

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Vienna, Austria, 21-23 April 2008)

Topic (iii): Improvement of quality through data editing

Assessing the performance of the thousand pounds automatic editing procedure at the ONS and the need for an alternative approach

Supporting Paper

Prepared by Alaa Al-Hamad, Daniel Lewis (Office for National Statistics, UK)
and Pedro Luis do Nascimento Silva (University of Southampton, UK)

I. INTRODUCTION

1. In recent years the ONS has been undergoing a large efficiency programme to its business statistical surveys with a vision mainly concerned with increasing the efficiency of data processing. This programme has three aims:

1. to cope with the ever increasing demand for statistical information within a shorter time period, at lower cost and at a much higher level of detail and quality than before;
2. to meet financial pressures; and
3. to generate savings for reinvestment into improved outputs.

2. Currently one of the most costly components of the survey processing procedures in ONS is data cleaning. This is because to fulfil our role as a National statistical agency successfully, the organisation is always under pressure to produce high-quality data in order to maintain public confidence in official statistics. This leads to ONS business surveys having high cleaning costs – on average consuming around 40% of the total survey budget. Furthermore, most of the resources within data cleaning are consumed by data editing. In practice, editing consists mainly of two stages:

- error localisation / detection;
- error verification / correction.

3. The very high costs are generally due to use of labour intensive processes for error verification / correction, which often involve re-contacting respondents. The requirement for manual intervention is introduced through the nature of the errors commonly found in business surveys. Some errors can be corrected through the application of logical editing rules, and some through recalculation of totals, but a large proportion are often simply “suspect data” that need to be queried directly with the businesses that provided the responses. These suspect data are rarely found to be in error - studies have found the “hit rate” to be as low as 20% and this is found to be true for ONS surveys. Therefore, these suspect data are often simply confirmed by the respondents and hence pass through the data editing process unchanged. The benefits from querying these confirmed data are often intended to be informative rather than statistical, i.e., they help explain movements in the data.

4. In the ONS a number of methods have been used over the years to cut costs and achieve more efficient editing processes. Some of these methods prioritise suspect data by their importance, i.e. selective editing and (to a degree) the Hidioglou-Berthelot method (Hidioglou and Berthelot (1986)). Others correct these suspect data automatically through recalculation and the application of logic. However, although such methods are employed for some surveys, it is thought they are not fully utilised

to achieve the maximum efficiency across all business surveys. In fact, some surveys have not been reviewed for a number of years which can result in increased resource costs and the possibility of lower quality data.

5. The aim of the wider study on which this paper is based is to review the current editing process for a large ONS business survey, Annual Business Inquiry part 2 (ABI/2), and to suggest ways in which efficiency can be improved. The approach adopted considers a holistic view of the editing process, rather than only some of its individual stages. Detailed specifications for the current editing processes have been examined, as well as their connections to data collection mode, data capture and the current editing methodology. The broader aim is mainly to learn from this exercise, and assess what possible improvements could be made to other ONS business surveys. However this paper will focus only on one aspect of this study, the consideration and use of automatic editing procedures in an attempt to recommend a more efficient alternative to the current procedure.

II. AUTOMATIC EDITING £000 ERRORS IN THE UK

6. Automatic editing in the ONS is a procedure used for *locating* and *correcting* possible systematic errors in the returned survey data automatically. The ultimate aim of automatic editing is to achieve resource efficiency without impacting on the quality of the statistics produced. Therefore the current procedures are designed to run before any validation is carried out on the survey returns. This assists in reducing the number of validation failures triggered, and therefore reduces the need for manual editing. Currently two main automatic editing procedures are used for ONS business surveys, namely: automatic £000 editing (mainly for turnover) and automatic employee totalling. Our focus in this paper will be on £000 editing.

7. The automatic editing of £000 errors was first introduced in ONS in 2001 after research showed that the bias introduced to business surveys estimates through applying the automatic editing procedure is negligible and hence it will not impact adversely on the quality standards of our survey estimates. The assessment of the size and the significance of the bias used the bias ratio measure as suggested by Särndal, Swensson and Wretman (1992).

$$BR(\hat{\theta}) = \frac{B(\hat{\theta})}{[\hat{V}(\hat{\theta})]^{1/2}}$$

8. To qualify for automatic editing of £000 errors a contributor must have been responding to the survey for at least two periods and a total turnover value from the previous period must be present (either returned, imputed or constructed). The method developed for editing £000 for number of business surveys is as follows:

For each contributor, i , calculate the ratio of current period's pre-edited total turnover to previous period's post-edited total turnover:

$$Ratio_{t,i} = \frac{y_{t,pre-edited,i}}{y_{t-1,post-edited,i}}$$

IF

$$k_1 < Ratio_{t,i} < k_2$$

where k_1 and k_2 are pre-defined parameters, set to ensure an acceptable bias ratio.

THEN

$$y_{t,post-edited,i} = Round_{0dp} \left(\frac{y_{t,pre-edited,i}}{1000} \right)$$

Otherwise original values are used in validation.

III. ISSUES WITH THE CURRENT EDITING PROCEDURE

9. Our investigation to the current £000 automatic editing procedure has highlighted a number of issues. The first is, it places strong belief in the value of turnover from the previous survey or from the register. Hence only increases of around 1,000 times will become suspicious, and lead to detection followed by automatic editing. However, if the value of turnover on the previous survey was reported wrongly in pounds rather than thousands of pounds, then this rule would not detect that type of error, instead such an error will be detected during the manual validation process.

10. In addition, businesses whose records are affected by automatic editing are not informed that they made a mistake in filling in the values in pounds rather than in thousands of pounds. So for businesses taking part in successive rounds of the ABI/2, automatic editing of their returns means that they will not learn from their previous mistakes.

11. Furthermore another issue stems from the fact that this rule is the first step in the data editing process, and hence, data for businesses which were changed by the automatic editing may still be changed again later during the subsequent manual editing operation, so that the process does not necessarily make as many efficiency savings as supposed.

12. Exploratory analysis in Silva (2007) found that many errors are caused by respondents reporting in pounds rather than thousands of pounds, there are also errors caused by respondents reporting in pounds and pence (and are therefore 100,000 times too large). The pounds and pence errors are not currently automatically detected. The analysis also found that only a small proportion of businesses had a previous year turnover value available (21.6% of the businesses being examined), so that in most cases the register turnover value was used for the rule instead. Register turnover is generally thought to be a less good predictor of the current survey turnover value. Many cases which, on manual inspection, are obviously thousand pound errors were not picked up by the current procedure because the predicted turnover value was not close enough to the true current turnover value.

IV. AN ALTERNATIVE DETECTION RULE FOR £000 AUTOMATIC EDITING

13. Given the possible shortcomings of the current ONS method for automatically editing thousand pounds errors, Silva (2007) investigates alternative detection rules. The proposed suggestion is to use a statistic which identifies the difference in the number of digits between the reported and predicted turnover values for each business. This may be achieved using the following statistic.

$$Diff_i = abs \left\{ ceiling \left[\log_{10} (y_{i,raw}) \right] - ceiling \left[\log_{10} (y_{i,pred}) \right] \right\}$$

where $y_{i,raw}$ is the returned unedited turnover value and $y_{i,pred}$ is the predicted turnover value (previous year's turnover where available, otherwise register turnover). The *ceiling* function returns the smallest integer larger than or equal to the value and *abs* returns the absolute value of the expression.

14. £000 errors may then be identified as those businesses for which the value of $Diff_i$ is equal to 3. In addition to this, pounds and pence errors may be identified as those businesses where $Diff_i$ is equal to 5. Note that, unlike the current ONS detection rule, this method identifies discrepancies of 3 or 5 digits in both directions. This means that the rule can also identify cases where the predicted value appears to have a £000 error.

V. CASE STUDY USING BUSINESS SURVEY DATA

15. In order to test the benefits of using the alternative detection rule for errors in measures of units, a small study was undertaken using 2005 ABI/2 data. The aim of the study was to compare the performance of the alternative rule with the current ONS £000 rule. To test performance accurately would require access to unedited survey returns as well as the true values for all respondents. In practice, true values are generally unavailable and even truly unedited values are not always stored. The available data for the study contained unedited and final turnover variables. However, at least some of the 'unedited' values had already gone through automatic editing. Therefore, it was necessary to construct a more realistic unedited turnover variable by readjusting those businesses that were flagged as having gone through automatic editing. In the absence of the true values, the only option was to assume the final turnover value was correct. As with the data used in Silva (2007), there are many businesses for which previous year's turnover was not available, so that for the majority the predicted turnover is the register turnover value.

16. The two automatic editing detection methods were applied to the data. The proposed alternative method was set to identify discrepancies of 3 or 5 digits. For the purposes of this comparison, only positive differences (indicating errors in the current year rather than the previous year) were automatically imputed. The following results show the percentage of true errors identified by each method and the proportion of false hits.

Table 1: Percentage of true errors that were identified by automatic editing rules

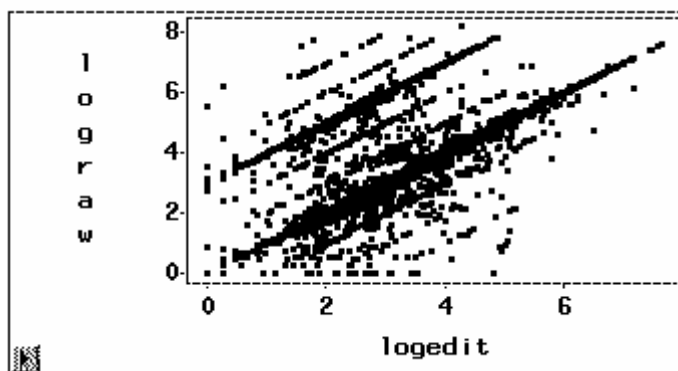
	Current ONS rule	Proposed new rule
% true errors identified	37%	44%

Table 2: Percentage of errors identified by automatic editing rules that were false hits

	Current ONS rule	Proposed new rule
% false hits	0.3%	2.5%

17. The above results show that the proposed new rule does identify more errors than the current rule for these data, but also leads to more false hits. The proposed new rule identifies 44% of all errors from this running of the survey. It is useful to investigate whether this 44% represents all errors due to respondents using incorrect units of measurement. The following graph, plotting the log of the unedited turnover values against the log of the final edited values, clearly shows the £000 errors.

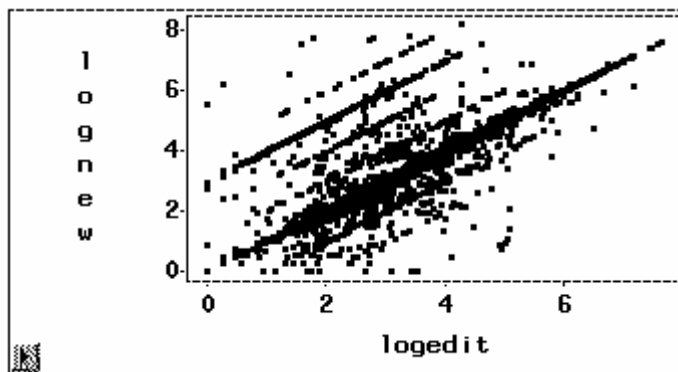
Figure 1: Plot of unedited (raw) turnover against fully edited turnover



18. There is a clear line parallel to the main scatter of values and 3 units higher up the graph. This is caused by the £000 errors. There is evidence of other lines, including one 5 units higher than the main body of data which would correspond to pounds and pence errors. The next graph plots turnover which

has been subject to automatic imputation following detection by the alternative automatic editing rule against the fully edited turnover.

Figure 2: Plot of automatically edited turnover using new rule against fully edited turnover



19. Automatic editing has dealt with a large proportion of errors caused by respondents using incorrect units of measurement. However, there is clear evidence of further £000 errors that have not been detected even by the new rule. As would be expected from the results in table 1, the corresponding graph using the current ONS automatic editing method shows that slightly less of these errors have been picked up.

20. Analysis of the data shows that there are businesses which appear to be £000 errors that have not been identified by either detection rule. The reason for this is that the predicted turnover (based on either the previous year's turnover or the register turnover value) often differs from the final turnover by a power of 10, so that comparing unedited turnover with predicted turnover results in a difference of 2 or 4 digits rather than 3. To improve further on the detection of this kind of error would require a better method for predicting turnover.

VI. CONCLUSION

21. This paper has investigated the method of automatic editing currently used in ONS to detect and automatically impute £000 errors in financial variables. A number of issues were identified with the current method, which led to the proposal of an alternative detection rule for this type of automatic editing. The proposed alternative counts the number of digits difference between unedited turnover values and predicted turnover values. This is in contrast to the current method, which looks at the ratio between unedited and predicted values and sets a threshold within which values are considered to be £000 errors. In a simple case study, using ABI/2 data, the alternative method does identify more errors than the current method. However, both methods fail to detect a large number of errors that look to have been caused by respondents using the incorrect units of measurement. This is due to the predicted turnover values often differing from the finally edited values by a power of 10, so that the observed £000 errors are not identified by the rules. Predicted turnover values are currently derived by taking the previous year's turnover value if available, or the register turnover value if not. Whilst the new detection rule may deliver small improvements to this automatic editing procedure, greater improvements may be forthcoming if it is possible to identify a better method of predicting turnover.

VII. REFERENCES

Hidioglou, M. A. and Berthelot, J.-M. (1986) *Statistical Editing and Imputation for Periodic Business Surveys*. Survey Methodology, June 1986, Vol 12. No 1. pp73-83.

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

Silva, P.L.N. (2007) *Editing and imputation for the ABI/2 survey – interim report 1*. Internal ONS report.