparison with sampling error.

This resulted in efficiency gains that were less than for MIDSS, but not much less, and still very worthwhile. On average, over the pilot period, some 33% of validation failures did not need to be followed up, and a further 8% only needed to have contributors' comments scrutinised.

3.2 Annual Business Inquiry (part 1)

The Annual Business Inquiry part 1 (ABI/1), is the component of this annual survey that gathers data on employment. It has, like MIDSS, two key variables, and the same scoring mechanism was used. The main difference is that, being an annual survey, there is a much lower level of sample overlap, and hence there are previous period data available for only a small proportion of the sample.

A possible alternative source for expected values is the Inter-departmental Business Register (IDBR), which provides the sampling frame for this and other business surveys, and which includes data on employment, though for small businesses this is not very frequently updated. Data from this source were tried as expected values when previous period data were not

available, but this produced many zero-scoring contributors with small editing changes which cumulatively exceeded the allowed maximum difference.

The selective method was therefore tried just on the small proportion of the sample for which previous period data were available – the large sample size for this survey meant that this could still produce worthwhile efficiency gains. The method has now been piloted, and is currently being evaluated.

3.3 Monthly Wages and Salaries Survey

The Monthly Wages and Salaries Survey (MWSS) represents the greatest departure from the previous surveys. Its main use is in the production of the Average Earnings Index (AEI), which is a very different kind of output from those previously considered. The AEI is an index of rates of pay, and is produced both including and excluding bonuses. Data are gathered from businesses on their number of employees and the amount paid to them, so errors in either of these could lead to errors in the key variable of pay rates.

It was decided that the range of potential errors could best be captured by calculating scores on the basis of the three measures: average weekly paybill including bonuses; average weekly paybill excluding bonuses; and average pay rate excluding bonuses.

The scoring mechanism for each of these variables needed to reflect the contribution to the index of any change in the variable resulting from editing. Formulae for the contribution could be derived from Bird and Youll (2000), but these formulae involved parameters that were dependent upon the whole dataset for the current period – information that would not be available at the time of editing. However, these parameters were not expected to change rapidly over time, and so it seemed reasonable to use the values from the previous time period.

The scores also needed to be standardised to the key domains, defined by 2-digit SIC and whether public or private. The scores were therefore calculated as a percentage of the rate for the domain from the previous period.

As for the previous surveys, thresholds were set at a level designed so that the differences made to the indices, both including and excluding bonuses, at domain and overall levels, should be small in comparison with sampling error. This was initially done by requiring the differences to be less than 10% of the standard error, but it was found that, for the index excluding bonuses, there were large inconsistencies between overall and domain levels. The same threshold as for the index including bonuses was therefore tried, and produced satisfactory results. As before, a contributor was selected for follow-up if any of the three scores was above the threshold.

The implementation of this methodology on the live system has required rather more programming work than the previous surveys. This is because the scoring mechanism is not only very different and more complex, but also requires data from various sources including a separate results system based on different software.

The piloting of this methodology is expected to start later this year.

3.4 New Earnings Survey

The New Earnings Survey (NES) is based on a sample of employees, but the data are obtained from employers' records. Five key pay and pay rate variables were identified. The key domains were defined by occupation and whether full or part time.

Pre-edited data at the time of data capture were not available, and it was therefore not possible to set thresholds for a specified allowable level of difference to the estimates. A methodology was therefore developed for prioritisation only, without setting thresholds.

There is currently no weighting in the estimation from this survey, so the score was initially tried as $|(expected\ value - returned\ value)|/domain\ total\ for\ previous\ year$. However, some

domains can be very small, so it was proposed to divide the domains into large and small categories, and combine the small domains into one for prioritisation purposes.

The expected value was taken as the previous year's value where this was available, and where the occupation was unchanged, and otherwise as the domain median from the previous year. The largest of the scores for the five key variables was used for the prioritisation.

The details of the method are currently being finalised, and it is likely to be tried later this year.

4 Issues in selective editing for further consideration

As noted above, selective editing has been successfully applied with individual scores for up to five key variables. However, there are some surveys that have a very large number of variables. These will need some method of selecting variables or combining scores to keep the number manageable and maintain an efficient procedure. Work is currently proceeding on the possibilities of applying selective editing to a survey of this kind.

Some surveys have key variables that are by their nature particularly volatile over time, for example stocks or capital expenditure. In these circumstances, the previous period's value, though still probably the best predictor of the current value, may not be a good enough predictor for selective editing to be efficient. It is planned to investigate this question later this year.

It is possible that a contributor could provide data that are increasingly in error, but with changes from one period to the next that consistently produce selective editing scores below the threshold. A monitoring system has recently been set up for the MIDSS survey to investigate this possibility.

Thus far, the threshold values for selective editing have been set on the basis of past datasets including both pre- and post-edited data, and then confirmed during a pilot on the live system. It is possible that the characteristics of the survey data may subsequently change, and there is a need for methods for monitoring the continuing appropriateness of the threshold values over time.

Additionally, the initial threshold values were set on a very conservative basis as regards the allowable difference to the survey outputs. Some further investigation of the relationship between changes in the threshold level and the resulting differences to survey outputs is needed to assess the scope for further gains in efficiency. This and the previous issue are being addressed in a small project on threshold-setting.

So far, selective editing has been implemented as a filter added on to existing validation systems. It would however be possible to apply the selection process to all data, with no validation at all, (except probably for new contributors). This remains to be investigated.

References

Bird D. and R. Youll (2000), The UK Average Earnings Index: A Modified Monthly Chain-Linked Time Series, *Proceedings of the Second International Conference on Establishment Surveys*, American Statistical Association.

Lawrence, D. and R. McKenzie (2000), The General Application of Significance Editing, *Journal of Official Statistics*, 16, pp. 243-253.

Särndal, C.-E., B. Swensson, and J. Wretman (1992), *Model Assisted Survey Sampling*, New York: Springer.