2009 European Establishment Statistics Workshop – EESW09

Investigating selective editing ideas towards improving editing in the UK Retail Sales Inquiry

Pedro Luis do Nascimento Silva
Daniel Lewis
Alaa Al-Hamad
Ping Zong

1. INTRODUCTION

Data editing often takes up a large part of the resources spent on business surveys – Granquist (1995), Granquist & Kovar (1997). *Traditional micro-editing* for business surveys involves setting a number of edit rules to detect suspicious returned values / records. Flagged returns are then handled by staff looking for omissions, errors or surprising variations that may require data checking, respondent probing or explanation. In common with many statistical agencies, the UK Office for National Statistics (ONS) has been searching for ways to improve the efficiency of the editing process in its business surveys.

The UK Retail Sales Inquiry (RSI) is a typical ONS monthly survey of around 5,000 businesses providing measurements of retail turnover (monthly) and employment (quarterly). Current RSI editing involves a combination of *traditional micro-editing* based on a specified set of edit rules and some *macro-editing* performed by survey analysts who examine provisional survey estimates to check for remaining errors or surprising variations. As part of an agency-wide drive to improve editing performance (Black, 2009), this survey was targeted for potential redesign of its editing approach by applying *selective editing* methods – see for example Hedlin (2003, 2008), Luzi et al. (2007) and references cited therein.

This paper reports the findings of a study which investigated how alternative selective editing methods might be used in the RSI, estimated their likely impact on the key survey outcomes, and made recommendations about which methods should be used in a redesigned editing system for this survey. The study used data from 36 months of the RSI collected between January 2005 and December 2007. *Estimate-related scoring methods* (Hedlin, 2003) were considered, with several options for the item and unit level scores examined, for both cases when the targets of inference were the overall total or the domain totals.

Performance of various scoring methods was assessed by simulating their application to data from each of 35 months (February 2005 to December 2007), calculating a large set of impact measures for each method and each period, and summarizing these measures across all periods. A detailed account of the study is available in Silva (2009). The outcome of the study is clear: the RSI should benefit substantially from the application of selective editing methods to replace its current traditional micro-editing approach, both in terms of expected savings in the editing costs and of reduced burden on businesses, as well as for targeting editing efforts more effectively towards locating the largest and most influential errors.

2. SELECTIVE EDITING APPROACH CONSIDERED

Following Luzi et al (2008) we define *Selective Editing* as the approach which aims to split survey records into two streams labelled *critical* and *non-critical records*. *Critical records* (those expected to have an important impact on the final estimates) are submitted to any relevant edit rules and referred to reviewers for resolution of the suspected data quality issues. *Non-critical records* are submitted to a smaller set of edit rules (or even no edit rules), and any edit failures are either ignored or dealt with by automatic imputation. The goals of selective editing include reducing survey cost, processing time and respondent burden (by limiting re-contacts), and avoiding or reducing over-editing (after a certain point, editing can create as many errors as it removes) by recognizing that most errors have a small impact on estimates. It is also expected to help focus attention on records with the highest expected impact on estimates.

In selective editing, cases (records) are prioritized for editing (or follow-up) using *score functions*. There are two competing approaches to setting up score functions – see Hedlin (2003). *Estimate related score functions* are computed taking account of specified target survey estimates. *Edit related score functions* depend on a specified set of edits for a given survey. Score functions are calculated for each variable or item in a record, and subsequently aggregated to generate a *unit level score*.

In many business surveys, the key target parameters are population totals denoted $Y_j = \sum_{k\in U} y_{jk}$ which are often estimated by the weighted sample sums $\hat{Y}_j = \sum_{k\in s} w_k y_{jk}$, where y_{jk} is the 'true' survey measurement for variable j at unit k, w_k is a survey weight for unit k, U is the population and s is the sample.

The above notation carries an implicit assumption that the 'true' survey measurements can be obtained for all units included in the sample, which frequently is not possible for two main reasons: nonresponse and measurement error. Nonresponse is not considered here given our focus on editing the responses actually received in a survey. Berger (2009) investigated scoring methods for prioritizing chasing of non-respondents in a business survey context.

Measurement errors occur when the responses used for estimation do not correspond to the true values. Assuming full response, the absolute difference in the estimate $\hat{Y}_j$ due to using a reported value z_{jk} rather than the true value y_{jk} is given by $d_{jk} = w_k \left| z_{jk} - y_{jk} \right|$. Hence a frequently used score function for selective editing for variable j is given by

$$s_{jk} = w_k \left| z_{jk} - \hat{y}_{jk} \right| \Big/ \hat{T}_j \qquad (1)$$

where $\hat{y}_{jk}$ is a *predicted value* which replaces the unknown true value y_{jk}, and $\hat{T}_j$ is a standardizing factor, often the estimated total for variable j in a previous survey wave for the publication domain containing unit k.

Under an estimate-related selective editing approach, the key idea is to keep the amount of measurement error in the estimated total under control while still editing only a subset of the records / responses. For a survey collecting a single item (variable), the selective editing rule would then be to edit only records with values of s_{1k} larger than a specified threshold c_1. For surveys collecting several items (variables), item scores have to be combined to define a *unit level score*, which would then be used in a similar way to separate critical and non-critical records. Hedlin (2008) suggested a unified approach to define *unit level scores*. The idea is to view unit level scores as *Minkowski distance measures*, defined by

$$s_k(p) = \left(\sum_{j=1}^{J} s_{jk}^{p} \right)^{1/p} \qquad (2)$$

where $p \geq 1$ is a parameter used to specify the score function to be used. In this study, the values 1, 2 and 10 for p were tried for the unit level score.

Hedlin (2001, 2003) has already investigated selective editing ideas for ONS business surveys. Selective editing is currently used on a number of ONS business surveys in conjunction with traditional micro-editing rules. For these surveys, business returns are first submitted to the specified edit rules. Any business failing edit rules is edited only if it is judged to have an impact on published estimates. It is easy to see that this approach is subtly different from the approach considered in this study, where we closely followed the approach adopted at the Australian Bureau of Statistics (ABS), where all businesses are scored and only units having scores above a specified threshold are edited, with very little or even no use of traditional micro-editing rules – see ABS (2007).

For RSI the parameters of interest were Total Turnover (j=1) and Total Employment (j=2), either for the retail sector as a whole (overall estimates) or for some publication domains defined by grouping businesses in similar industry groups (17 domains were considered). The options for predictor values $\hat{y}_{jk}$ considered in this study are described in the sequence. First, the **previous value** (for turnover, this was the value from the previous month's survey, and for employment, the value from the last quarter) was defined as:

$$\hat{y}_{jk} = \begin{cases} y_{jk}^{t-1} \text{ if available for unit } k \\ y_{jk}^{0} \text{ otherwise} \end{cases} \tag{3}$$

where y_{jk}^{t-1} is the observed value in the previous month or quarter for unit *k* and y_{jk}^{0} is the register value (of turnover or employment, respectively) for the same unit.

The second option for the predictor value was simply the **register value**, namely $\hat{y}_{jk} = y_{jk}^{0}$. The third option was the **pseudo-imputed value**, obtained as:

$$\hat{y}_{jk} = \begin{cases} \hat{R}_j \times y_{jk}^{t-1} \text{ if available for unit } k \\ y_{jk}^{0} \text{ otherwise} \end{cases} \tag{4}$$

where $\hat{R}_j$ is the '*imputation link*' derived for variable *j* in period *t* using data from the responding businesses. These imputation links were provided by the ONS.

One potential drawback of this choice of predictor value is timeliness, because the scoring would need to wait for the calculation of the imputation links, which itself requires that at least part of the sample for the current month is available. The standard approach used by ONS to compensate for non-response is ratio imputation by multiplying a previous value by these imputation links, obtained using a trimmed mean of ratios of current and previous period returns. Therefore the outcome should be robust even if unedited data were used to perform the calculation each month prior to the start of the selective editing. In terms of this study, the imputation links available were calculated after the full sample was collected, so that it does not mimic exactly how this option would perform in the future.

The fourth and last option for the predictor value was to use **edited values for the current month**, namely $\hat{y}_{jk} = y_{jk}^{t}$, where the values y_{jk}^{t} correspond to the edited values available for the past waves of the survey. Following Hedlin (2003) this last choice of predictor was used only as a benchmark, since in practice it is not a feasible option because it depends on the 'true' or final edited values. However scores based on this option for the predictor provide a 'best possible' scenario on how selective editing methods based on estimate-related scores might perform, under the strong assumption that past editions of the survey achieved 'clean' responses for all units after performing their full editing procedures.

Silva (2009) and Zong (2009) examined whether the estimated totals $\hat{T}_j$ in the denominator of the scores in (1) should be calculated at the overall survey level or separately for each publication domain. They found that using totals calculated at the domain level is essential to keep bias under control for all of the publication domains.

Having calculated scores for each unit, it is necessary to define thresholds above which units are deemed to have failed selective editing and will therefore have their records edited and may be re-contacted by editing staff. Thresholds were set separately for each publication domain, such that the Absolute Relative Bias defined in (6) would be kept under control:

$$ARB\left(\hat{T}_{jd,se}\right) = 100 \times \sum_{k \in s_d} w_k \left| z_{jk} - y_{jk} \right| I\left[s_k(p) < c_d\right] \Big/ \hat{T}_{jd,cur} \tag{5}$$

where $\hat{T}_{jd,se}$ is the estimate of total for variable *j* in domain *d* after selective editing, s_d is the observed sample in domain *d*, $I\left[s_k(p) < c_d\right]$ is the indicator function taking value 1 when the unit level score $s_k(p)$ is less than the threshold c_d, so that the unit is not flagged for selective editing, and $\hat{T}_{jd,cur} = \sum_{k \in s_d} w_k y_{jk}$ is the estimate of total for variable *j* in domain *d* under the current editing procedure.

A range of values c_d were tested for each domain to find the threshold which resulted in the least amount of editing under the condition that the *ARB* should be less than 1% for the total turnover. Notice that the *ARB* defined in (5) is an upper limit on the absolute value of actual relative bias of the estimator of total after selective editing, defined as

$$RB\left(\hat{T}_{jd,se}\right) = 100 \times \left(\frac{\hat{T}_{jd,se} - \hat{T}_{jd,cur}}{\hat{T}_{jd,cur}} \right) \tag{6}$$

This follows because in practice it is likely that some of the errors not detected and corrected under selective editing could be in opposite directions, and would thus partly cancel out in (6), which is not the case when *ARB* is calculated using (5). With the bias this low, there should be no important differences between the estimates based on selective editing and the estimates obtained using the current RSI edit rules.

The impact of selective editing in the RSI was assessed using a number of quality indicators described in Silva (2009). The main indicators used include relative bias and absolute relative bias defined in (6) and (5) above, and the edit failure rate and edit hit rate, defined as follows:

$$\text{Edit failure rate} = 100 \times \frac{\text{Number of units failing at least one edit}}{\text{Number of responding units}} \tag{7}$$

$$\text{Edit hit rate} = 100 \times \frac{\text{Number of units with changes as a result of failing edits}}{\text{Number of responding units}} \tag{8}$$

3. RESULTS OF SELECTIVE EDITING STUDY FOR RSI

Analysis of the various performance indicators showed that the best performing option for predictor values was achieved by using **previous values** defined in (3). Amongst the options for the parameter p governing the unit level scores, larger savings were achieved by using p=10, but this choice was also the one leaving larger biases. In the following discussion the choice p=1 was adopted for being the most conservative (i.e. giving less bias).

One of the key questions in implementing selective editing is how much editing resource can be saved. Using the unit level scores with p=1 and the previous value as predictors of the true response, average savings of 55% on the current editing load could be expected at the end of quarter months, when both employment and turnover are investigated. For other months when only turnover is investigated, the average savings would be 74%.

The study showed that the current micro-editing rules have a high failure rate, always over 30% and often over 50% (especially for end of quarter months). If all of these businesses are re-contacted, it is clear that this will take up a large portion of the resources spent on the survey. It also showed that the current edit rules are fairly inefficient, with less than 10% of businesses having their turnover values changed as a result of editing.

Analysis of the bias resulting from selective editing showed that bias was fairly small for turnover across all domains, as expected. The median bias for all domains was very close to zero, and the bias in absolute value for the worst periods and the worst domains seldom exceeded 1%. For the employment total, bias seldom exceeded 2% in absolute value.

Because biases can be positive or negative, there is some degree of cancellation in the relative biases. The absolute relative biases give a better indication of the size of the maximum potential bias that could be expected from selective editing. The median *ARB* for total turnover in each domain is within the target of 1%, and very often is below 0.5%. Even for the worst months the *ARB* is below 2%. For the employment total the median *ARB* is always below 2%, with the worst periods reaching a maximum which is less than about 2.5%.

While the above figures may appear large, for some, the bias resulting from the current micro-editing rules is actually larger than that arising from using selective editing. Because selective editing focuses on those businesses which have the largest impact on the total estimates, it is also possible to improve on the accuracy of the traditional micro-editing currently used by the survey and make some substantial savings at the same time.

4. CONCLUSION AND FUTURE WORK

The results of the RSI selective editing study are very promising. The difference in edit failures between selective editing and the current edit rules translated into average savings of over 50% for almost all domains. At the same time, selective editing reduces the impact of the measurement errors on the final estimates when compared with the current micro-editing.

The bias results reported here did not consider the potential cumulative impact of selective editing due to the fact that previous data used for calculating scores are likely to contain more unedited values, leading less accurate scores. Lewis et al (2009) took this effect into account in calculations that showed that the expected impacts of selective editing should be very similar to those reported here, both in terms of savings and expected bias.

This paper has discussed work to improve the efficiency of ONS business surveys through applying a 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 approach. Using domain-specific threshold values proved important to keep bias under control for all publication domains. Our findings are robust in the sense that they illustrate how the procedure would operate for a large number of survey periods.

5. REFERENCES

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

Black, O. Improving validation for business surveys – the Eden Project. Paper presented to EESW09.

Berger, Y. G. (2009) Priority response chasing. Southampton: technical report.

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.

Hedlin, D. (2008). Local and global score functions in selective editing. Conference of European Statisticians, Work Session on Statistical Data Editing. Vienna, United nations statistical commission Economic Commission for Europe**:** 8.

Lewis, D., A. Al-Hamad, et al. (2009). Assessing the impact of selective editing on data quality. Conference of European Statisticians, Work Session on Data Editing. Neuchatel, Switzerland. To be presented.

Luzi, O., et al. (2007). Recommended practices for editing and imputation in cross-sectional business surveys - draft version 1.4, ISTAT Instituto Nazionale di Statistica, CBS Statistics Netherlands, SFSO Swiss Federal Statistical Office: 73.

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.