WP.42
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 (viii): Selective and macro editing

ASSESSING THE IMPACT OF SELECTIVE EDITING ON DATA QUALITY

Invited Paper

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

I. INTRODUCTION

1. Data editing takes up a large amount of the resource spent on business surveys – see Granquist (1995), Granquist & Kovar (1997). In common with many statistical agencies, the UK Office for National Statistics (ONS) has been looking at ways of improving the efficiency of the editing process. Traditionally, micro-editing has involved setting a variety of edit rules to detect suspicious returned values. These rules can be split into two main types – hard edits and soft edits. Hard edits specify conditions that must be met for the returned value to make sense, for example employment must be positive. In contrast, soft edits identify suspicious values which may or may not be in error. Soft edits often involve setting thresholds either for a particular returned value itself or for some function of the returned value and other related values. For example, if the ratio of a returned value to the previous period's response for a business is larger (or smaller) than a specified threshold the business would fail the edit. Businesses failing such an edit are often re-contacted to confirm their response.

2. ONS has recently undertaken two initiatives to improve the efficiency of business survey micro-editing. The first is a SAS program known as Snowdon developed by ONS Methodology Directorate and enhanced with assistance from Southampton University. Snowdon is a tool to analyse the effects of changes to edit rules already existing on a survey. The procedure involves reading in the existing edit rules and then making changes to the thresholds of those which fail large numbers of businesses. The effect of changing these thresholds is assessed by calculating a range of indicators comparing the impact of the existing and revised edit rules on the data. The indicators include edit failure rates, hit rates, saving (i.e. the difference in the number of businesses failing edits between existing and revised edit rules) and bias. These indicators are derived using the unedited and edited versions of the response data. Any business which failed the existing edits but which would not fail the revised edits is assigned the unedited value rather than the edited value. Indicators such as the bias, for example, can then be estimated by calculating the difference between estimates based on the current edit rules and those that would have been obtained if the revised edit rules were used. For more detailed information on the Snowdon approach see Silva et al (2008).

3. The second initiative is a more fundamental review of the editing process for ONS business surveys. This follows a visit by Eden Brinkley and Lynn Bismire of the Australian Bureau of Statistics (ABS) in 2007 which described their work to improve data editing in ABS business surveys. One of the key methodological aspects of the ABS approach was using selective editing as the main method of micro-editing. Selective editing is currently used on a number of ONS business surveys in conjunction with traditional micro-editing rules following ideas developed by Hedlin (2001, 2003). For these surveys,

[1] Prepared by Daniel Lewis & Alaa Al-Hamad of ONS (daniel.lewis@ons.gov.uk , alaa.al-hamad@ons.gov.uk) and Pedro Luis do Nascimento Silva of Southampton University.

business returns are first submitted to the normal edit rules. Any business failing edit rules is then selectively edited and only those judged to have an impact on published estimates are re-contacted to confirm their figures. In the ABS approach, all businesses are selectively edited and there is very little use of any other micro-editing rules. The focus of the strategy is to only spend resource re-contacting those businesses which are expected to have an impact on the quality of published estimates. In ONS, a project was set up to review ONS editing processes in the light of the ABS approach – see for example ABS (2007).

4. As part of the project reviewing editing processes, ONS Methodology Directorate and Southampton University are exploring selective editing methodology for ONS business surveys, beginning with the Retail Sales Inquiry (RSI). This paper discusses the methods proposed for RSI and describes how they were evaluated using the same indicators developed for Snowdon. The paper then reports on new analysis to assess the cumulative impact of implementing selective editing for RSI.

II. SELECTIVE EDITING APPROACH FOR RETAIL SALES INQUIRY

5. The RSI is a monthly survey collecting information on turnover every month and employment every three months. The initial development of selective editing was restricted to the turnover variable. Businesses responding to RSI give approximately monthly turnover figures which are then converted to an adjusted turnover value using date information from the business to calculate the average weekly turnover for the month. This standardises the periodicity of the turnover responses, since not all businesses give their monthly turnover information for the same dates. Adjusted turnover was the variable used for selective editing.

6. Selective editing works by prioritising businesses for editing by calculating scores for each responding business. In the method chosen for RSI, scores are calculated by predicting the effect each business would have on survey estimates if unedited values were used instead of edited values – i.e. the effect on estimates of not re-contacting the business. For this study, the parameter estimate of interest was total turnover. In practice, the published estimates for RSI are seasonally adjusted indices derived from the total turnover estimates, but basing selective editing on total turnover should be sufficient to appropriately focus the editing effort.

7. The main selective editing study was carried out by Pedro Luis do Nascimento Silva of Southampton University and Ping Zong of ONS Methodology Directorate. In this work, a variety of selective editing scores were tested of the form:

$$d_i = 100 \times \frac{w_i \left| z_i - \hat{y}_i \right|}{\hat{T}_d}$$

where w_i is the estimation weight for unit i

z_i is the unedited adjusted turnover value for unit i

$\hat{y}_i$ is the predicted adjusted turnover value for unit i

$\hat{T}_d$ is the previous period's total turnover estimate for the domain

This score estimates the difference between the unedited and edited turnover values as a percentage of the domain estimate. In this way we can estimate the effect on the estimate of not editing turnover for the business. Since in practice we will not know the edited values for the current period, we need to estimate them in some way. Silva (2009) examined a range of options for predicting adjusted turnover, including the adjusted turnover value from the previous period, the annual turnover value from the business register (divided by 52 to approximate weekly turnover) and a pseudo-imputed value (calculated by multiplying the previous value by the link factors used for ratio imputation in RSI). Of these, the previous period's turnover value was shown to be the best option.

8. Silva (2009) and Zong (2009) examined whether the selective editing score should be calculated at the survey level or separately for the 17 important published domains (industry groupings). They found

that setting the score at the survey level did not produce good results for estimates of some of the published domains. Therefore, it was decided to set scores at the domain level.

9. Having calculated scores for each business, it is necessary to define thresholds for the scores, above which businesses are deemed to have failed selective editing and will be re-contacted to confirm their figures. The thresholds were set separately for each domain. The choice of threshold was determined in Silva (2009) using the absolute relative bias:

$$ARB(\hat{T}_{d,sel.edit}) = \left| \frac{\hat{T}_{d,sel.edit} - \hat{T}_{d,current}}{\hat{T}_{d,current}} \right|$$

where $\hat{T}_{d,sel.edit}$ is the estimated total turnover for the domain using selective editing

$\hat{T}_{d,current}$ is the estimated total turnover for the domain using the current RSI edit rules

A range of threshold values were tested for each domain to find the threshold which resulted in the least amount of editing under the condition that the absolute relative bias should be less than 1%. With the bias this low, there should be no significant difference between the estimate based on selective editing and the estimate using the current RSI edit rules. In this way, the most appropriate threshold for each domain was found.

10. The result of selectively editing RSI data was assessed using quality indicators originally developed for Snowdon. Selective editing was compared with current RSI micro-editing rules. For selective editing, any business with a score below the threshold was assigned the unedited adjusted turnover value (since they would pass the selective editing rule) and any business with a score above the threshold was given the edited adjusted turnover value. For the current RSI micro-editing rules, businesses which failed at least one of the rules were assigned the edited value and businesses which failed no rules were assigned the unedited value. Note that RSI data are also subject to macro editing. Any suspicious looking aggregate estimates can lead to businesses being re-contacted later in the process to confirm their figures. This means that errors missed by the current micro-editing rules can be measured by comparison with the final edited value. The main indicators used to assess the quality of the selective editing were the edit failure rate, edit change rate, relative bias and absolute relative bias. Absolute relative bias is defined above. The other indicators are defined as follows:

$$\text{Edit failure rate} = \frac{\text{Number of businesses failing at least one edit}}{\text{Total number of responding businesses}}$$

$$\text{Edit change rate} = \frac{\text{Number of businesses whose values changed as a result of failing edits}}{\text{Total number of responding businesses}}$$

$$\text{Relative bias, } RB(\hat{T}_{d,sel.edit}) = \frac{\hat{T}_{d,sel.edit} - \hat{T}_{d,current}}{\hat{T}_{d,current}}$$

Selective editing was implemented and analysed using 36 months of RSI data (January 2005 to December 2007). Because the selective editing score required previous month's data, it was not possible to calculate scores for January 2005. Therefore, results were calculated for the 35 months starting in February 2005.

III. RESULTS OF INITIAL SELECTIVE EDITING STUDY FOR RSI TURNOVER

11. One of the key questions in implementing selective editing is how much editing resource can be saved. Figures 1 and 2 compare the edit failure rate and edit change rate for the current RSI micro-editing rules and for selective editing over the 35 months of the study.

Figure 1: Edit failure rate and edit change rate for current micro-editing rules

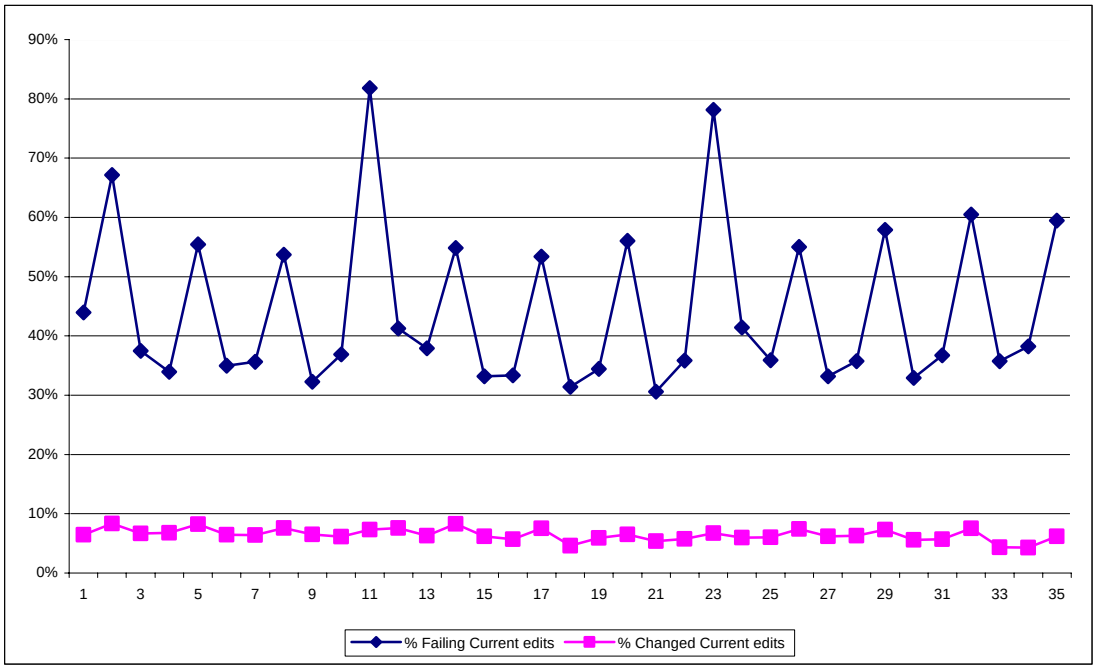


Figure 1 shows that the current micro-editing rules have a high failure rate, always over 30% and often over 50%. Note that the peaks occur on the quarter months when employment is also collected in the survey. If all of these businesses are re-contacted, it is clear that this will take up a large amount of the resource spent on the survey. The graph also shows that the current rules are fairly inefficient, with less than 10% of businesses having their adjusted turnover value changed due to editing.

Figure 2: Edit failure rate and edit change rate for selective editing

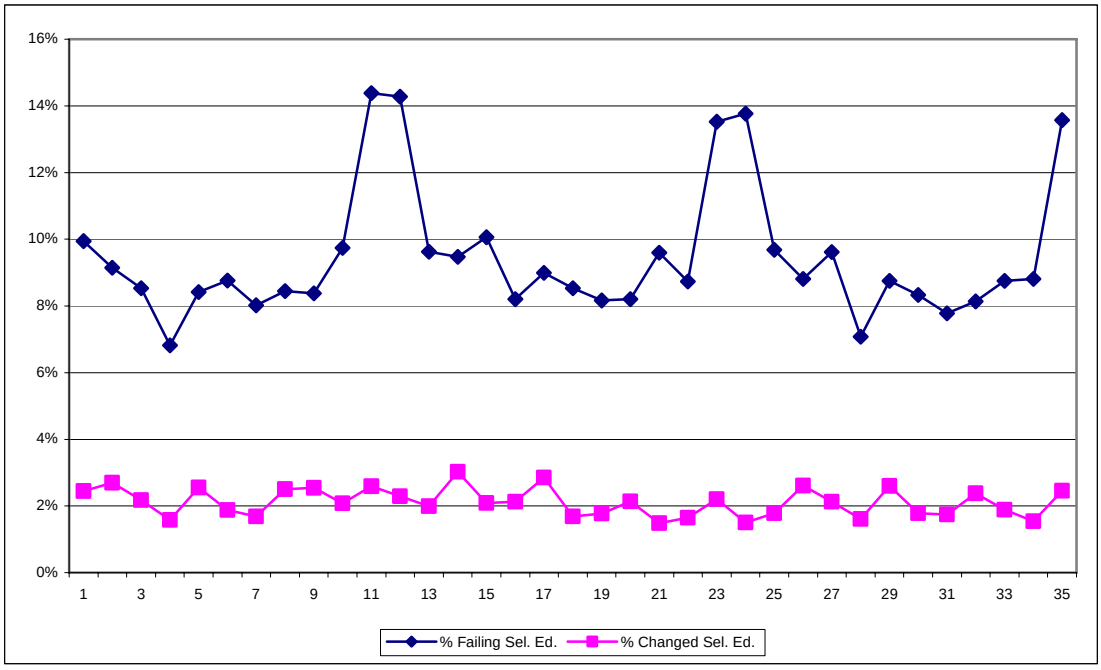


Figure 2 shows that the number of businesses failing selective editing is a lot smaller. In the month with the highest proportion of failures, less than 15% fail editing. This shows that the aim of reducing resource spent on micro-editing should be easily met. Note that more businesses fail selective editing in December and January. This is caused by high seasonal turnover in December reducing the accuracy of using previous month's turnover in the score. The proportion of businesses whose adjusted turnover value is changed due to selective editing is a little lower than was the case for the current micro-editing rules.

12. Selective editing works by focussing on those businesses whose returns could have a significant impact on the quality of published results if they are not edited. If the method has worked well, the businesses which were edited under the current edit rules but not edited under selective editing should not contribute a significant amount to published estimates. To test this, we can look at the relative bias and absolute relative bias of estimates resulting from selective editing. Figures 3 and 4 show how estimates of these biases vary over the 35 months of the study for each of the 17 key published RSI domains.

Figure 3: Boxplots of relative bias (%) for RSI total turnover estimate using selective editing by publication domains

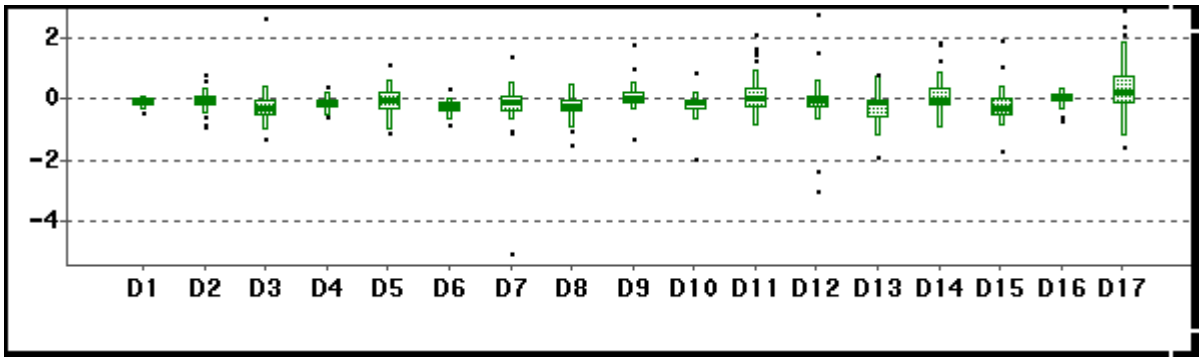
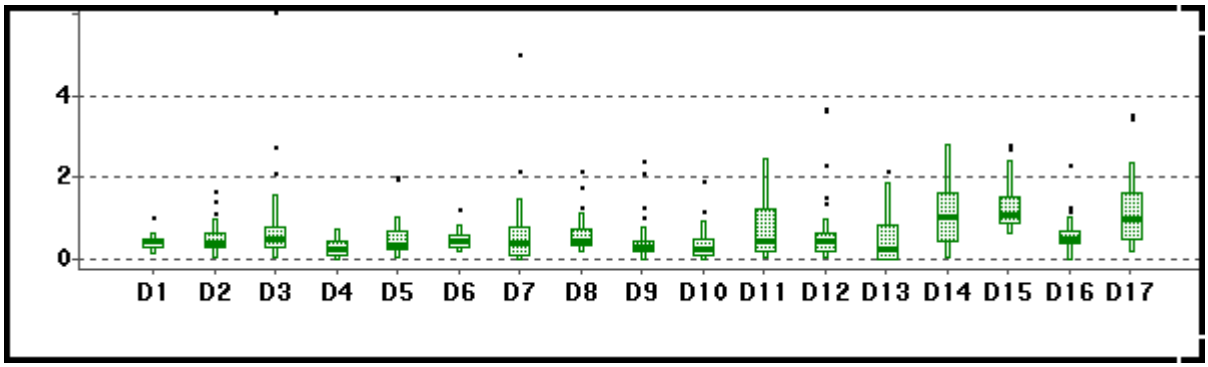


Figure 4: Boxplots of absolute relative bias (%) for RSI total turnover estimate using selective editing by publication domains



The graphs show a low level of bias for all domains. In figure 3 we can see that the median bias for all domains is close to zero. The bias for the worst periods and the worst domains is not often larger in absolute value than 1% and only very occasionally larger than 2%. Because biases can be positive or negative, there is some degree of cancellation in the relative biases. The absolute relative biases in figure 4 could be said to give a good indication of the size of the maximum potential bias that could be expected from selective editing. The median absolute relative bias in each domain is within the target of 1%. The worst months show some larger biases, but in the majority of cases the relative absolute bias is below 2%.

13. It is also possible to get an idea of the bias in the current RSI micro-editing by estimating the relative bias and absolute relative bias. In this case the bias comes from errors missed by micro-editing but identified and changed by macro-editing. Figures 5 and 6 show these estimated biases for each domain.

Figure 5: Boxplots of relative bias (%) for RSI total turnover estimate using current RSI micro-editing by publication domains

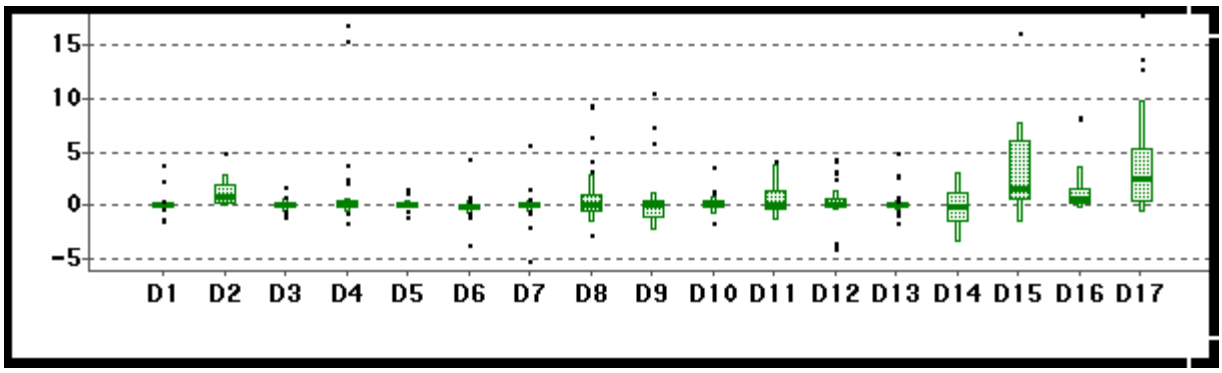
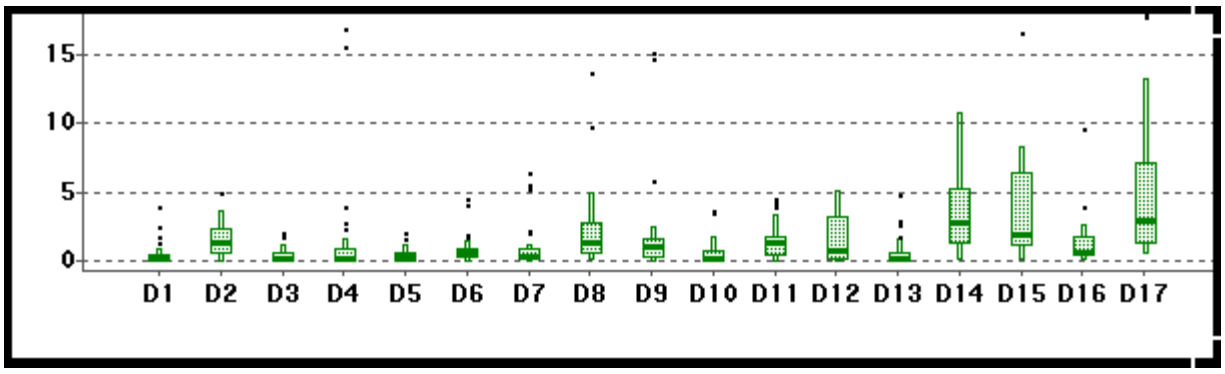


Figure 6: Boxplots of absolute relative bias (%) for RSI total turnover estimate using current RSI micro-editing by publication domains



Note that in order to view the graphs clearly, any biases (or absolute biases) greater than 20% were removed from figures 5 and 6. This resulted in 33 outliers being removed in both cases, with the largest being well over 100%. Figures 5 and 6 show that the biases resulting from the current micro-editing rules are actually larger than those from using selective editing. Because the selective editing focuses on those businesses which have the largest impact on the total turnover estimate, it has been possible to improve on the accuracy of the traditional micro-editing rules used by the survey.

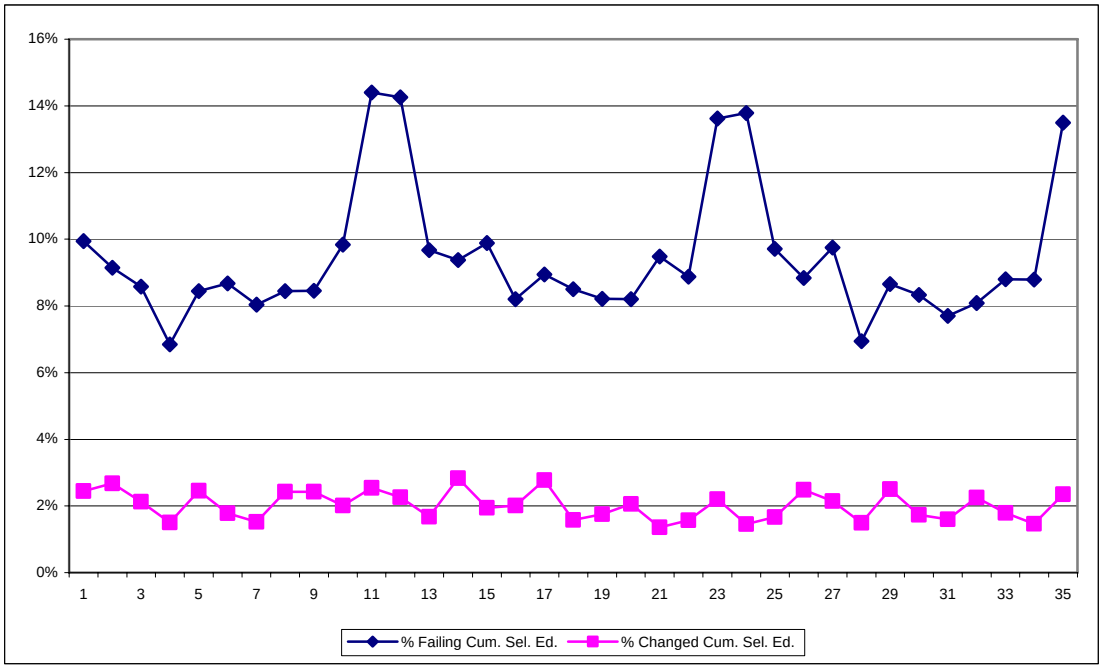
14. The results of the initial selective editing study for RSI were very promising. The difference in edit failures between selective editing and the current edit rules translated into average savings of over 50% for each domain. At the same time as making this saving, using selective editing removes more of the error, as a proportion of the final estimates. However, in calculating scores for this study, the previous period's edited values were used. These values came from the current editing process (the combination of micro-editing and macro-editing). In practice, fewer businesses will have their values changed under selective editing, so the edited values available will be of different quality. Section IV reports on a study to see how much the cumulative impact of selective editing will affect the results of the initial analysis.

IV. ASSESSING THE CUMULATIVE IMPACT OF SELECTIVE EDITING

15. In order to get an idea of the impact that selective editing will have on data quality in practice, we ran selective editing for RSI using the same method described above and the same data, with the difference that the edited adjusted turnover variable was changed according to the results of selective editing. For any business failing selective editing, the new edited value was set equal to the edited value from the current RSI edit rules (both micro-editing and macro-editing). The new edited value for any business passing selective editing was set equal to the unedited value. This new selectively edited variable was then used in the next period as the previous data needed to calculate the score function (for both the previous value and in calculating the previous period's total turnover estimate). This will inevitably affect the accuracy of the scores and may therefore affect the results of the selective editing.

16. In order to assess the cumulative impact of selective editing, we calculated the same quality indicators as used in the original study. Figure 7 shows the edit failure rate and edit change rate for the cumulative selective editing study.

Figure 7: Edit failure rate and edit change rate for selective editing with scores calculated using previous selectively edited data



Comparison with figure 2 shows that the edit failure and edit change rates are virtually unaffected by the change in the way selective editing scores were calculated. This suggests that the savings reported above could be expected to be indicative of what will be achieved when selective editing is implemented.

17. To test the impact on data quality of selective editing with scores calculated using data that will be available in practice, figures 8 and 9 show the relative bias and absolute relative bias of total turnover estimates resulting from the cumulative selective editing study.

Figure 8: Relative bias (%) for RSI total turnover estimate using selective editing with scores calculated using previous selectively edited data

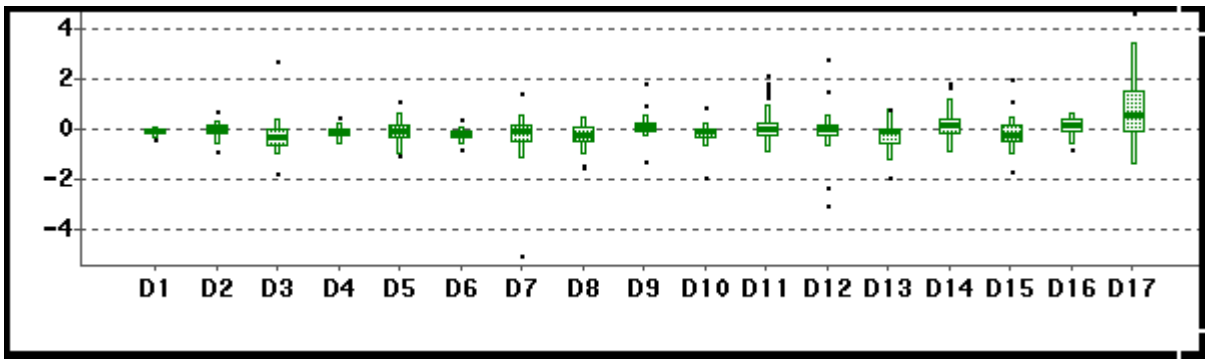
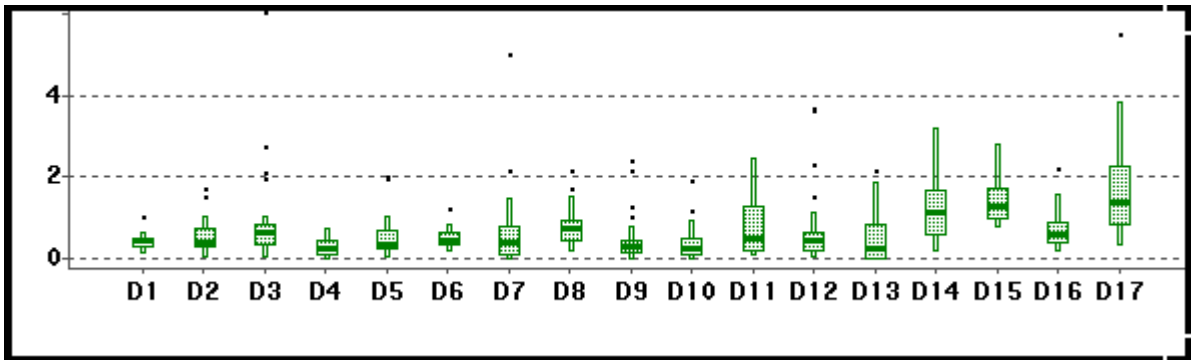


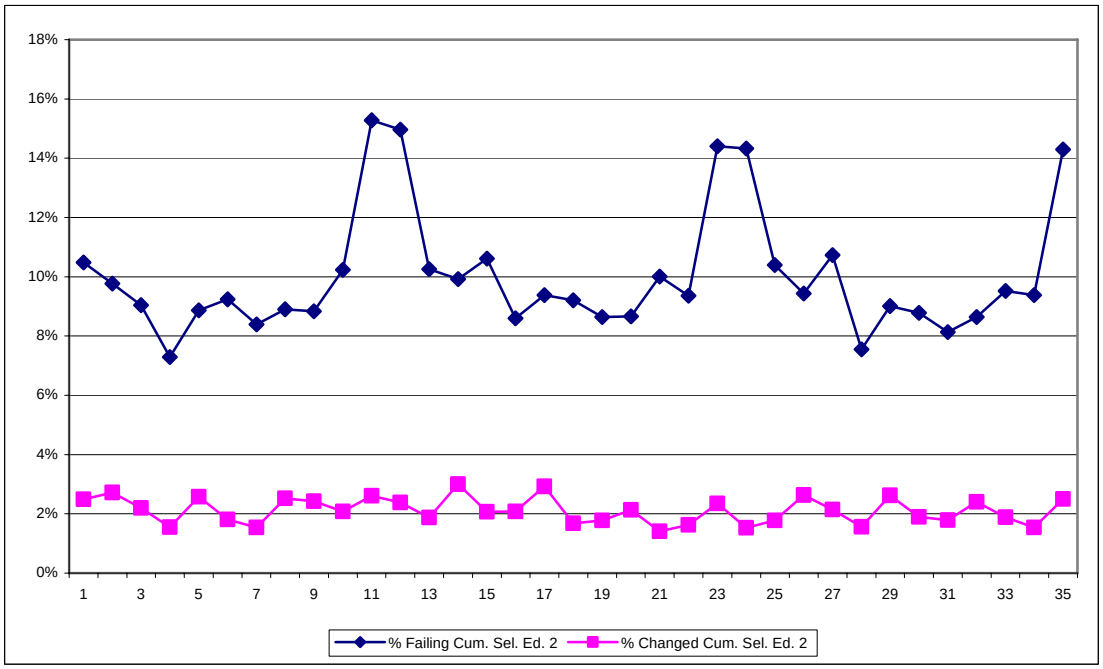
Figure 9: Absolute relative bias (%) for RSI total turnover estimate using selective editing with scores calculated using previous selectively edited data



The bias graphs are very similar to figures 3 and 4, however there are small increases in bias in many domains. In domains 15 and 17, the median absolute relative biases are now a little above 1% (1.3% and 1.4% respectively).

18. Figures 8 and 9 show that in practice the bias from selective editing may be expected to be slightly larger than the initial results showed. If we wish to keep the relative absolute bias under control, we have the option of altering the thresholds for the domains with the largest bias. We re-ran the cumulative selective editing results using slightly lower thresholds for domains 15 and 17. This had the desired effect of reducing the median absolute relative biases, to 1.0% and 0.9% respectively. The change in threshold means that slightly more businesses would need to be edited (although still many less than are edited with the current RSI edit rules). Figure 10 shows the edit failure rate and edit change rate for cumulative selective editing using the new thresholds for domains 15 and 17.

Figure 10: Edit failure rate and edit change rate for selective editing with scores calculated using previous selectively edited data and with revised thresholds



The increase in edit failure rate shown in figure 10 is noticeable when compared to figures 2 and 7. However, the increase is still small, especially when compared with the size of the saving made from using selective editing instead of the current RSI micro-editing rules.

19. The results above show that the conclusions of the initial RSI selective editing study are very slightly optimistic compared to what could be expected in practice. However, the cumulative impact of selective editing is minor, especially when compared to the current micro-editing rules. The slight increase in bias can easily be controlled by fine tuning the thresholds.

V. CONCLUSION AND FUTURE WORK

20. This paper has discussed work to improve the efficiency of ONS business surveys through developing a new selective editing strategy, which is used as a replacement for traditional micro-editing rules. The monthly Retail Sales Inquiry was chosen as the first survey to test this new method. The results show that savings of around 50% could be expected in the resource needed to edit the adjusted turnover variable. The impact of selective editing on RSI estimates should be insignificant, with the median relative absolute bias kept within 1% for all important published domains. The initial study did not consider the practical cumulative impact of selective editing, the fact that previous data used for calculating scores will contain many more unedited businesses so that the scores may be less accurate. However, a further study to take this into account showed that the true results of selective editing should be very similar to those originally reported, both in terms of savings and expected bias. For any domains where the bias becomes unacceptable, it is possible to rectify this by slightly lowering the selective editing threshold.

21. ONS has also recently reduced the resource needed for micro-editing in some surveys using a program called Snowdon which assesses the impact on survey estimates of changing thresholds for traditional edit rules. Because the Snowdon method relies only on small changes to existing rules, the savings which have been achieved up to now have been in the region of 5% to 10%. However, the Snowdon approach offers the possibility of savings without making major changes to the editing process. The indicators developed for Snowdon have also been useful in analysing the effect of selective editing on survey data. In practice it is useful to have both approaches to help improve the efficiency of micro-editing in ONS.

22. The RSI also collects information on employment every three months. Work is currently underway to develop a selective editing strategy that takes account of both adjusted turnover and employment. Following completion of selective editing methodology for RSI, work will begin to develop selective editing for the multivariate Annual Business Inquiry. This is likely to involve testing a range of alternative methods to cope with the complex nature of the survey. When these methods are finalised it should be possible to implement the new selective editing approach on a range of ONS business surveys.

REFERENCES

ABS (2007). The Editing Guide: the key to validating data from businesses. Canberra, Australian Bureau of Statistics.

Granquist, L. (1995). Improving the traditional editing process. Business survey methods. B. G. Cox, D. A. Binder, B. N. Chinnappa et al. New York, John Wiley & Sons Inc: 385-401.

Granquist, L. and J. G. Kovar (1997). Editing of survey data: how much is enough? Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins et al. New York, John Wiley & Sons Inc: 415-435.

Hedlin, D. (2001). Selective editing for business surveys. Southampton, University of Southampton: Technical report, 50p.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the UK Office for National Statistics. Journal of Official Statistics 19(2): 177-199.

Silva, P. L. N. (2009). Investigating selective editing ideas towards improving editing in the UK Retail Sales Inquiry. Internal report for ONS.

Silva, P. L. N., Bucknall, R., Zong, P. and Al-Hamad, A. (2008). A generic tool to assess the impact of changing edit rules in a business survey – an application to the UK Annual Business Inquiry part 2. Paper presented at the UNECE work session on Statistical Data Editing, Vienna, April 2008.

Zong, P. (2009). Analysis of selective editing for the Retail Sales Inquiry (RSI) based on domains. Internal ONS report.