STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 14
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (ii): Measuring and evaluating data editing quality

SCORE FUNCTIONS TO REDUCE BUSINESS SURVEY EDITING AT THE ONS

Invited paper

Submitted by the University of Southampton, United Kingdom[1]

I. INTRODUCTION

1. As noted by Granquist and Kovar (1997) extensive micro-editing is not usually a good resource allocation, because of the costs involved and the limits to what it can achieve. In view of this, improving micro-editing essentially means making it less costly by checking and editing less. This paper analyses the use of score functions through a practical example, the Monthly Inquiry for the Distribution and Services Sector (the MIDSS) conducted by the Office for National Statistics (ONS) in the U.K. We are concerned with the efficacy of manual editing, as opposed to 'automatic editing', where the edit failures are followed up by automatic imputation. Prioritising units through some score function seemed to fit most readily into the existing organisation of the ONS. Instead of studying all aspects of quality, e.g., timeliness, the scope was limited to 'when the errors that would not have been corrected if a selective editing approach had been put in place do not affect the estimates of important target parameters'.

II. THE EDITING PROCESS AT THE ONS

2. Most business surveys at the ONS are mail-out/mail-back surveys. The scanned pictures of the returned questionnaires and the interpreted numbers are passed on to a division whose main task is editing and validation. The vast majority of the respondents that fail an edit will be called back. After having been checked, data are passed on to another division who will compute estimates, look at changes in estimates of domain totals and drill down. If a suspect unit is found in this secondary editing process, a query is sent back to the data validation division, where values from earlier periods of the survey will also be looked at this time and may be changed.

3. Raw and edited data were available for five periods of the MIDSS, November 1999 to March 2000. The raw data were captured immediately after the OCR was run, while the edited data were extracted after the editing and validation had been done, but before adjustments for non-standard reference periods and imputation. I refer to the difference between the raw and the edited value for one item as the ***change***.

4. The main variable of the MIDSS is turnover (total revenue from provision of goods and services, less trade discounts and taxes). Each quarter an employment item is added to the questionnaire. The target parameters of the MIDSS are domain totals and month-on-month changes in domain totals. About 25% of all MIDSS questionnaires triggers at least one edit failure for the turnover variable in November, January and February and about 45% for either turnover or employment in December and March. About 4% and 12%, respectively, of the units are actually changed. A problem with the evaluation of many micro-editing systems is the paucity of process data. Here I suggest some useful graphical analyses that require

[1] Prepared by Dan Hedlin, Department of Social Statistics, University of Southampton, U.K. (deh@socsci.soton.ac.uk).

only the availability of raw and edited data values, and the date when the record passed the editing system.

5. In Figure 1, values that passed through the editing process unchanged show up as the line through origin with slope 1. The confidentially rules of the U.K. Statistics of Trade Act do not allow us to show the scales of the axes. The other clearly visible line below, with the same slope, are instances where businesses have responded to the turnover item in actual pounds rather than in the requested £000. ONS corrects these errors automatically and £000 and larger errors were removed from the data available for the analyses reported here. The graph also shows several other, vaguer lines with the same slope. They are scanning errors where one or more digits are introduced or dropped by the OCR system. ONS is in the process of improving the reliability of the scanning.

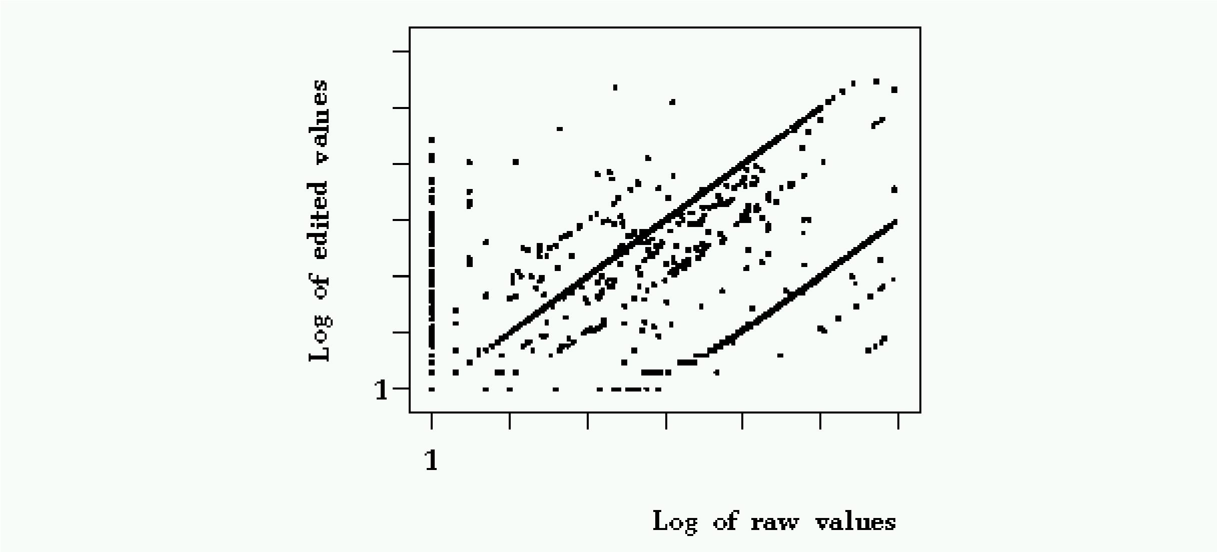

Figure 1. Logarithms of raw and edited turnover values with unity added. March 2000.

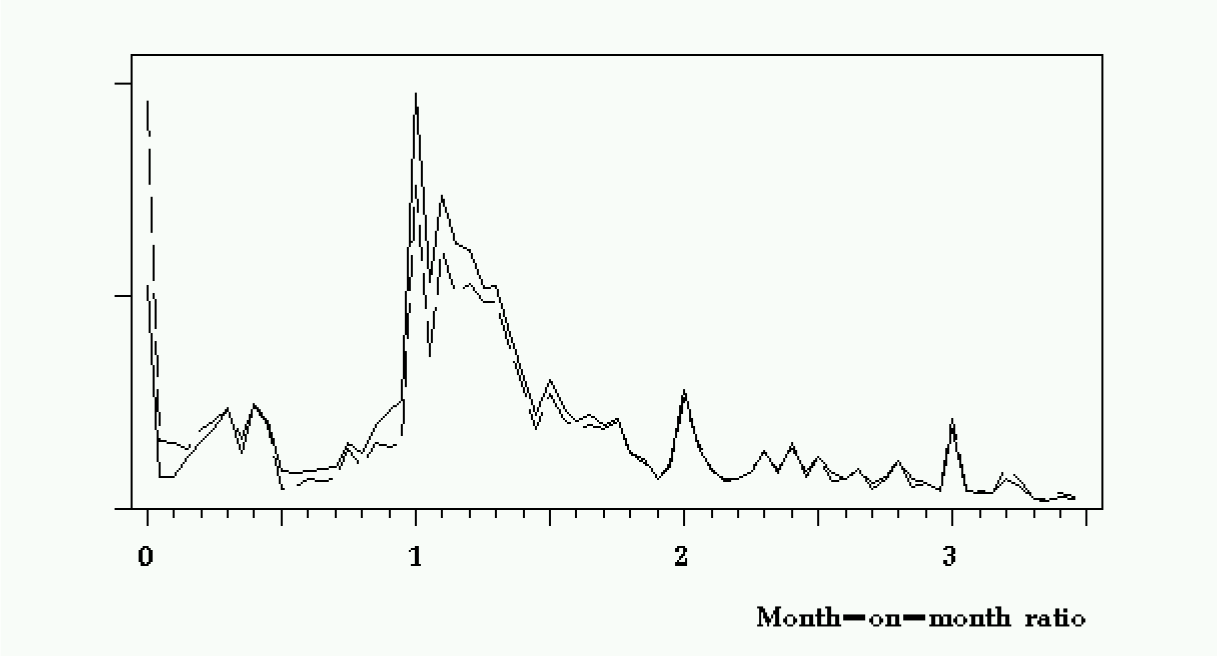

Figure 2. Individual month-on-month growth ratios for turnover, March on February. The solid curve is a frequency curve of ratios of edited values for March on edited values for February. The dashed curve is a frequency curve of ratios of raw values for March on edited values for Feb.

6. Figure 2 shows the distribution of month-on-month ratios for MIDSS turnover, March on February. Two ratios were computed for each unit that had failed an edit in March: one set of ratios for raw March values on edited February values, and one for edited March values on edited February values. For example, the small spike at 2 shows that some respondents have reported a March value that is twice as large as their February value. As expected there is a peak at unity. This and other smaller peaks at 'even' values such as 1.5, 2 and 3 indicate that many respondents think in terms of approximate growth *ratios* and work out the requested actual turnover based on the growth ratio. The most important edit for the MIDSS is a ratio edit where month-on-month ratios outside (0.5, 2) fail the edit. Figure 2 shows that one of the main editing processes is that of raw zero ratios (i.e. when the raw value is coded as zero and previous edited value is strictly positive) having been moved in to this interval. Apart from this, the shapes of the distributions are similar. The hump on the interval (0.1, 0.5) is striking. Why would 60% month-on-month reductions be more common than 50% ones? The heaping on even ratios, and perhaps also the hump, indicates measurement errors in the edited data.

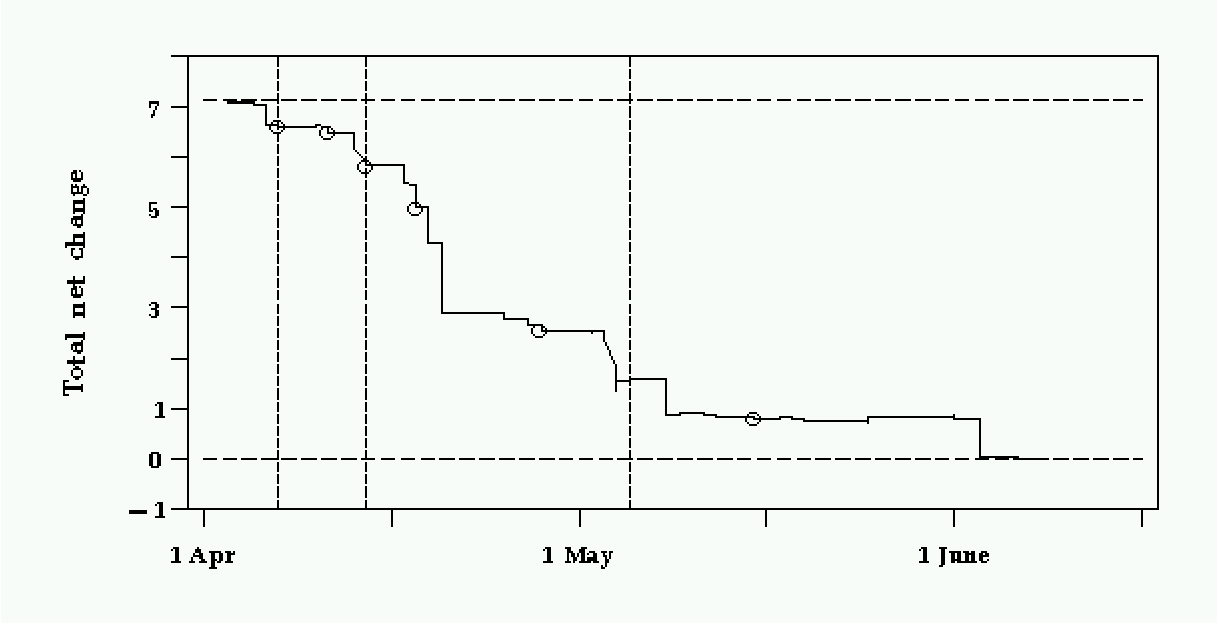

Figure 3. MIDSS March. The total net change of the turnover variable in £billion made during the editing process. Every 500th edit failure is circled up to the 3000th one. In all 3295 turnover values failed at least one edit. The vertical lines indicate the deadline for respondents, the start of intensive nonresponse follow-up, and delivery of estimates to users.

7. Figure 3 allows us to follow the actual editing process for turnover in March. The total net change, i.e. the sum of all changes, that were made to the variable, except for £000 and larger errors, was about £7 billion. The final estimated total was about £37 billion. The starting point along the y-axis is at the total net change and the end-point is zero, both indicated by horizontal lines. The first vertical curve is the respondents' deadline. After the second vertical line, which indicates the start of non-response follow-up, the curve falls steeper, thus showing larger changes to the data. Some editing is done long after delivery date, indicated by the rightmost vertical line. When estimates for April are sent to the customers (e.g. the National Accounts), revised March estimates which take most of the late changes into account will also be delivered. In theory, the total net change could have been nearly zero if changes of different signs cancelled out. However, the scanning errors visible in the lower right part of Figure 1 account for the vast majority of the large changes from large to smaller values, changes that dwarf other smaller changes of both directions. Plots for other months were similar and are not shown here.

III. TESTING SCORE FUNCTIONS ON THE MIDSS

8. The issues in this section are how to measure and report the efficacy of an editing method and how to set the threshold that determines how far the editing should go. Two types of score function were applied to the MIDSS data, an ***'estimate-related'*** and an ***'edit-related'*** function. The objective of the former is to predict the impact that a suspected error has on the estimates, see e.g. Lawrence and McDavitt (1994). The latter method orders suspected errors by size by using the distance from the

observation to the bulk of the data in some sense. In the estimate-related method an item score is calculated that represents the change in the estimate if a raw data value z_k for unit k is replaced with the edited value y_k. If the target parameter is the total and the estimator is the Horvitz-Thompson estimator $\hat{t}_y = \sum_{k=1}^{n} w_k y_k$, where n is the sample size and w_k is some weight for the *kth* unit, then the predicted absolute difference in the estimate caused by using a raw value z_k, which may or may not be suspect, rather than y_k, is

$$\hat{\delta}_{z_k} = w_k \left| z_k - \hat{y}_k \right| \ , \qquad (1)$$

where $\hat{y}_k$ is a prediction of y_k. For the purposes of computing an estimate-related item score, it is at least for sub-annual surveys often enough to predict y_k by simply using the most recent edited value from previous periods of the survey. If there is no such value, there may be a register value or an imputed value that could approximate y_k. Note that the $\hat{\delta}_{z_k}$ can be calculated for all raw values z_k, not only for those that fail an edit. Thus, the calculation of (1) is not dependent on the existing editing system. Another advantage is that the calculations are simple and explainable to non-statisticians. A potential problem is that the choice of score function depends on the estimator and the target parameter. An estimate-related method may therefore also need an additional score that, e.g., predicts the impact on estimates of change. Lawrence and McKenzie (2000) discuss this and other generalisations of (1).

9. A different idea is to put the selective editing on top of a micro-editing system and prioritise units by how many edits they fail and by 'how much' they fail; I refer to this technique as 'edit-related'. A distance from a datum point to each edit failure is calculated. For example, a ratio edit that fails all raw values that are more than twice as large as or less than half of the previous edited value can be represented as a cone in a scatter plot that displays new raw values against old edited values, with the edges of the cone spanning out from (0, 0) with slopes 2 and 1/2. The points outside the cone are edit failures. For each raw value there will be one magnitude of failure per edit the raw value has been subjected to. Earlier investigations showed that the existing edits could be simplified and still detect all non-trivial changes to the data. A raw value was classified as an edit failure if either the ratio of it to previous edited value was outside (0.5, 2); or if the ratio of the raw value to the corresponding frame variable was outside (0.4, 6) and (0.2, 2) for turnover and employment, respectively; or if the raw value was zero or missing. As a ratio edit measures relative movement it seems reasonable to take the angle between the line from the origin to the point representing the suspected value and the closest of the lines of the cone as a measure of magnitude of failure. For the third edit, $\hat{y}_k$ was taken as the magnitude of failure for the units that failed this edit. Since these magnitudes have variances of very different order and the first two are correlated, the Mahalanobis distance is a reasonable way of combining these to an item score. Let

$$d_{z_k} = \sqrt{\left(\mathbf{e}_k - \bar{\mathbf{e}}\right)' \mathbf{S}^{-1} \left(\mathbf{e}_k - \bar{\mathbf{e}}\right)}, \qquad (2)$$

be the Mahalanobis distance for a raw data value z_k, where $\mathbf{e}'_k = \left(e_{1k}, e_{2k}, e_{3k}\right)$ is the vector of magnitudes of failures for z_k, $\bar{\mathbf{e}}$ and $\mathbf{S}$ are the mean and the variance-covariance matrix of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$, m being the number of records that failed at least one edit. The data from current period were used for $\mathbf{e}_k$, but for $\bar{\mathbf{e}}$ and **S** I used data from the previous period only, even for units whose current period data were present they were processed (doing otherwise would have been impractical). The item score was

$$\hat{d}_{z_k} = \sqrt{\left(\mathbf{e}_k - \hat{\bar{\mathbf{e}}}\right)' \hat{\mathbf{S}}^{-1} \left(\mathbf{e}_k - \hat{\bar{\mathbf{e}}}\right)}, \qquad (3)$$

where $\hat{\bar{\mathbf{e}}}$ and $\hat{\mathbf{S}}$ denote approximations of $\bar{\mathbf{e}}$ and **S** obtained from the previous period. Alternatives to using all records from the previous period would be to use, for example, only data from the same industry or to pool data over several periods of the survey. Robust versions of $\hat{\bar{\mathbf{e}}}$ and $\hat{\mathbf{S}}$ could have been used (e.g. Little and Smith, 1987). The advantage of the Mahalanobis distance over the Euclidean distance is that the former is down-weighted by the variances and covariances in $\hat{\mathbf{S}}$; this shortens the distance for magnitudes of failures that showed large spread last period. This is desirable for the following and other reasons. Since a raw value that failed an edit with a large margin on a Euclidean scale will only be given a large distance on the Mahalanobis scale if this type of edit failure is rather atypical, edit failures that tend to occur together and hence tend to give redundant information will have relatively small Mahalanobis distances. Most score functions used in practice seem to be a mixture of edit-related and estimate-related methods. They seem to use an existing micro-editing system, on top of which an editing strategy similar to the estimate-related method has been put to reduce the amount of error signals.

10. For the MIDSS, four methods were compared:

A. The current micro-editing method.

B. The edit-related score function (3). Raw values that failed at least one edit in the current system but none of the three edits used for this experiment where given a zero score.

C. The estimate-related score function (1). The most recent study variable was used as the predicted value. If there was not any recorded study variable value for the unit, the corresponding frame variable (turnover or employment) was used, and if this one was missing, the score was set to missing.

D. Ideal micro-editing. Here the edited values are assumed known before the editing starts. The difference between this and method 3 is that here y_k is used instead of $\hat{y}_k$ in (1). The ideal micro-editing method is included here as a point of reference, a 'best possible' method for prioritising.

11. For a multipurpose survey, the ideal method would not target one particular estimate. The general idea is to prioritise in the best possible order given the edited values, which may be by the relative error size. An 'ideal' version of method B using $\bar{\mathbf{e}}$ and $\mathbf{S}$ instead of $\hat{\bar{\mathbf{e}}}$ and $\hat{\mathbf{S}}$ in (3) was also studied but the difference in outcome was small.

12. Figure 4 shows the total net change for Methods A – D for turnover for a domain here called 'domain 312'. Each curve is a scatter plots of the points $(i,\ \Delta_i)$, $i = 0, 1, 2, \ldots, m$, with interpolated straight lines between points, where m is the number of values that failed at least one edit and Δ_i is the weighted sum of outstanding changes when the first i edit failures have been attended to:

$$\Delta_i = \sum_{k=i+1}^{m} w_k \left(z_k - y_k\right). \qquad (4)$$

13. Note that if $i = m$ then all values are checked and $\Delta_i = 0$. Figure 4 shows a domain for which the selective editing approaches are not obviously superior (they were for most other domains). The current method signalled 125 out of 377 turnover values as suspect. The curve that represents the current method is ordered by the actual date when the editing took place. The other three curves are ordered by descending item score. The flat parts of the curves represent raw turnover values that failed at least one edit but were accepted as correct.

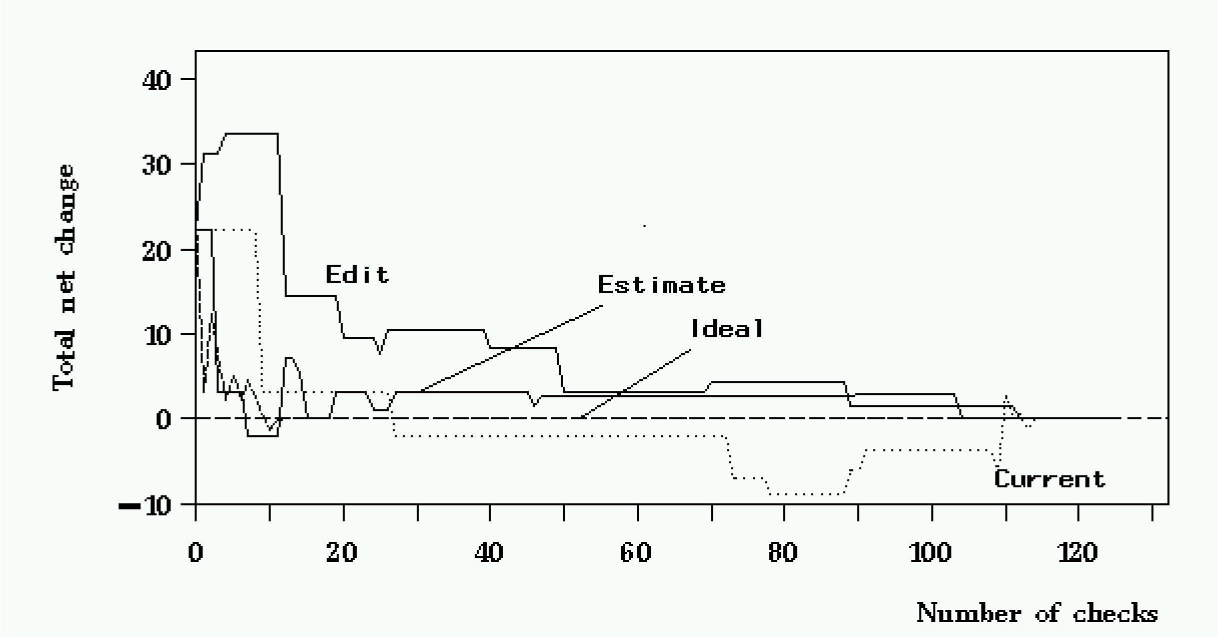

Figure 4. The total net change in £m in domain 312 against number of edit failures. Turnover, March 2000. Four editing methods: A) the current method, B) the edit-related method, C) the estimate-related and D) the ideal method.

14. Plots of $(i,\,\Gamma_i)$, where $\Gamma_i = \sqrt{\sum_{k=i+1}^{m} w_k\left(z_k - y_k\right)^2}$, may also be revealing. If Δ_i approaches zero quickly, the reason is either that the absolute values of changes associated with small scores are small or that they are large but with different signs and hence cancel out. In the latter situation selective editing methods may be unstable, in particular for domains with few changes.

15. To quantify the information in Figure 4, I isolate two components of an editing process: how well it detects errors, and, within the group of detected errors, how well it prioritises the errors. I relate the measure of editing effectiveness to the ideal editing method. As a measure of the first component, I suggest the ratio *a*/*b* where *a* is the number of edit failures for a certain item that the studied method need to go through in order to find all errors and *b* is the corresponding number for method D. For domain 312, Method C required that 104 edit failures were checked before the last one of all 13 changes was found, i.e. a ratio of 104/13=8. Table 1 shows the ratio (called measure 1) for each method. Methods B and C are similar in this respect but only slightly better than the current editing. Other domains showed a similar pattern.

Table 1. Two measures of editing efficacy for all domains for the MIDSS. Turnover, March. Measure 1 indicates error localisation capability, and Measure 2 how well the methods prioritise the errors.

			Measure 1			Measure 2		
Domain	Number of checks	Number of changes	A. Current method	B. Edit-related	C. Estimate-related	A. Current method	B. Edit-related	C. Estimate-related
312	125	13	8.8	8.6	8.0	0.8	0.8	1.0

16. For the second component, the sequence of found errors is extracted from the ideal method and the studied method, and some measure of correlation between the sequences, weighted with w_k, is computed. The beta coefficient in a robust regression relationship without intercept, with homoscedastic errors, and weights suggested by Welsch (1980) and analysed by Ryan (1997, Ch. 11) was calculated. Measure 2 in Table 1 refers the slopes for each editing method. Method C showed excellent performance for 312 (and other domains). This may not be surprising as the 'ideal' editing method D targets the total

and prioritises in a way similar to (1). To see if Method C would find errors in a more general sense, the ideal method was replaced with another 'ideal' editing method that first found the largest relative error, $|z_k - y_k|/y_k$, then the second largest one, and so on. Method C still performed very well, which confirms that for these data it found large errors effectively, not only those that impact on certain estimates. Recall that two of the edits used for Method B check relative movements and that Method C measures absolute movement. This seemed to be the main reason why Method B was slightly less efficient than C: for these data, absolute movement detected large errors better than relative movement. Note that this analysis can be applied to any selective editing method so long as the method explicitly orders the values by editing priority.

17. To see whether the difference between two estimates is 'negligible' or not, one can look at the coverage probabilities and see if they are 'nearly' the same under selective editing and current editing. The exact criterion used here was whether the difference between the two estimates was less than 10% of the standard error:

$$BR\left(\hat{\theta}_{proposed}\right) = \frac{\left|\hat{\theta}_{proposed} - \hat{\theta}_{current}\right|}{\sqrt{\hat{V}\left(\hat{\theta}_{current}\right)}} < 0.10 , \qquad (5)$$

with $\hat{\theta}_{current}$ and $\hat{\theta}_{proposed}$ being the estimate of a parameter θ under the current and proposed editing method, respectively, and $\hat{V}\left(\hat{\theta}_{current}\right)$ is the estimated variance of the total under the current method. The reason for setting the limit of $BR\left(\hat{\theta}_{proposed}\right)$ at 10% (an extremely conservative limit) is that if the values produced by the current method are regarded as the target that any proposed method should come close to, then BR statistic can be seen as a bias ratio, that is, a ratio of the bias of $\hat{\theta}_{proposed}$ to the estimator variance, which is related to the coverage probability in such a way that ratios smaller than 10% give negligible distortions of the coverage probability (see Särndal, Swensson and Wretman, 1992, p. 163-165). The criterion (5) served well as an objective rule that the ONS could agree on. I used the ratio estimator for $\hat{\theta}$ with the frame variable turnover as auxiliary, an estimator that is reasonably accurate.

18. After having prioritised the records to be edited, there must be some mechanism that classifies the score into, for example, two groups: 'needs editing' and 'does not need editing'. This can be done by predetermining a threshold above which data items will be selected for editing. A suitable threshold is the smallest score that will produce domain estimates that pass (5). To review and set thresholds adaptively, a ***progress graph*** can be produced regularly during the data collection period. Figure 5 gives progress graphs for domain 312 produced on April 15, 20 and 30. For example, on 15 April fourteen turnover values had been checked. The corresponding points given by (4) have been sorted by descending score and graphed. If the progress graph looks like any of the curves in Figure 5 the threshold may be increased for the rest of the data collection, as the editing of units with the smallest scores (but still above the original threshold) have not led to any significant change at all. If, on the other hand, the curve does not level out, the threshold may be lowered. If this is done early enough in the data collection period, there may be time to go back to questionnaires that fell short of the original threshold. Progress graphs give staff a means for control and hence may lead to greater staff acceptance. However, care must be taken so that small blips will not lead to premature changes of thresholds. Note that while the progress graph is persuasive, an outcome like that of Figure 5 does not prove that scores below the thresholds would not contribute significantly to the cumulative change had they been edited. The only way of estimating the contribution of small scores is to edit a random subsample below the threshold. The objective of the subsampling procedure is to estimate an upper bound for the net total of the errors that could have been corrected with more extensive micro-editing. This type of sampling and estimation problem has been studied in the context of auditing, that is, verification of financial accounts (Thompson, 1997).

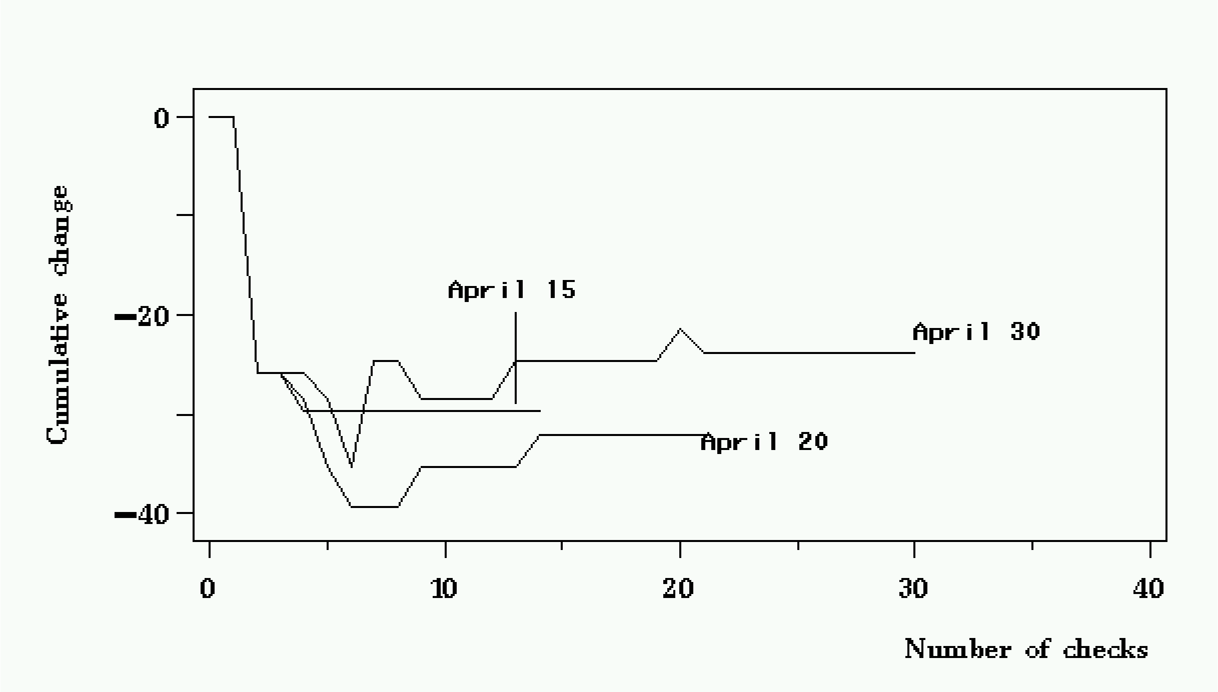

Figure 5. Three progress graphs for domain 312.

19. For the MIDSS the initial thresholds for both turnover and employment were set to 0.12% of the domain totals. It was the highest uniform threshold that made all domains pass (5) on both domain and overall level, this for all periods from December to March. With the 0.12% threshold Method C would have picked only 38 of the 125 edit failing values for domain 312. If the changes are interpreted as correction of errors, then Method C will in this case leave a total net error of about £5m uncorrected, 0.14% of the domain total. If a questionnaire is checked whenever either turnover or employment exceeds the threshold, then about 50% of the questionnaires that were actually checked in December and March would have been selected for editing. Errors that pass the bias ratio criterion (5) have no real impact on estimated totals but may have an impact on estimates of change. However, domain estimates of month-on-month change based on data obtained with Methods A and C, respectively, were very similar. In an additional test the frame variable turnover was used to predict the edited value instead of the previous edited turnover value. Although, the frame variable is often about 12 months old there was not much difference in outcome.

IV. DISCUSSION

20. Two score functions based on very different rationales have been studied. Both proved effective, but, somewhat surprisingly, the estimate-related method was seen to work better for a wide range of different estimates, apart from the estimates it targeted. One reason of the slightly less successful prioritisation of the edit-related method was the inefficiency of the underlying micro-editing system, which the estimate-related method does not depend on. It is possible that the edit-related method would work better for multi-purpose data with other edits. What tipped the balance, however, was rather the ease of implementation and understanding of the former method. The estimate-related method combined with the graphics shown here was a winning approach. It has now been implemented for the MIDSS and been subjected to pilot studies for several other business surveys at the ONS (Underwood 2001). About 50% of the editing effort could be spared for the MIDSS.

21. The most difficult problem with selective editing is to find a suitable threshold, above which units will be selected for editing. This paper advocates the use of a ‘progress graph’ that allows the monitoring of the impact of units just above the threshold. If the amount or structure of errors suddenly changes, for example due to some structural change that impacts on the respondents’ reporting capacity, the predetermined thresholds may not be appropriate. These changes should be reflected in the progress

graph. A subsample of records below the threshold could also be taken on a regular basis to estimate the total remaining error.

Acknowledgment: The author is grateful to several members of staff at the U.K. Office for National Statistics for valuable comments and for providing data.

References

Granquist, L. and Kovar, J.G. (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: Wiley, 415-435.

Latouche, M. and Berthelot, J.M. (1992). Use of a score function to prioritise and limit recontacts in business surveys. Journal of Official Statistics, 8, 389-400.

Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. Journal of Official Statistics, 10, 437-447.

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

Little, R.J.A., and Smith, P.J. (1987). Editing and Imputation for Quantitative Survey Data. Journal of the American Statistical Association, 82, 56-68.

Ryan, T.P. (1997). Modern Regression Methods. New York: Wiley.

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

Thompson, M.E. (1997). Theory of Sample Surveys. London: Chapman & Hall.

Underwood, C. (2001). Implementing Selective Editing in a Monthly Business Survey. Paper presented at the Sixth Government Statistical Service Methodological Conference, London, 25 June.

Welsch, R.E. (1980). Regression Sensitivity Analysis and Bounded-Influence Estimation. In Evaluation of Econometric Models, eds. J. Kmenta and J.B. Ramsey. New York: Academic Press, 153-167.