Journal of Official Statistics, Vol. 19, No. 2, 2003, pp. 177–199

Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics

Dan Hedlin[1]

A score function associates a number to each item response. The function indicates the relative importance of allocating manual resources to review responses, thereby allowing the survey analyst to prioritise editing efforts. By applying different score functions to the Monthly Inquiry for the Distribution and Services Sector the article discusses how to measure the effectiveness of a score function, and what it is that makes this method effective at reducing manual editing. One finding is that the effectiveness of the technique is rather insensitive to the type of score function. In many situations it is useful to specify a threshold that splits the responses into two groups where manual editing is directed at responses with scores above the threshold. The article suggests a simple graphical tool that allows the analyst to assess the threshold.

Key words: Selective editing; estimate-related; edit-related; threshold; progress graph.

1. Introduction

The act of checking and correcting respondent data in surveys is usually referred to as editing. Many national statistical institutes use a ‘‘micro editing’’ approach for business surveys. Micro editing (also called input editing) focuses on the individual record or questionnaire, as opposed to macro editing where checks are used on aggregated data. In micro editing, the respondent data are passed through *edits* (edit rules, checks) that typically aim at detecting unusual item responses. For example, many repetitive business surveys use ratio edits whereby a response from a business is compared to its prior response. If the relative movement is more than, say, *a%* or less than *b%*, the incoming datum point fails the edit, and the questionnaire will be inspected manually. The business may be called back. As business data are volatile, a sizeable proportion of the respondents will confirm reported large movements. Thus, micro editing will lead to many false signals unless the edits have been carefully designed (Thompson and Sigman 1999).

Since editing is one of the most time-consuming processes in the production of official statistics (Granquist and Kovar 1997, p. 418), more efficient methods have been discussed and implemented for many surveys. With selective editing the incoming units are prioritised, and those that have been given priority are selected for editing. The prioritisation step often involves the computation of a score for each datum point that reflects the

[1] University of Southampton, Department of Social Statistics, Highfield, Southampton, U.K. Email: deh@socsci.soton.ac.uk

Acknowledgments: The author is grateful to several members of staff at the U.K. Office for National Statistics for valuable comments and for providing data. I am also indebted to Pascal Rivière and several other colleagues, and also to the associate editor and the referees for making many suggestions for improvement.

importance of investigating this datum point. Some types of score can be computed for all data, others only for those that have failed at least one edit (see Section 3). A score may be computed for each item on the questionnaire, and the item scores may be combined to a unit score. Questionnaires with unit scores above a predetermined threshold are inspected manually; other units are left unattended or passed on to another process, for example some other type of editing or automatic imputation. Alternatively, there may be more than two levels of priority, with a high threshold defining the highest priority, and so on. The set of scores for one item can be viewed as the range of a *score function* of the unedited data (the *raw values*) and the background data, such as past edited data. Granquist and Kovar (1997) give an overview of selective editing and further references.

There are other ways of reducing measurement error than micro editing. The crucial role of questionnaire design cannot be overstated. Editing is fixing errors after they have occurred; obviously, it is better to try to prevent errors from occurring in the first place. Dippo, Young, and Sander (1995), Paxson, Dillman, and Tarnai (1995) and Dillman (2000) discuss questionnaire design for business surveys. Incorrect data that are not obviously different from correct values are sometimes referred to as inliers (as opposed to outliers), for example systematic errors that many respondents make repeatedly. Inliers cannot usually be detected with micro editing (DesJardin and Winkler 2000). Other crucial measurement error and related issues that are not addressed by micro editing include ''profiling,'' which refers to the process of delineating large businesses with complex structures (Pietsch 1995), and classification of activities, which determines what domains a sampled business will contribute to (Nijhowne 1995). As noted by Granquist and Kovar (1997), extensive micro editing is rarely a good resource allocation, because of the costs involved and the limits to what it can achieve. Discussions of survey process improvement and efficient resource allocation in a wider perspective are to be found in Linacre and Trewin (1993) and Lyberg et al. (1997). Improving micro editing essentially means making it less costly by editing less than traditionally has been the case for business surveys while maintaining quality. Granquist (1998) argues that cutting the traditional editing will most likely *improve* quality, if resources are sensibly reprioritised from micro editing to other means of quality enhancement. We are concerned with the efficacy of manual editing, as opposed to ''automatic editing,'' where the edit failures are followed up by automatic imputation. See Fellegi and Holt (1976) and Little and Smith (1987) for this strand of research.

There are other viable approaches to reducing editing. For example, a reduction of sample sizes in conjunction with more efficient estimation reduces the total amount of editing and other post-editing survey operations.

The aim of an extensive project at the Office for National Statistics (ONS) in the U.K. was to find substantial savings in the data collection and editing processes for business surveys without adversely impacting on quality. One task was to explore selective editing. Prioritising units through some score function seemed to fit most readily into the existing organisation of the editing process. In this article I analyse the use of score functions through a practical example, the Monthly Inquiry for the Distribution and Services Sector (MIDSS) and discuss extensions to other business surveys at 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.'' There are also indirect effects on quality, including the fact that a more efficient editing process would reduce the response burden, which may have considerable quality effects in the long run.

Section 2 introduces the MIDSS and suggests some graphs that will help to understand an editing process. How to set the threshold for the selection step and how to compare and evaluate different selective editing methods is discussed in Section 3. In Section 4 it is explored under what circumstances it can be expected that selective editing will be more efficient than the complete follow-up required by conventional micro editing. The article concludes with a discussion in Section 5.

2. The Editing Process at the ONS

2.1. Background

Most business surveys at the ONS are traditional 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 are called back. There are typically many operators involved in each survey, so the effect of measurement errors due to correlation between decisions taken by the same operator may be expected to be small. The revised data are passed on to another division that computes estimates, compares past and present domain totals and investigates conspicuous changes. If a suspect unit is found in this secondary editing process, a query is sent back to the data validation division, where back data of the survey will also be examined this time and may be changed.

Raw and edited data were available for five periods of the MIDSS, November 1999 to March 2000. There were about 11,000 responding businesses per month. 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 nonstandard reference periods and imputation. I refer to the difference between the raw and the edited value for one item as the *change*.

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. The design is stratified simple random sampling. There are more than 20 edits for turnover and/or employment. As seen in Table 1, a very large

Table 1. Fail and change rates for the MIDSS. The proportion of questionnaires that failed at least one edit (first column) and the proportion of all questionnaires where any variable value was changed as a result of the editing (second column)

	Fail rate (% of questionnaires)	Changed (% of questionnaires)
Nov 1999	24	4
Dec 1999	45	12
Jan 2000	26	4
Feb 2000	26	3
Mar 2000	45	11

proportion of all MIDSS questionnaires trigger at least one edit failure for the turnover variable in November, January, and February or for either turnover or employment in December and March. Despite this, few units are actually changed.

2.2. *Graphing the current editing process*

A problem with the evaluation of many micro editing systems is the paucity of process data (Granquist 1995). Here are some useful graphical analyses that require only the availability of raw and edited data values, and the date when the record passed through the editing system.

Figure 1 shows simulated raw and edited MIDSS turnover data for March on a logscale, base 10, with unity added to each turnover value. The U.K. Statistics of Trade Act does not allow us to display the real data. Values unchanged in the editing process show up as the line through the origin with slope 1. Below that line, there is another clearly visible line with the same slope. These are instances where businesses have responded in actual pounds rather than in the requested £000. These errors are corrected automatically (Underwood, Small, and Thomas 2000); and £000 and larger errors were removed from the data available for the analyses reported here. There are also several other, less distinct lines with the same slope. Most of these are scanning errors where one or more digits are introduced or dropped by the OCR system. For March 2000, 4% (130 out of 3,300) of all raw turnover values that failed an edit had one more or one less digit than the corresponding edited values and all other digits were the same: they are most likely scanning errors. The ONS is in the process of improving the reliability of the scanning. Sometimes units report annual rather than the requested monthly turnover. Other errors are inaccurate reporting detected mainly through the comparison with the previous response or the frame

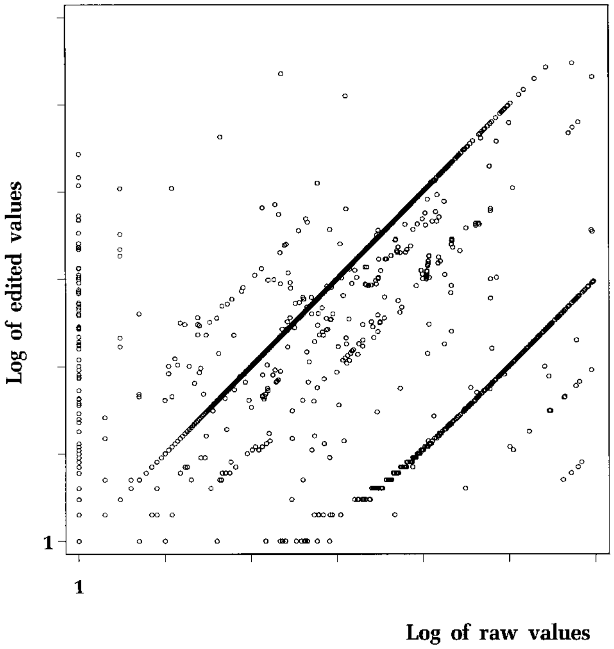

Fig. 1. Logarithms of simulated raw and edited MIDSS turnover values with unity added

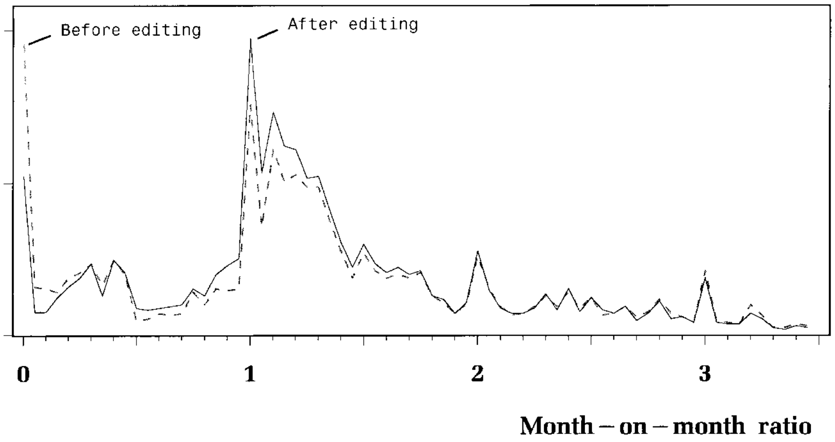

Fig. 2. Individual month-on-month growth ratios for turnover, March on February. Frequency curves for ratios of raw values for March to edited values for February (before editing), and for ratios of edited values for March to edited values for February (after editing)

turnover value. Figure 1 also displays zeroes or missing raw values that were coded as zeroes and were changed to positive values in the editing process. Note that the changes that are made to the data in the editing process form a highly skewed distribution even without the £000 and larger errors. MIDSS data for other months were very similar, and the general data structure was the same as the one exhibited in Figure 1.

Figure 2 shows the distribution of month-on-month ratios for MIDSS turnover, March on February. Two ratios were computed for each turnover value 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 implies that some respondents reported a March value twice as large as their February value. The peak at unity and other small peaks at ''even'' values such as 1.5 and 2 indicate that many respondents think in terms of approximate growth *ratios* and work out the requested actual turnover from the growth ratio. As there were 10% more weekdays in March 2000 than in February of the same year, and 12% more weekdays and Saturdays, we might expect the mode of the growth ratios in Figure 2 to be slightly above 1. Indeed, there is a small peak at 1.10. 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 makes one of the main editing processes visible: that of raw zero ratios (i.e., when the raw value is coded as zero and the previous edited value is strictly positive) having been moved into the interval (0.5, 2). Apart from this, the shapes of the distributions are similar. The hump on the interval (0.1, 0.5) is striking. So while few respondents reported March values that were 0.1 or 0.5 times their February values, relatively many had March to February ratios in between. One would have expected a ''regular'' appearance, for example the curve being monotonous on the interval (0.1, 1). Why would 60% month-on-month reductions be more common than 50% ones? This hump and the other features mentioned showed up in all months for which data were available. The heaping on even ratios, and perhaps also the hump, indicates measurement errors in the edited data.

Figure 3 displays the actual editing process for turnover in March. The total net change,

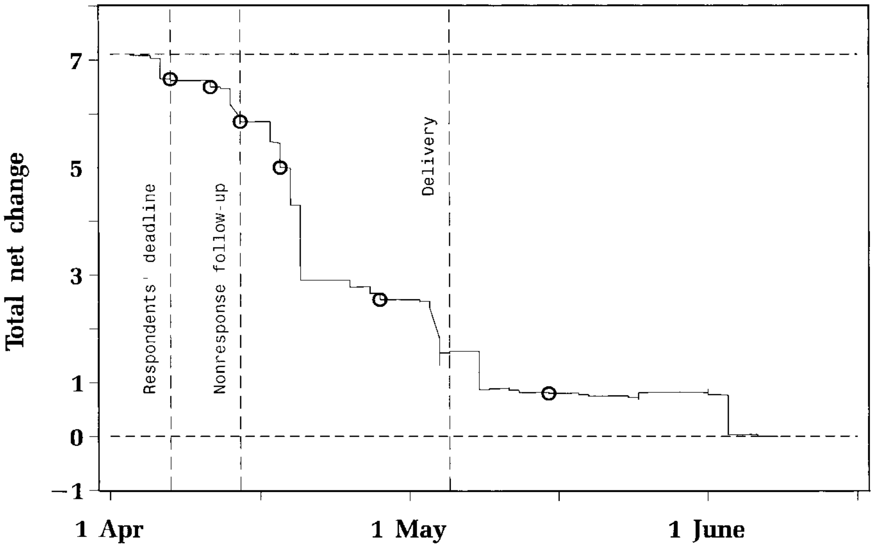

Fig. 3. The total net change of the turnover variable in £billion made during the editing process. Every 500th edit failure is circled up to the 3,000th one. In all, 3,295 turnover values failed at least one edit. The MIDSS, March 2000

i.e., the sum of all changes that were made to the turnover 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 the total net change and the end-point is zero, both indicated by dashed horizontal lines. The curve falls steeper from the start of intensive nonresponse follow-up, thus showing larger changes to the data. Some editing is done long after delivery date. When estimates for March are sent to the customers at the beginning of May (e.g., the National Accounts), revised February estimates, which take most of the late changes into account are also delivered. Graphs for other months were similar. 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 smaller changes in both directions.

3. Testing Alternative Score Functions on the MIDSS

The issues in this section are how to measure and report the efficacy of a selective editing method and how to set the threshold that determines how far the editing should go. I consider two types of score function, an ''*estimate-related*'' and an ''*edit-related*'' function. The objective of the estimate-related method is to predict the effect that a suspected error has on the estimates. 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. Consider the total of $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$ over N population units, and a widely used class of estimators of the total,

$$\hat{t}_y = \sum_{k \in sample} w_k y_k \qquad (1)$$

where w_k is some weight for the *k th* unit. 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

$$\hat{\delta}_{z_k} = w_k|z_k - \hat{y}_k| \tag{2}$$

where $\hat{y}_k$ is a prediction of y_k. The DIFF function of Latouche and Berthelot (1992) is similar to (2). For the purposes of the estimate-related item score (2), 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. This will be discussed further in Section 4. 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 (2) does not depend on the existing editing system. Another advantage is that the calculations are simple and explainable to nonstatisticians. 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 for example predicts the effect on estimates of change. Lawrence and McDavitt (1994) and Lawrence and McKenzie (2000) discuss generalisations of (2).

A different idea is to put selective editing on top of a micro editing system and prioritise raw values 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, so there will be one *magnitude of failure* per edit the raw value has been subjected to. 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, as in Figure 4, where

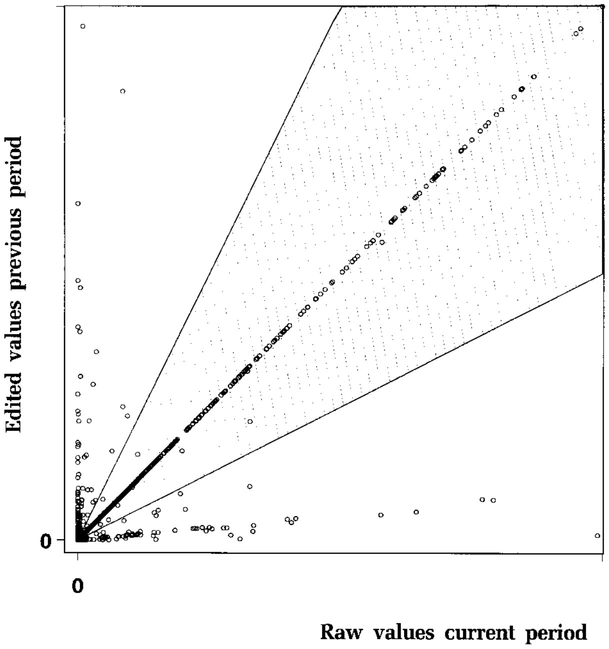

Fig. 4. A graphical representation of a ratio edit. Edited values previous period against raw values current period, whereby raw values larger than twice the previous edited value or smaller than half the previous edited value fail the edit. Points inside the cone pass the edit

points outside the cone are edit failures. Earlier investigations showed that a subset of the 20 specified edits found all changes that the full set of edits detected. A raw value was classified as an edit failure if either the ratio of it to the most recent edited value was outside (0.5, 2); or if the ratio of it 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 as a measure of magnitude of failure the angle between the line from the origin to the point representing the suspected value and the closest of the lines of the cone associated with the ratio edit. For the third edit, $\hat{y}_k$ (turnover or employment) was taken as the magnitude of failure for edit-failing responses. Since these magnitudes have variances of very different order and the first two are correlated, the Mahalanobis distance (e.g., Krzanowski 1990) is a reasonable way of combining these to an item score. Let

$$d_{z_k} = \sqrt{(\mathbf{e}_k - \bar{\mathbf{e}})' \mathbf{S}^{-1} (\mathbf{e}_k - \bar{\mathbf{e}})} \tag{3}$$

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

$$\hat{d}_{z_k} = \sqrt{(\mathbf{e}_k - \hat{\bar{\mathbf{e}}})' \hat{\mathbf{S}}^{-1} (\mathbf{e}_k - \hat{\bar{\mathbf{e}}})} \tag{4}$$

where $\hat{\bar{\mathbf{e}}}$ and $\hat{\mathbf{S}}$ denote approximations of $\bar{\mathbf{e}}$ and $\mathbf{S}$ obtained from the previous period. Alternatives to basing the computation of $\hat{\bar{\mathbf{e}}}$ and $\hat{\mathbf{S}}$ on all records from the previous period would have been to use, for example, only data from the same industry or to pool data over several periods of the survey. Also, robust versions of $\hat{\bar{\mathbf{e}}}$ and $\hat{\mathbf{S}}$ could have been used (e.g., Little and Smith 1987). As a general example of the advantage of the Mahalanobis distance over the Euclidean distance $[(\mathbf{e}_k - \hat{\bar{\mathbf{e}}})'(\mathbf{e}_k - \hat{\bar{\mathbf{e}}})]^{0.5}$, consider a raw value that fails several edits. Then the Mahalanobis distance will be shorter than the Euclidean distance if the covariance between the magnitudes of failures is large, i.e., if these edit failures tend to occur together, in which case they tend to give redundant information. Hence the edit-related method has the attractive property of assigning a value to the edits that gives a measure of the importance of the edit failures taken together. Edit-related methods have the additional advantage of being more general than estimate-related methods as they do not target one particular estimate. Only a subroutine needs to be added to a micro editing system, although the inverse of $\mathbf{S}$ must be computed. However, the measure of importance may be poor if the $\mathbf{e}_k$ do not offer enough predictive power in terms of likelihood and size of errors in the raw data values. Hence, unlike the estimate-related method, the edit-related method has the disadvantage of depending on the specified edits.

Most score functions used in practice 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 (Latouche and Berthelot 1992; Lawrence and McDavitt 1994; Granquist and Kovar 1997; UN 1997; and Lawrence and McKenzie 2000).

For the MIDSS, four methods were compared:

A. The current micro editing method.
B. The edit-related score function (4).
C. The estimate-related score function (2). The most recent study variable was used as the predicted value. If there was no recorded study variable value for the unit, the corresponding frame variable (turnover or employment) was used, and if even this one was missing, the score was set to missing. The weights w_k were those of a model-assisted combined ratio estimator for a stratified simple random sample (Särndal, Swensson, and Wretman 1992, Ch. 7).
D. Ideal micro editing. Here the edited values are assumed known before the editing starts. The difference between this and method C is that y_k is used instead of $\hat{y}_k$ in (2). The ideal micro editing method is included here as a point of reference, a ''best possible'' method for prioritising.

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 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 (4) was also studied but the difference in outcome was small.

The selective editing approaches B–D were simulated with the raw and edited MIDSS data that had failed at least one edit in the current system. The current edit failures were taken to define what values were suspect and the edited values were regarded as true values. While these are assumptions used in all the studies on selective editing referred to above, it is quite possible, for example, that with selective editing time pressure can be reduced and therefore the edited values may be more accurate than with current editing. Recall that Figure 2 indicated the presence of measurement errors.

Figures 5 and 6 show the total net change for Methods A–D for turnover for two domains. They are scatterplots of the points (i, Δ_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:

$$\Delta_i = \sum_{k=i+1}^{m} w_k(z_k - y_k) \tag{5}$$

In Figure 5, when the 375 businesses that failed at least one edit for turnover had been looked at (33% of a total of 1,137 responding businesses in the domain), the current method had reduced $\Delta_0 \approx$ £200m to $\Delta_{375} = 0$. The edit-failing values corresponding to the curve that represents the current method are ordered by the actual date when the editing took place, with $i = 1$ being the earliest verification. The other three curves are ordered by descending item score, i.e., by descending editing priority. With Methods B and C, $|\Delta_i| < 0.02\%$ of the domain total for $i > 150$. The flat parts of the curves represent raw values that failed at least one edit but were accepted as correct. It is the asymmetry observed in Figure 1 and the effective prioritisation that is reflected in the nearly convex shape of three of the functions in Figure 5.

Figure 6 shows domain 312 for which the selective editing approaches are not obviously superior. The current method signalled 125 out of 377 turnover values as suspect.

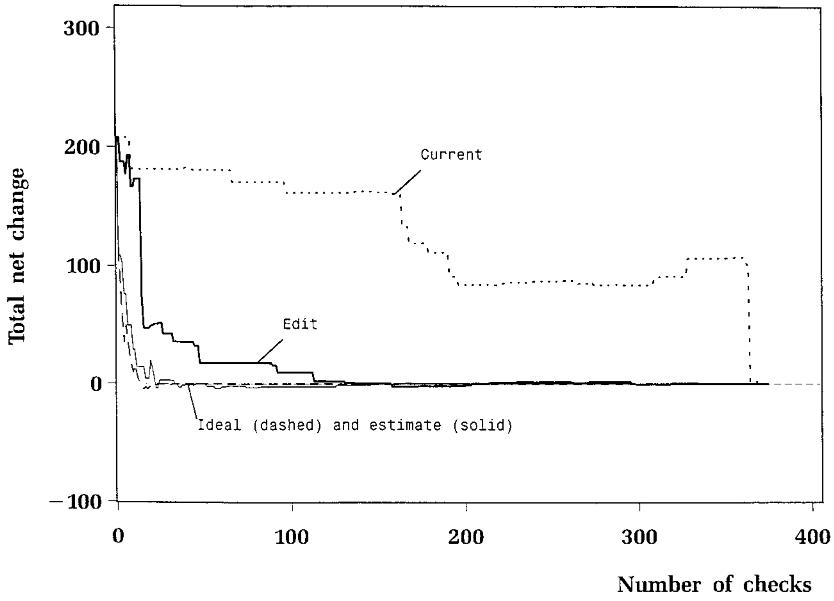

Fig. 5. The total net change in £m in domain 333 against number of verified edit-failing values. The current, edit-related, estimate-related and ideal editing method. Turnover, March 2000

However, both Figures 5 and 6, and similar plots for other domains, do show that methods B and C can successfully prioritise suspect values.

Plots of (i, Γ_i), where $\Gamma_i = \sqrt{\sum_{k=i+1}^{m} w_k(z_k - y_k)^2}$, may also be revealing. Even if $\Delta_i \approx 0$ for some i, there may be several $k > i$ with large errors $z_k - y_k$ in absolute terms that have cancelled out due to different signs. If so, selective editing methods are unstable,

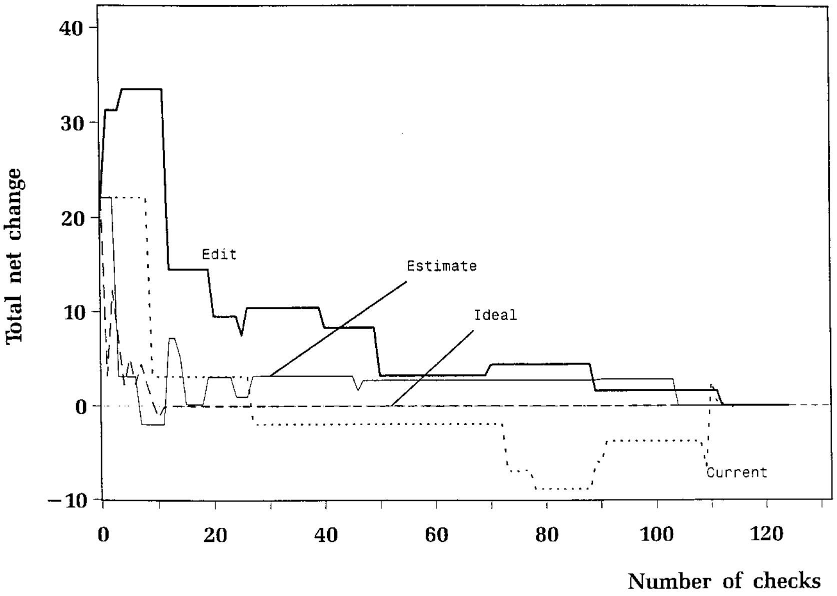

Fig. 6. The total net change in £m in domain 312 against number of verified edit-failing values. The current, edit-related, estimate-related and ideal editing method. Turnover, March 2000

in particular for domains with few changes. We observed substantial cancelling out only for the current method in domain 312.

3.1. A measure of prioritisation efficacy

Although Figures 5 and 6 give a graphical display of the effectiveness of editing methods, we need also some way of quantifying the information in these graphs. To do this, I isolate two components of an editing process: how well it detects errors regardless of size, and, within the group of detected errors, how well it prioritises the errors by size; both measures are compared 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 needs to go through in order to find all errors, and *b* is the corresponding number for the ideal Method D. For example, in domain 312 Method C required 104 edit failure investigations to find all 13 changes, i.e., a ratio of 8. Table 2 reports the ratio (measure 1) for each domain and method. Methods B and C are similar in this respect but only slightly better than the current editing.

For the second component, I suggest that the sequence of detected changes be extracted from the ideal method and the studied method, and some measure of correlation between the sequences, weighted with the corresponding case weight w_k, be computed. For example, Table 3 provides the sequences for the ideal and estimate-related methods for domain 312. I use the beta coefficient in a regression relationship without intercept, with homoscedastic errors, and regression weights suggested by Welsch (1980) and analysed by Ryan (1997, Ch. 11). With this bounded-influence estimation approach, observations with large influence are down-weighted. The regression weight for the pair k in the sequences of edit changes is

$$c_k = \begin{cases} 1 \text{ if } |DFFITS_k| \le 2n^{-0.5} \\ 2n^{-0.5}|DFFITS_k|^{-1} \text{ if } |DFFITS_k| > 2n^{-0.5} \end{cases} \tag{6}$$

where n is the number of observations in the sequences, and the $DFFITS_k$ is a well-known regression diagnostic that measures how much a prediction of the dependent variable for this observation's value of the independent variable would change in terms of standard deviations of the predicted value if the regression line is refitted without observation k. Measure 2 in Table 2 refers to the slopes for each domain and editing method. Now Method C shows excellent performance throughout. This may not be surprising as the ''ideal'' editing Method D targets the total and prioritises in a way similar to (2). To see if Method C would find errors in a more general sense, Method D was replaced by another ''ideal'' editing method with the kth score equal to $w_k|z_k - y_k|/y_k$. Method C still performed very well in terms of Measure 2 with upper quartile, median and lower quartile slopes being (0.99, 0.9, 0.6) and (0.97, 0.7, 0.3) for Methods C and B, respectively. 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 prioritised large errors better than relative movement did.

Note that the analyses whose results are reported in Figures 5 and 6 and Table 2 can be applied to any selective editing method as long as the method explicitly orders the values by editing priority.

Table 2. Two measures of editing efficacy for all domains in the MIDSS. Turnover, March. Measure 1 indicates error localisation capability, and Measure 2 indicates how well the methods prioritise the errors. The rows are sorted by the number of turnover values that were manually investigated

			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
333	375	39	9.6	9.3	9.1	0.4	0.8	0.9
328	347	49	7.0	7.0	5.4	0.3	0.2	1.0
326	241	21	11.4	11.4	8.7	0.2	0.9	1.0
327	220	30	7.0	4.8	6.5	0.4	0.4	1.0
331	197	31	6.0	6.0	6.0	0.3	0.6	1.0
324	196	22	8.9	8.2	7.7	0.0	1.2	1.0
335	190	30	6.3	5.9	6.3	0.1	0.5	1.1
312	125	13	8.8	8.6	8.0	0.8	0.8	1.0
309	106	14	7.3	6.6	5.9	0.2	0.7	1.0
323	100	12	8.0	5.1	6.6	0.3	0.4	1.0
332	92	12	7.3	7.1	5.3	0.0	1.2	1.0
307	75	17	4.0	4.4	4.4	0.2	0.7	1.0
315	75	7	9.3	10.7	8.1	0.4	0.8	1.0
316	71	14	4.9	2.6	4.7	0.2	0.9	1.0
300	68	10	5.1	6.5	5.5	0.6	0.8	1.0
302	62	10	5.9	2.4	4.9	0.0	1.0	1.0
310	61	9	6.0	2.8	4.3	0.5	0.6	1.0
311	60	8	6.8	3.8	4.6	0.2	0.9	1.0
330	60	8	7.1	7.5	7.5	0.0	1.0	1.0
301	57	6	7.0	3.8	6.2	0.1	1.0	1.0
308	49	3	10.7	11.0	10.3	0.1	0.1	1.0
313	49	7	5.9	1.9	2.0	0.0	1.0	1.0
319	45	7	6.3	5.6	4.1	0.4	1.0	0.9
304	41	6	5.7	4.2	3.3	0.4	0.9	1.0
317	41	5	8.0	4.2	1.6	0.4	1.0	1.0
305	38	2	6.0	16.5	14.0	0.4	0.4	1.0
318	38	7	4.7	3.1	3.1	0.4	0.9	1.0
321	38	4	6.3	5.5	1.8	0.7	1.0	1.0
306	37	5	5.6	4.6	5.2	0.3	2.6	1.0
314	28	5	4.6	1.8	3.0	0.0	1.8	1.0
329	27	2	10.0	7.0	12.0	1.0	1.0	1.0
325	26	5	4.6	4.8	5.2	0.7	1.0	1.0
320	19	2	6.0	1.5	2.0	1.0	1.0	1.0
303	18	4	4.3	1.8	3.0	1.0	1.0	1.0
334	15	1	8.0	5.0	8.0	1.0	1.0	1.0
322	8	1	7.0	6.0	4.0	1.0	1.0	1.0
Total	3,295	428						

3.2. Setting thresholds

The item responses can be edited in order of priority, given by the score, if it is feasible to wait until all data have come in. Otherwise there must be some mechanism that classifies the score into, say, two groups: ‘‘needs editing’’ and ‘‘does not need editing.’’ A suitable

Table 3. The order in which all 13 changes in domain 312 were found with the ideal and the estimate-related method. Weighted absolute changes

Sequence number	1	2	3	4	5	6	7	8	9	10	11	12	13
Ideal method	19,000	9,161	5,095	5,000	2,905	2,831	2,216	2,117	2,000	1,604	1,132	87	8
Estimate-related method	19,000	5,095	9,161	2,000	5,000	2,905	2,117	2,216	1,604	1,132	8	87	2,831
DFFITS	1.21	.57	−.90	.13	−.19	.00	.01	.00	.02	.01	.00	.00	−.16

dividing point between two groups, a *threshold*, is the smallest score that will produce domain estimates that are not ''different'' (by a criterion given below) from the estimates obtained with fully edited data. The threshold can be set adaptively or be based on a model of historical data. Lawrence and McKenzie (2000) discuss a model where the $\hat{\hat{\delta}}_{z_k}$ obtained by (2) are assumed uniformly distributed random variables for units that would not be edited under a selective editing approach. The most common procedure, though, is probably to select a ''test data set,'' which could be from some previous periods of the survey, or be an early batch of data in the current survey period, apply full micro editing and determine suitable, conservative thresholds that will be in use for some time.

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:

$$BR(\hat{\theta}_{proposed}) = \frac{|\hat{\theta}_{proposed} - \hat{\theta}_{current}|}{\sqrt{\hat{V}(\hat{\theta}_{current})}} < 0.10 \tag{7}$$

with $\hat{\theta}_{current}$ and $\hat{\theta}_{proposed}$ being the estimates of a parameter θ under the current and the proposed editing method, respectively, and $\hat{V}(\hat{\theta}_{current})$ the estimated variance under the current method. The reason for setting the limit of $BR(\hat{\theta}_{proposed})$ at 10% is that if the values produced by the current method are regarded as the target that any proposed method should come close to, then the 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 et al. 1992, pp. 163–165). I used the ratio estimator for $\hat{\theta}$ with the frame variable turnover as auxiliary. This estimator is reasonably accurate.

The 10% limit is conservative. In a simulation study of empirical coverage probabilities for ratio estimators, Wu and Deng (1983) found that standard estimators for the ratio estimator variance, such as those used at the ONS, gave empirical coverage probabilities in the range of 79% to 93%. A bias ratio less than 10% gives a loss of coverage probability less than 1%, which therefore is entirely negligible compared with other shortcomings of common variance estimation. Lawrence and McDavitt (1994) used the bias ratio criterion with a 20% limit.

To review and set thresholds adaptively, a *progress graph* can be produced regularly during the data collection period. Figure 7 gives progress graphs for domain 312 produced on April 15, 20, and 30. For example, on April 15 fourteen turnover values had been manually investigated. The corresponding points given by (5) with Δ_0 set to zero have been sorted by descending score and graphed. If the progress graph looks like any of the curves in Figure 7, the threshold may be increased for the rest of the data collection, as the editing of units with the smallest scores above the original threshold have not led to any significant change in the estimates. If, on the other hand, the curve does not level out, the threshold may be lowered. Progress graphs give staff a means of 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 7 does not prove that scores below the thresholds would not contribute significantly to the cumulative change if they were edited. The only way of

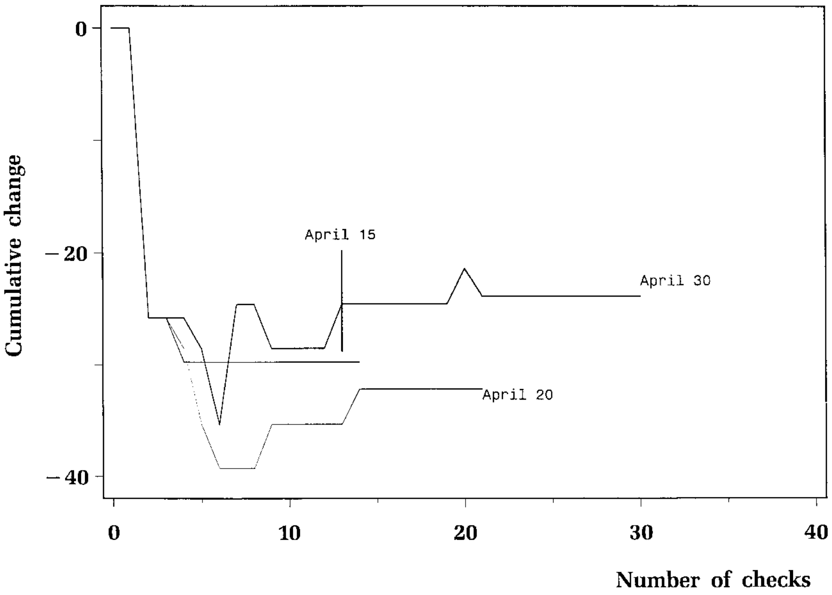

Fig. 7. Three progress graphs for domain 312. The cumulative change in £m against number of verified edit-failing values

estimating the contribution of small scores is to edit a random subsample below the threshold in order to estimate the net total of the errors that could have been corrected with more extensive micro editing, or an upper bound for this total. This type of sampling and estimation problem has been studied in the context of auditing, that is, verification of financial accounts (see Panel on Nonstandard Mixtures of Distributions 1989; Thompson 1997; and references therein).

To find the initial set of domain thresholds the December to March periods were used as test data sets. For implementation reasons it was felt in the early stages of the project that the set of domain thresholds needed to be simple, rather than having a different threshold for each variable and domain. The strategy was to set the threshold for both turnover and employment at 0.12% of the corresponding domain total from the previous period with a view to meeting (7) for both variables on both domain and overall level for all periods from December to March. Once in use, the domain thresholds will be monitored and altered with the progress graph as one of the tools.

With the 0.12% threshold Method C selected only 38 of the 125 edit-failing values for domain 312, which includes the first eight of the changes in Table 3. With the remaining five changes left unattended the bias was about £5m (0.14% of the domain total for March). The inflation of the variance due to the five outstanding changes was entirely negligible. Table 4 shows that 60% of the turnover values in all MIDSS domains that are currently investigated manually can be left unattended. There were almost the same number of domains with positive and negative sums of the outstanding changes. For higher thresholds, though, there were more domains with positive sums due to the asymmetric distribution of changes. The difference between the columns in Table 4 indicates the loss in predictive power using proxy values for prediction in (2). If a questionnaire were manually investigated whenever either turnover or employment exceeds the

Table 4. Per cent of all MIDSS turnover values that failed at least one edit in the current system and that would have been manually investigated with selective editing and ideal scoring, respectively

	Estimate-related scoring	Ideal scoring
Dec	37	8
Jan	42	8
Feb	39	8
March	41	9

threshold, then only about 50% of the questionnaires that were actually checked in December and March would have been selected for editing. This rate is in line with the applications of selective editing reviewed by Granquist and Kovar (1997), all of which have reported savings of this size or more. With differential thresholds the reduction of manual work would have been larger.

3.3. Other parameters

While errors that pass the bias ratio criterion (7) have no real effect on estimated totals, they may have an effect on estimates of change. To probe into this, I computed domain estimates of month-on-month change based on data obtained with Methods A and C, respectively. The target parameter was the ratio of the difference between the current and previous level to the previous level. The levels were estimated with a ratio estimator. The two sets of estimates of change were very similar for both turnover and employment for each domain in each period from January to March. Since (2) targets errors that have an effect on sample sums and since estimates of both change and total are based on sample sums, this may not be surprising. To see if Method C captures large errors in a more general sense, estimates of domain medians were computed for fully edited and selectively edited data. Again, these estimates were not different under Methods A and C.

4. Extending to Other Periods of the MIDSS and Other ONS Business Surveys

Having seen that the estimate-related score function (2) gave excellent results for several periods of the MIDSS, the broader issue under what general circumstances selective editing can be expected to be successful is now addressed. The issue was crucial in the debate concerning whether this technique could be relied on in future periods of the MIDSS and if it could be considered for other business surveys at the ONS. In particular, there were four concerns: *i*) How sensitive is the estimate-related method regarding the accuracy of the predictions of the y_k? *ii*) Are the results shown in previous sections mainly due to scanning errors, which do not require respondent follow-up? *iii*) Looking at (2), it is obvious that some errors will slip through. For example, a business that always reports a zero value would never be selected for editing. Other systematic errors, such as a business that reports cumulative rather than monthly turnover, may also slip through. *iv*) Several ONS business surveys have many items. Would this lead to a larger proportion of questionnaires being selected for editing and not generate the large savings that the MIDSS tests indicated?

To start with the first concern, note that if the scores are only used for dividing observations into two groups, then the order and levels of scores are not important, except for their relation to the threshold. In an additional test, the frame variable turnover, instead of the previous edited value, was used for $\hat{y}_k$ in (2). Although the frame variable is often about 12 months old there was not much difference in outcome. This is in line with the empirical findings of Lawrence and McKenzie (2000). In fact, as long as the prediction error $y_k - \hat{y}_k$ is in general smaller than the real error $z_k - y_k$, the score $\hat{\delta}_{z_k} = w_k|(z_k - y_k) + (y_k - \hat{y}_k)|$ given by (2) will be a good proxy for the weighted real error. This may not happen if it is feasible to find and correct errors $z_k - y_k$ smaller than the ''natural'' movement from period $t-1$ to t: $y_{k,t} - y_{k,t-1}$ (which equals $y_k - \hat{y}_k$ for $\hat{y}_k = y_{k,t-1}$). However, if the absolute month-on-month change for businesses is, say, around 10–40%, then it will be difficult for editing clerks to detect spurious changes of about the same or smaller order. On the other hand, if there is information that can be used to detect these relatively small errors, then it could also be used for enhanced prediction of y_k. While more accurate prediction of y_k through a regression model with frame variables and past values of the survey variable as predictors did not improve the efficiency of the estimate-related method for the MIDSS, similar models may well give benefits for other surveys with particularly seasonal or volatile data.

Contrary to the second concern, for a survey with a given amount of manual micro editing, selective editing has the best potential to reduce the manual editing if all errors are small and symmetrically distributed. This claim is borne out by the following argument. Assume that the raw value for a responding unit k, z_k, is decomposable into a true value θ_k, an often neglected measurement error ε_k that micro editing cannot resolve, and a measurement error ξ_k that *can* be corrected with micro editing:

$$z_k = \theta_k + \varepsilon_k + \xi_k = y_k + \xi_k \tag{8}$$

The ξ_k will have a mixed distribution with a spike at zero. Let $\tau_k = (\xi_k | k \text{ is not selected for editing})$. The mean of z_k after selective editing is

$$E_M(z_k) = E_M(y_k) + E_M(\tau_k) \tag{9}$$

with the expectations taken over the error model distribution. Assume that $E_M(\tau_k) = \mu_\tau$ for all k, for some constant μ_τ. Let $\bar{\tau}_r$ be the average of the τ_i for a set of r unedited units, $i = 1, 2, \ldots, r$. Since the statistics most often used in business surveys are functions of averages, $\bar{\tau}_r$ must in this context be small in relation to important estimates for selective editing to work. By the law of large numbers, $\bar{\tau}_r$ approaches μ_τ under mild conditions for increasing r with relatively good pace if $V_M(\tau_k) = \sigma_\tau^2$ is small; more slowly if σ_τ^2 is large. This suggests that situations where selective editing will not reduce manual editing are those when one or more of the following points are true.

1. The expected value of τ_i for nonselected units, μ_τ, is far away from zero;
2. The variance of τ_i for nonselected units, σ_τ^2, is large;
3. Statistics are based on few units, some of which may have errors that could have been corrected with micro editing.

To rephrase the point made above, note that the ideal situation is when none of 1–3 holds, in which case few units need to be selected for editing. The negation of point 1 will in practice often mean that the distribution of τ_i is fairly symmetric. Systematic and scanning errors are likely to satisfy points 1 and 2 (cf. Figure 1), so they actually make the selective editing perform worse in the sense that the reduction of manual work seen in Table 4 would have been even larger without these errors.

If the unconditional variance $V_M(\xi_k)$ is negligible compared to $V_M(\varepsilon_k)$ so that the measurement errors of the edited and the raw values are approximately the same, then any micro editing is ineffective and could be abolished altogether. The measurement errors not amenable to micro editing may indeed be fairly large. Figure 2 indicated the presence of measurement errors. Also, there are some very large changes that are not included in the published estimates (see Figure 3). In absence of an analysis of measurement errors through for example a split plot experiment where the most experienced staff would edit one random part of the sample, and the regular process is used for the remainder, I conducted a limited simulation study based on an untested model. Normally distributed errors with zero mean and standard deviation equal to 2.5% of the edited value were added to each edited value in February and March. For the vast majority of the domains the sum of the changes that Method C left outstanding was less than 20% of the sum of the error terms (as an average over 1,000 simulations). This confirms that the effect that Method C has on the estimates is entirely negligible compared to other errors. Alternatively, multiple imputation could be used to create several possible ‘‘true values’’ for each recorded value (Heitjan and Rubin 1990). The variation among the estimates based on different imputations would allow a comparison of the effect of possible measurement errors and the difference between editing methods.

When it comes to the third concern about error types that for logical reasons cannot be captured with the estimate-related method the selective editing solution is to include specially designed scores for these – if judged necessary. No clear example of these error types could be observed in the data. Rather than furnishing a score for every conceivable error type, a limited number of scores that demonstrably add prediction power would be desirable. For example, a score that signals large differences between the reported value and the corresponding frame value may add some robustness to the process. It was decided, however, to use graphical editing for these potential errors rather than selective editing through score functions (DesJardin and Winkler 2000). Also, some of them can be corrected automatically. For example, cumulated turnover reported several months in a row could be replaced with the month-on-month differences without manual intervention if employment stayed the same.

To address the fourth concern, if selective editing can be used for surveys with many items, recall that the aim of this work was to find savings in the editing process without adversely affecting the quality of the estimates. This was interpreted strictly. Consequently, all variables must be edited until some quality criterion is met for each of them, in our case (7). If a score function with a suitable threshold is applied to each variable and if a questionnaire is manually inspected if any item score exceeds its threshold, it is interesting to know the probability that at least one raw value on a form scores higher than its threshold. Suppose there are ν variables on the form. If ν is large, say 100, it may appear very likely that at least one of them will overstep the threshold. However, as we shall see, that is a function of the strength of dependency between the item scores. Denote by A_i the event that variable i on a questionnaire will score higher than the threshold. Let

B_ν be the event that the questionnaire will be selected for editing, i.e., $B_\nu = \bigcup_{i=1}^{\nu} A_i$. Assume $P(A_i) = \alpha$ for all i. We shall derive $P(B_\nu)$ under two different assumptions:

1. The events A_i are independent.
2. The events A_i are exchangeable in the sense that all combinations of the same number of variables are equally likely to score higher than the threshold, that is $P(A_i A_j) = P(A_k A_l)$ for $i \neq j$ and $k \neq l$, $P(A_i A_j A_k) = P(A_l A_m A_n)$ for $i \neq j \neq k$ and $l \neq m \neq n$, and so on. $P(A_i A_j)$ may be arbitrarily close to or even equal to $P(A_i)P(A_j)$ for $i \neq j$. Assumption 2 specialises to Assumption 1 if all events are mutually independent.

Under Assumption 1,

$$P(B_\nu) = 1 - (1-\alpha)^\nu = \sum_{k=1}^{\nu} (-1)^{k-1} \binom{\nu}{k} \alpha^k$$

This probability approaches 1 quickly with increasing values of ν.

Under Assumption 2 it can be shown by induction that,

$$P(B_\nu) = \sum_{k=1}^{\nu} (-1)^{k-1} S_k, \text{ where } S_k = \binom{\nu}{k} P(A_1 A_2 \ldots A_k)$$

Denote $P(A_i|A_j)$, $i \neq j$, by β_1, $P(A_i|A_j A_k)$, $i \neq j \neq k$, by β_2, and so on. Then

$$P(B_\nu) = \sum_{k=1}^{\nu} (-1)^{k-1} \binom{\nu}{k} \alpha\beta_1\beta_2 \ldots \beta_{k-1}$$

For any survey it is reasonable to assume that $P(A_i) \leq P(A_i|A_j)\, i \neq j$. As an approximation it is also reasonable to assume that $P(A_i|A_j) \approx P(A_i|A_j A_k) \approx P(A_i|A_j A_k A_l) \ldots$, $i \neq j \neq k \neq l$. We shall assume that all β_i are exactly equal, and denote their common value by β. While the assumptions that $\beta = \beta_1 = \beta_2 = \ldots \beta_{\nu-1}$ and $\alpha \leqslant \beta$ are not a direct consequence of Assumption 2, it can be shown that the feasible region for these parameters will be small if these assumptions are violated. For example, if $\alpha > \beta$, then it is a consequence of Assumption 2 that α must be very small unless there are only a few variables on the form. This is also true if, for example, β_1 is not close to β_2. Under the given assumptions we obtain after some algebra

$$\begin{aligned} P(B_\nu) &= \sum_{k=1}^{\nu} (-1)^{k-1} \binom{\nu}{k} \alpha\beta^{k-1} \\ &= \frac{\alpha}{\beta}[1 - (1-\beta)^\nu] \end{aligned}$$

This probability is close to α/β if $\nu \geq 4$. In the MIDSS $\alpha \approx 0.4$ and $\beta_1 \approx 0.75$ for both employment and turnover, so $\alpha/\beta \approx 0.5$. Even if $\beta_1 < \beta_2 < \ldots$, and if the growth of the β's is not too large, the probability that a questionnaire is selected will be bounded well below 1 (the growth cannot be large under Assumption 2).

This reasoning shows that even if the score function is applied to each variable, it will not necessarily lead to all questionnaires having to be edited. Further, it is arguable that assigning scores to each variable is unnecessary and unrealistic. First, if the errors of some variables are highly correlated then a score for only one of them could be sufficient, although unless the correlation is 1, one or two large errors may well slip through. A stronger argument is that there are at least two reasons why scoring a large number of items will in practice not necessarily meet criterion (7) for all items: first, with an abundance of signals some may be overlooked; second, due to changes in distribution from period to period, the domain thresholds for individual items based on past data may not meet the criterion for the current period. Again, the progress graph will give advice on this. One referee believes that a unit score, if it is a sensible function of item scores, should be more consistent from period to period, and hence the domain thresholds should be determined on the basis of the unit score and not on a collection of separate item scores.

In sum, there is no reason why a score function technique should not greatly reduce the current micro editing for future periods of the MIDSS. For other surveys, though, the distribution of the outstanding errors for different levels of thresholds needs to be looked at, and surveys with many variables may need other editing or imputation methods as well.

5. Discussion

A selective editing approach has been examined in great detail and has been shown to be successful. 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.

Two score functions have been studied: one estimate-related method that prioritises by the predicted effect a suspect value will have on particular estimates, and one edit-related method that is explicitly based on specified edits and prioritises suspect values by the Mahalanobis distance of the magnitude of edit failures. It is interesting that both proved effective although they are based on very different rationales. Somewhat surprisingly, the former method was seen to work rather better for a wide range of different estimates, apart from the estimates it targeted. One reason for the slightly less successful prioritisation of the edit-related method was that the underlying micro editing system was inefficient in its new role. The estimate-related method does not depend on the edits. 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. Past experience at many other national statistical institutes has shown that it may not be enough to demonstrate potential efficiencies to gain acceptance; the new methods must also be presentable to a wide audience. The estimate-related method combined with the graphics shown here turned out to be a successful approach.

For selective editing to be effective, some conditions must be met. First, the errors that are not attended to under a selective editing approach need to have an expected value not too distant from zero and have small variance, or they need to be smaller than those errors that are not likely to be detected with traditional micro editing. Second, if there are many variables per unit, and if the selective editing approach is required not to miss any influential error that traditional micro editing is believed to have the capability of finding, then there must be strong dependencies between the occurrences of errors for different variables of the same unit. Otherwise the requirements need to be relaxed for certain variables or selective editing has to be combined with imputation. The accuracy of the prediction of errors, however, is not very sensitive for selective editing to be effective.

Systematic errors that can be corrected with micro editing will also limit the scope of selective editing, since they tend to have a strongly asymmetric distribution. Such an error structure should however be detected in the editing process. For example, it may be known to data editing clerks that many businesses underreport certain costs because they tend to use too narrow a definition of the activities whose costs are to be reported. This knowledge should be passed on to the ‘‘owners’’ of the questionnaire and thus contribute to the improvement of the survey process. Several authors have pointed out that one of the main purposes of editing in a repetitive survey should be to identify shortcomings of the methods used (Granquist and Kovar 1997). Thus, editing has a strong potential for quality improvement.

Since a single editing method cannot be expected to solve all editing problems, the methodology unit at the ONS has added graphical editing as a complement. Also, automated editing and imputation of the type of question that consists of one total and several components that sum up to the total have been implemented for some surveys. After scanning, the data pass through automated editing, and estimate-related scores are computed for both raw values and automatically edited values.

This article advocates that the threshold, above which units will be selected for editing, should be based initially on a test dataset, and then followed up by an adaptive approach. The ‘‘progress graph’’ allows the effect of units just above the threshold to be monitored. If the amount or structure of errors suddenly changes, for example due to some structural change that affects the respondents’ reporting capability, the predetermined thresholds may not be appropriate. A subsample of records below the threshold should also be selected on a regular basis to estimate the total remaining error.

6. References

DesJardins, D.L. and Winkler, W.E. (2000). Design of Inlier and Outlier Edits for Business Surveys. Proceedings of American Statistical Association, Second International Conference on Establishment Surveys, 547–556.

Dillman, D.A. (2000). Procedures for Conducting Government-Sponsored Establishment Surveys: Comparison of the Total Design Method (TDM), a Traditional Cost-Compensation Model, and Tailored Design. Proceedings of American Statistical Association, Second International Conference on Establishment Surveys, 343–352.

Dippo, C.S., Young, I.C., and Sander, J. (1995). Designing the Data Collection Process. In Business Survey Methods, B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds). New York: Wiley, 283–301.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Granquist, L. (1995). Improving the Traditional Editing Process. In Business Survey Methods, B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds). New York: Wiley, 385–401.

Granquist, L. (1998). Efficient Editing – Improving Data Quality by Modern Editing.

Paper presented at the Conference on New Techniques and Technologies for Statistics, Sorrento, Italy, 4–6 Nov.

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

Heitjan, D.F. and Rubin, D.B. (1990). Inference from Coarse Data via Multiple Imputation with Application to Age Heaping. Journal of the American Statistical Association, 85, 304–314.

Krzanowski, W.J. (1990). Principles of Multivariate Analysis. A User's Perspective. Oxford Science Publications.

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.

Linacre, S.J. and Trewin, D.J. (1993). Total Survey Design – Application to a Collection of the Construction Industry. Journal of Official Statistics, 9, 611–621.

Little, R.J.A. and Smith, P.J. (1987). Editing and Imputation for Quantitative Survey Data. Journal of the American Statistical Association, 82, 58–68.

Lyberg, L.E., Biemer, P.P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N, and Trewin, D.J. (eds) (1997). Survey Measurement and Process Quality. New York: Wiley.

Nijhowne, S. (1995). Defining and Classifying Statistical Units. In Business Survey Methods, B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds). New York: Wiley, 49–64.

Panel on Nonstandard Mixtures of Distributions (1989). Statistical Models and Analysis in Auditing, Statistical Science, 4, 2–33.

Paxson, M.C, Dillman, D.A., and Tarnai, J. (1995). Improving Response to Business Mail Surveys. In Business Survey Methods, B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds). New York: Wiley, 303–316.

Pietsch, L. (1995). Profiling Large Businesses to Define Frame Units. In Business Survey Methods, B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds). New York: Wiley, 101–114.

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.

Thompson, K.J. and Sigman, R.S. (1999). Statistical Methods for Developing Ratio Edit Tolerances for Economic Data. Journal of Official Statistics, 15, 517–535.

UN (1997). Statistical Data Editing. Volume No. 2. Methods and Techniques. Conference of European Statisticians. Statistical Standards and Studies No. 48. United Nations Statistical Commission and Economic Commission for Europe.

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

Underwood, C., Small, C., and Thomas, P. (2000). Improving the Efficiency of Data Validation and Editing Activities for Business Surveys (CD-ROM). Proceedings of American Statistical Association, Second International Conference on Establishment Surveys, Buffalo, 18–21 June.

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

Wu, C.F. and Deng, L.Y. (1983). Estimation of Variance of the Ratio Estimator: An Empirical Study. In Scientific Inference, Data Analysis, and Robustness, G.E.P. Box, T. Leonard, C.F. Wu (eds). New York: Academic Press, 245–277.

Received May 2001
Revised September 2002