

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

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Ottawa, Canada, 16-18 May 2005)

Topic (ii): Implementing editing strategies and links to other parts of processing

**A MANAGEMENT INFORMATION SYSTEM FOR CONTROLLING EDITING QUALITY IN
A SURVEY WITH MULTIPLE REQUIREMENTS**

Supporting Paper

Submitted by the Office for National Statistics, United Kingdom¹

I. INTRODUCTION

1. In 1998, as part of a major efficiency programme, ONS undertook the development and application of new methodological approaches to substantially reduce the total costs of the survey operation. This work was restricted to the development of short-term improvements, or quick wins, to existing systems rather than total re-engineering (Tate et al., 2001). This comprised the implementation of automatic editing on some types of error and the methodology of selective editing (Lawrence and McKenzie, 2000). Automatic editing is now implemented on 8 surveys, and selective editing on 5 surveys; these editing procedures have created savings of about £250K from the total survey costs.

2. Selective editing has been applied to the New Earnings Survey (NES) since 2002. NES is a panel survey in which 240,000 questionnaires are sent out to 80,000 employers annually, with approximately 169,000 responses used in results. In practice, about 75% to 80% of employees remain in the sample from one year to the next. NES is a complex survey that produces statistics on average pay in many domains, and also an estimate of the number of employees earning below the National Minimum Wage (NMW). Selective editing is applied to five key pay variables, two are directly captured: Basic Weekly Pay and Gross Annual Pay; whilst three variables are derived: Gross Hourly Pay, Hourly Pay Excluding Bonuses and Gross Weekly Pay. The variables are scored and their scores compared to specified thresholds. The threshold values are set to ensure that the overall editing bias, which results from not editing all records potentially in error, of the estimator of a population parameter of interest is small in relation to the standard error of the estimate. The thresholds are determined using past pre-edited and edited data.

3. Analysis of past NES data has shown that selective editing identifies potential errors that can have a substantial impact on estimates of averages, but fails to identify many errors that have an impact on low pay estimates. For this reason, separate edit checks were set up for low pay data in NES. This causes a conflict over the allocation of the limited editing resource to cater satisfactorily for the multiple requirements of the survey. A fixed number of data analysts are employed solely to scrutinise and edit all types of error in NES, and work to a fixed deadline by which to clear the errors. For both the 2003 and 2004 survey cycles, the volume of validation failures was far higher than expected and eventually overwhelmed the workflow. During the later stages of the editing period, the selective editing thresholds were raised in order to target those records with the most significant influence on the final estimates. The timing of the change to the thresholds and the amount by which they were raised were not determined

¹ Prepared by Salah Merad and Heather Wagstaff (salah.merad@ons.gov.uk, heather.wagstaff@ons.gov.uk)

according to pre-specified criteria. It was done on a trial and error basis, and the impact of the change on data quality could not be assessed.

4. The objective of this paper is to develop a framework to support the management of the editing process in such a way that data quality is maintained. In Section II, we describe a simple decision rule that will indicate when available resources will be insufficient to cope with current and expected workload. In Section III, we propose a method to determine the amount by which thresholds can be raised in order to satisfy editing capacity constraints, whilst maintaining data quality. We will use NES to demonstrate the method. We conclude the paper with some remarks about the implementation issues that need addressing, and discuss briefly the strategy for the future.

II. DEFINING A DECISION RULE

5. We want to set up a simple decision rule that will inform data analysts at regular time intervals whether available resources are sufficient to cope with current and expected future workload. The rule involves comparing available man-hours with an estimate of the number of man-hours needed to complete the manual editing of all current and future records with errors. We first introduce some notation.

6. The editing process starts at time t_0 , and must be completed by time T . Time t , $t_0 \leq t < T$, can refer to a particular day or a week, depending on the survey cycle. In this paper we will refer to time in weeks. The editing process is comprised of several stages, and at each stage specific edit checks are applied. Some of these checks are hard (fatal), some are soft (statistical). As a simplification, we partition failed records into two types. Let records that fail only soft edit checks be of type E_{soft} , and let the remaining failed records be of type E_{other} . Records of the latter type are generally all manually edited. When selective editing is applied to a survey, only a subset of the records of type E_{soft} are manually edited. Records in this subset are said to be of type E_{SE} .

7. Let n be the total survey sample in the current survey cycle; $n_{t_0,t}$ be the number of responses that are returned by time t , $t_0 \leq t < T$; and $n_{t,T}$ be the number of responses that will arrive in the time interval $(t, T]$. Let \hat{r} be an estimator of the response rate of the survey; this estimator can be the response rate from the previous cycle. Then, an estimator of $n_{t,T}$ is given by $\hat{n}_{t,T} = \hat{r}n - n_{t_0,t}$. This estimator will be approximately unbiased if the subset of responses received by time t is a random sample of the set of all expected responses.

8. Let $p_{SE, t_0,t}$ and $p_{Other, t_0,t}$ be the proportions of failed records of types E_{SE} and E_{Other} , respectively, which are processed by time t . Let $n_{SE, t_0,t}^{(B)}$ and $n_{Other, t_0,t}^{(B)}$ be the number of records in the backlog with errors of type E_{SE} and E_{Other} , respectively, at time t . Records in the backlog are those that arrive before time t and are waiting manual editing.

9. We need to estimate the total man-hours needed to process the records in the backlog and future records with errors. To obtain this estimate we need to estimate a number of parameters. Let $p_{SE, t,T}$ and $p_{Other, t,T}$ be the proportions of records of types E_{Other} and E_{SE} , respectively, in the set of records that arrive in the time interval $(t, T]$. Their respective estimates, $\hat{p}_{Other, t,T}$ and $\hat{p}_{SE, t,T}$, can be computed using data from previous cycles of the survey and/or data up to time t from the current cycle. In surveys where

the questionnaire has been changed substantially, we recommend using data from the current survey cycle only, unless no processing has started for a particular type of error by time t . In the current NES editing system, some low pay edit checks are applied towards the end of the editing period. Therefore, we have to estimate the number of failures using the previous year's data. In the next section we evaluate the accuracy of the proportion $p_{SE, t_0, t}$ as an estimate of $p_{SE, t, T}$. The estimators of the number of records that require manual editing in relation to the different types of error are then given by $\hat{n}_{Other, t, T} = \hat{n}_{t, T} \hat{P}_{Other, t, T}$ and $\hat{n}_{SE, t, T} = \hat{n}_{t, T} \hat{P}_{SE, t, T}$. The estimator of the total man-hours needed to process the records in the backlog at time t and future records with errors is given by

$$H_{after\ t}^{(required)} = \left(\hat{n}_{SE, t_0, t}^{(B)} + \hat{n}_{SE, t, T} \right) \hat{m}_{SE, t, T} + \left(\hat{n}_{Other, t_0, t}^{(B)} + \hat{n}_{Other, t, T} \right) \hat{m}_{Other, t, T},$$

where $\hat{m}_{SE, t, T}$ and $\hat{m}_{Other, t, T}$ are estimators of the mean processing time per record of types E_{SE} and E_{Other} , respectively, for records in the backlog and future failures. Let $H_{t, T}^{(capacity)}$ be the available man-hours for editing in the time interval $(t, T]$. The decision rule is then:

If at time t $H_{after\ t}^{(required)} \leq H_{t, T}^{(capacity)}$, do nothing,

Else, reduce the number of records for manual editing by an amount $u_{t, T}$ so that

$$H_{after\ t}^{(required)} \leq H_{t, T}^{(capacity)}.$$

10. In order to reduce the number of records that are manually edited in the remaining editing period $(t, T]$, we propose to raise the selective editing thresholds by a safe amount. This can reduce the amount of manual editing substantially while maintaining the quality of the data, as we will show in the next section. To determine the amount $u_{t, T}$ by which to reduce the number of records that are manually edited, we first solve the equation

$$\left(\hat{n}_{SE, t_0, t}^{(B)} + \hat{n}_{SE, t, T}^{(capacity)} \right) \hat{m}_{SE, t, T} + \left(\hat{n}_{Other, t_0, t}^{(B)} + \hat{n}_{Other, t, T} \right) \hat{m}_{Other, t, T} = H_{t, T}^{(capacity)},$$

where the solution is denoted by $n_{SE, t, T}^{(capacity)}$. Then, $u_{t, T} = \left(\hat{n}_{SE, t_0, t}^{(B)} + \hat{n}_{SE, t, T} \right) - n_{SE, t, T}^{(capacity)}$.

Estimating the mean processing time per record with errors

11. In the above analysis we assumed that we can estimate separately the mean processing times for records of types E_{SE} and E_{Other} , which we denote by $m_{SE, t, T}$ and $m_{Other, t, T}$. The parameter $m_{SE, t, T}$ can be estimated using the ratio of the total processing time of records of type E_{SE} to the number of records of the same type and processed in the time interval $[t_0, t)$. However, because there are error sub-types within type E_{Other} that are only dealt with towards the end of the editing period, we cannot obtain a good estimate of $m_{Other, t, T}$ using only data from records processed in the interval $[t_0, t)$, especially for early times. We consider two ways to deal with this problem:

- (a) Estimate the mean processing time for the main distinct error sub-types within E_{Other} using data from the current survey cycle, when available in a large enough quantity, and data from the previous survey instance otherwise.
- (b) Estimate an overall mean processing time for type E_{Other} records using data from responses processed after time t in the previous cycle of the survey.

Solution (b) will yield a less accurate estimate of $\mathbf{m}_{Other,t,T}$ than solution (a), especially when a questionnaire has been amended and/or new edit checks have been introduced. However, solution (b) is easier to implement.

12. Note: the mean processing time could not be estimated separately for the different error types in previous cycles of NES, as the relevant information is not available. In the next cycle, $\mathbf{m}_{Other,t,T}$ will be estimated by computing the average processing time per record over all error types in the previous cycle. We plan to set up a system to make the appropriate measurements for future cycles.

III. REVISION OF SELECTIVE EDITING THRESHOLDS TO MEET RESOURCE CONSTRAINTS

13. Selective editing at ONS is implemented using the scoring method developed in Hedlin (2003). Let Y_1, Y_2, \dots, Y_K be the K key variables of the survey that are individually scored if a record fails a soft edit check related to any of the variables. Let $\mathbf{t}_{i,k}$ be the score of unit i for variable Y_k , and \mathbf{t}_k^* be the associated selective editing threshold. Records in which at least one of the scores exceeds the associated threshold are manually edited.

The score of unit i for variable Y_k is given by

$$\mathbf{t}_{i,k} = \frac{w_i (y_{i,k,l} - y_{i,k,l}^{(pred)})}{T_{D,k,l}^{(pred)}},$$

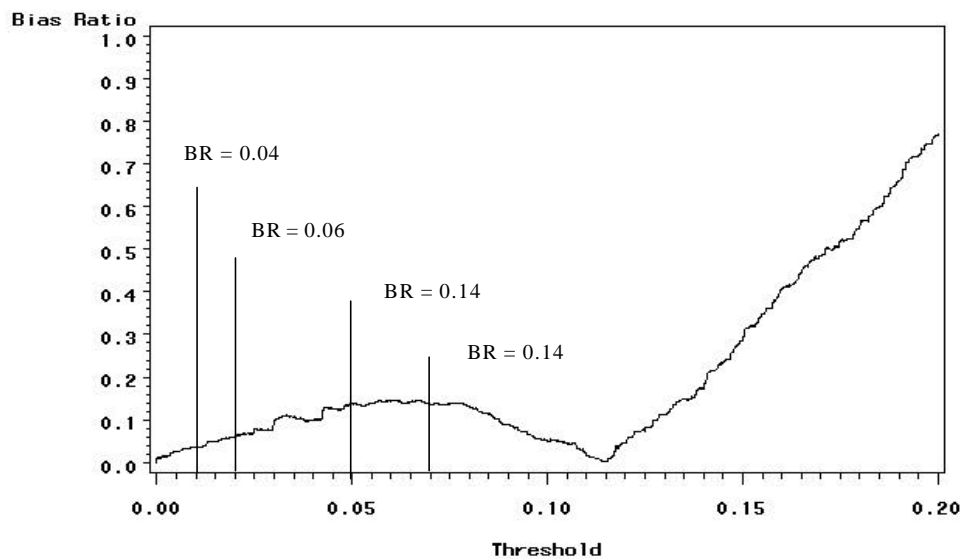
where w_i is the weight of unit i , $y_{i,k,l}^{(pred)}$ is the predicted value of variable Y_k for unit i at survey cycle l , and $T_{D,k,l}^{(pred)}$ is the predicted total of variable Y_k in domain D at survey cycle l , where D is a domain of interest. In NES, the predicted value, $y_{i,k,l}^{(pred)}$, is the value from the previous cycle, if available, otherwise, it is the domain median of the values from the previous cycle. The predicted total is the estimated domain total from the previous cycle.

14. The thresholds \mathbf{t}_k^* are determined on the basis of the (editing) bias ratio that results from not editing records with scores below the threshold. The threshold values for NES were obtained from the editing bias analysis carried out using 2001 data, which is prior to the introduction of selective editing to the survey. Graph 1 displays the (editing) bias ratio of the variable 'Weekly Basic Pay'.

15. When choosing a threshold value, we require:

- 1) The (editing) bias ratio to be small.
- 2) The graph of the (editing) bias ratio to be quite flat around the threshold; this ensures that the threshold value is robust (Jones 2002).
- 3) The percentage of domains in which the (editing) bias ratio exceeds 50% to be very low.

Note: the value of the (editing) bias ratio that is acceptable depends on the quality requirements of a survey. However, we require that the (absolute) bias ratio does not exceed 50%, as higher values would result in invalid confidence intervals (Särndal et al. 1992).



Graph 1: Absolute editing bias ratio as a function of threshold values for Weekly Basic Pay

16. We can see that, overall, the (editing) bias ratio increases for threshold values up to around 0.06, then decreases for values up to around 0.11, before increasing again steadily. The first increase occurs because most records with a very low score, below around 0.06, have a returned value lower than the edited value. The proportion of records in which the returned value is higher than the edited value increases with the score. When the threshold is above about 0.06, the differences between the returned and edited values start to cancel each other out, which causes the bias ratio to decrease. When the threshold is above about 0.11, the proportion of records in which the returned value is higher than the edited value becomes too high, and this causes the bias ratio to start increasing again.

17. Now, can we use Graph 1, which was obtained using 2001 data, to set thresholds for other survey cycles? We plotted the same graph using data from 2002 to 2004, and the shapes of the graphs were quite similar, even though in 2003 and 2004 selective editing was applied. The similarity between the graphs can be explained by the fact that most records that failed soft edits in 2001, but scored low, were unchanged after editing. This is also true for the other NES key variables. The threshold value that was recommended for 'Weekly Basic Pay' is 0.01. We can see from the graph that the resulting bias ratio is very low, and, from Table 1 that no domain has a bias ratio that exceeds 20%. The threshold is chosen on the basis of the overall bias ratio, but we can see that it also performs well at domain level. This is because the scoring function relates the discrepancies between returned and edited values to the domain estimates. The threshold value of 0.01 for 'Weekly Basic Pay' is very conservative; it can be increased safely to 0.07. However, it is risky to choose threshold values higher than 0.10, in case the graph of the actual bias ratio in future cycles of the survey shifts to the left of the 2001 graph by a non-negligible amount. In such cases, the bias ratio could exceed 50% for thresholds higher than about 0.10.

Bias ratio Range	Weekly Basic Pay		Gross Annual Pay	
	Threshold = 0.01	Threshold = 0.07	Threshold = 0.015	Threshold = 0.10
0%-10%	99.76	97.81	98.31	92.05
10%-30%	0.24	1.46	1.44	6.27
30%-50%	0.00	0.49	0.25	0.72
50%+	0.00	0.24	0.00	0.96

Table 1: Percentage of domains in different ranges of bias ratio; thresholds = (0.010, 0.015)

A method to determine new thresholds that satisfy capacity constraints in surveys where several variables are scored

18. In surveys where a large number of variables are scored, it is computationally more efficient to reduce the number of variables that are used to determine a new set of thresholds. This can be done by grouping together variables with highly correlated scores. In NES, it was noted that the scores of the four key variables Weekly Basic Pay, Gross Weekly Basic Pay, Gross Hourly Rate, and Gross Hourly Rate Excluding Overtime were highly correlated. In what follows, we will describe the method with two groups of variables, but it can be easily extended to more groups. We partition the scored variables in two groups, and a variable is chosen to represent each group. In NES, Group 1 is composed of the weekly and hourly variables, which we represent by the variable 'Weekly Basic Pay'. Group 2 is composed of 'Gross Annual Pay'. We represent the original thresholds of the two groups of variables by the pair (t_1^*, t_2^*) . The first component is associated with the first group, and the second component is associated with the second group. Because some survey variables can be more important than others, and assuming that the variables in a single group are of similar importance, we associate the coefficients a_1 and a_2 to group 1 and group 2, respectively, to reflect the importance of each group of variables. The coefficients are such that $a_1 + a_2 = 2$, that is they add up to the number of groups.

19. Because of the shape of the (editing) bias ratio graph and the uncertainty about the position of the actual shape of the graph in future cycles of the survey, we want a new pair of threshold values (d_1^*, d_2^*) such that:

- (a) it is as close as possible from the original pair (t_1^*, t_2^*) ;
- (b) resource constraints are satisfied;
- (c) each threshold value does not exceed the highest value that can be considered safe for the associated group of variables; we denote these bounds by b_1 and b_2 , respectively

In NES, using the graphs of the bias ratio against the threshold, which were obtained using past data, the safe upper-bound value for the weekly and hourly variables is set to 0.07, whereas for 'Gross Annual Pay' it is set to 0.10.

20. Hence, to find the new acceptable set of thresholds, we need to solve the following optimisation problem:

$$\text{Minimise}_{d_1^*, d_2^*} d\left((d_1^*, d_2^*), (t_1^*, t_2^*); a_1, a_2\right)$$

Subject to

$$\hat{n}_{SE,t,T}(t_1^*, t_2^*) - \hat{n}_{SE,t,T}(d_1^*, d_2^*) \geq u_{t,T},$$

$$d_1^* < b_1 \text{ and } d_2^* < b_2,$$

where $d\left((d_1^*, d_2^*), (t_1^*, t_2^*); a_1, a_2\right)$ is a distance function between the original pair and the new pair of thresholds, and where $\hat{n}_{SE,t,T}(t_1^*, t_2^*)$ and $\hat{n}_{SE,t,T}(d_1^*, d_2^*)$ are the estimates of the future number of records of type E_{SE} under the original and the new pair of thresholds, respectively. The distance function incorporates the importance coefficients for each group of variables. An example of such a distance

$$\text{function is } d\left((d_1^*, d_2^*), (t_1^*, t_2^*); a_1, a_2\right) = \frac{a_1 |d_1^* - t_1^*| + a_2 |d_2^* - t_2^*|}{2}.$$

21. Because we do not have explicit expressions for the estimators $\hat{n}_{SE,t,T}(\mathbf{t}_1^*, \mathbf{t}_2^*)$ and $\hat{n}_{SE,t,T}(\mathbf{d}_1^*, \mathbf{d}_2^*)$, we cannot solve the optimisation problem using the Lagrange method. Instead, we propose a simple heuristic method that finds a near-optimal solution. The heuristic is an iterative method based on the hill climbing technique. At each iteration we examine three candidate solutions. Let $(\mathbf{d}_1^{(k-1)}, \mathbf{d}_2^{(k-1)})$ be the solution at iteration $k-1$. Then, at iteration k we consider the neighbouring solutions $(\mathbf{d}_1^{(k-1)} + \mathbf{e}_1, \mathbf{d}_2^{(k-1)})$, $(\mathbf{d}_1^{(k-1)}, \mathbf{d}_2^{(k-1)} + \mathbf{e}_2)$, and $(\mathbf{d}_1^{(k-1)} + \mathbf{e}_1, \mathbf{d}_2^{(k-1)} + \mathbf{e}_2)$, where \mathbf{e}_1 and \mathbf{e}_2 are specified positive increments. The values of the increments will depend on the values of the scores. For NES, they can be set to 0.01 for 'Weekly Basic Pay' and to 0.02 for 'Gross Annual Pay'. Note that at iteration 0, $(\mathbf{d}_1^{(0)}, \mathbf{d}_2^{(0)}) = (\mathbf{t}_1^*, \mathbf{t}_2^*)$.
- (a) The algorithm is as follows: at iteration k :
 - (b) Construct the three candidate solutions from the solution at iteration $k-1$.
 - (c) For each candidate pair of thresholds do:
 1. Check whether both components are below the acceptable bounds \mathbf{b}_1 and \mathbf{b}_2 . If they are, then continue, otherwise reject the pair.
 2. Compute the proportion $p_{SE, \mathbf{t}, t}$ under the candidate pair of thresholds and the associated estimate $\hat{n}_{SE,t,T}$.
 3. Compute the difference between the original and the new estimates $\hat{n}_{SE,t,T}$. If this difference is more than the required amount $u_{t,T}$, then the candidate pair of thresholds is considered feasible.
 4. Compute the distance between the original solution and the candidate solution.
 5. Compute the value of the ratio of the reduction in the amount of manual editing to the distance between the original and the candidate solution. This ratio indicates the rate of improvement from the original solution.
 - (d) If at least one of the solutions is feasible (see step 3), then stop the algorithm. The solution with the highest ratio is chosen. Else, if not all three pairs are rejected (see step 1), then select the candidate solution with the highest ratio as the solution at iteration k . Set iteration to $k+1$, and repeat steps (a) to (c). Else, backtrack to iteration $k-1$, choose the next best solution and repeat steps (a) to (c).
22. If no solution that satisfies the editing capacity constraint can be found, then we need to consider alternative actions. We will not discuss these alternatives in this paper.

Simulations of the implementation of the method:

23. We carried out simulations using NES 2004 data to: (a) demonstrate the substantial reduction in manual editing that can result from raising thresholds; (b) evaluate the estimator of the proportion of future records of type E_{SE} . We consider each in turn.
- (a) Reduction of manual editing by raising thresholds:

24. Suppose that at the beginning of the editing period the pair of thresholds applied was (0.010, 0.015), and we decided to switch to the pair (0.02, 0.05) at week 5. This would result in reducing the number of records that are manually edited by about 12,000, which represents about 19% of the number of type E_{SE} records to be manually edited if the original thresholds were applied throughout.

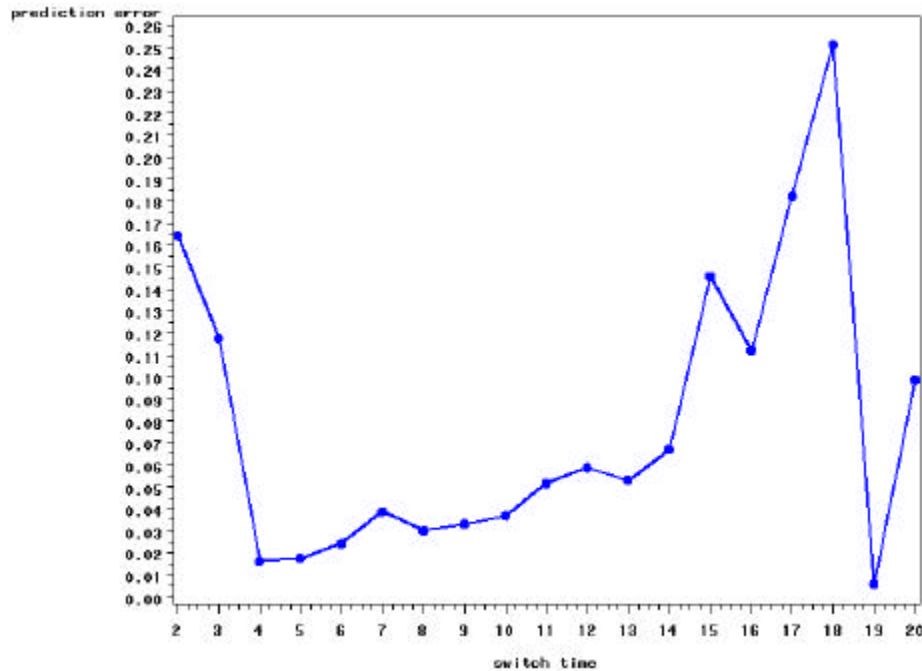
25. It can be shown that the (editing) bias ratio that results from switching from one pair of thresholds to another at a given time is lower than the maximum of the (editing) bias ratios under the individual pairs, provided that the switch is not made too early and not too late. Enough records are needed for terms to cancel each other in the expression of the editing bias. Therefore, by switching to the new pair of thresholds (0.02, 0.05) from week 5 onwards, the (editing) bias ratio for Group 1 variables is below 6% and for 'Gross Annual Pay' below 12%.

(b) Evaluation of the estimator of the proportion of type E_{SE} records:

26. To estimate the reduction in the number of records that need editing because of a change in threshold values at time t , we need to estimate $p_{SE,t,T}$. In Section II, we proposed to estimate this proportion by computing the equivalent proportion from data available by time t in the current cycle; that is, by $p_{SE,t,t}$. We evaluated the accuracy of the estimator $p_{SE,t,t}$ using data from NES 2004. The editing process lasted 20 weeks, so, for the weeks t , $t = 1, \dots, 19$:

- (a) we predicted the proportion in the remaining period $(t, T]$;
- (b) we computed the actual value of the proportion in the period $(t, T]$;
- (c) we computed the absolute relative difference between the predicted proportion and the actual proportion (we call the relative differences prediction errors).

27. As above, the starting pair of thresholds was (0.010, 0.015), and the second pair was (0.02, 0.05). We can see from Graph 2 that the prediction errors are small when the estimates are computed between week 5 and week 15, but increase substantially after that. Graph 2 shows that the prediction error for $\hat{p}_{SE,t,T}$ changes wildly towards the end; this is caused by the big changes in the proportion of type E_{SE} records. This proportion was 41% in week 18, 61% in week 19, and 45% in week 20. The 2004 data suggest that amending the thresholds between week 4 and week 15 would lead to a quite accurate estimate of the proportion $p_{SE,t,T}$. The pattern of arrival of questionnaires could be different in future cycles of the survey. However, we expect the quality of the estimates to be good, provided that an adequate volume of records is processed.



Graph 2: Prediction errors for the estimators of selective editing failure proportions as a function of the switch time between the pairs of thresholds (0.010, 0.015) and (0.02, 0.05).

IV. CONCLUDING REMARKS

28. Selective editing has brought substantial resource savings to ONS over recent years, but the methodology and its implementation must be constantly reviewed to respond to new challenges. In the next survey cycle of NES we face a lot of uncertainty, especially because of the redesign of the survey. We are planning to use the method for controlling the workload in a basic way. We are also planning to set up a system to make the measurements needed to obtain accurate parameter estimates, which will make the decision rule we defined in Section II more effective. The errors management system that we are putting forward should go some way towards controlling the quality of the data under resource constraints. However, we need to improve the whole processing operation; currently ONS is undergoing a statistical modernisation programme, this provides an opportunity for re-engineering the processing operation.

References

Hedlin, D. (2003) "Score Functions to Reduce Business Survey Editing at the U.K> Office for National Statistics". *Journal of Official Statistics*, Vol. 19, No.2, pp. 177-199.

Jones, D.L. (2002) "Selective Editing Thresholds: monitoring and measuring sensitivity". Seventh Government Statistical Service Methodology Conference. "Quality and Methodology in National Statistics", London.

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

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

Tate, P., Underwood, C., Thomas, P. and Small, C., (2001) "Challenges in Developing and Implementing New Data Editing Methods for Business Surveys". *Proceedings of Statistics Canada Symposium 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective"*.
