WP. 45
ENGLISH ONLY

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 (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

AN EDITING PROCEDURE FOR LOW PAY DATA IN THE ANNUAL SURVEY OF HOURS AND EARNINGS

Supporting Paper

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

I. INTRODUCTION

1. The development of the Annual Survey of Hours and Earnings (ASHE) to replace the New Earnings Survey (NES) was ONS's first major survey redesign as part of the modernisation programme. The new name, ASHE, will be adopted in 2005. The NES, which measures the earnings of employees in employment in the UK across the whole economy in April of each year, was designed to meet the policy needs of the 1970s and had changed little over the past thirty years. The NES had several deficiencies, in particular, it had no imputation, was unweighted and had coverage problems. ASHE provides an opportunity to improve the methodology of the survey, to meet users' requirements and to make use of new statistical tools.

2. ASHE is an employer survey based on a 1 per cent sample of employees from the Inland Revenue PAYE list, supplemented by a survey of businesses registered for VAT but not on the PAYE list. Around 240,000 questionnaires are sent out, and approximately 169,000 responses for employees are used in results. The sample design creates what is effectively a panel of employees. Requested data include their earnings, hours worked, and a description of their occupation. The hourly rate, which we refer to as the stated hourly rate, is also requested for employees who are paid on an hourly basis.

3. An important output of ASHE is to produce estimates of employees working below the National Minimum Wage (NMW). The NMW was introduced in the UK in April 1999, motivated, at least partly, by a desire to reverse a trend in wage inequality. Prior to 2004, the number of employees below the NMW was estimated using data from the New Earning Survey (NES) and the UK Labour Force Survey. Starting in 2004, only the ASHE data will be used. Classification to the low pay region (i.e.: below the NMW) is based on hourly pay derived from total pay, hours worked, and the reporting pay period. This derived variable is subject to several sources of error, and can result in incorrectly declaring employees earning less than the NMW.

4. As part of ONS's effort to introduce efficiency in its editing practices (Tate et al 2001), selective editing was implemented in NES 2003. Selective editing focuses on records potentially in error which have the most impact on the final estimates of totals and averages (Lawrence and McKenzie 2000). The

[1] Prepared by Salah Merad, Mike Hidiroglou, and Fiona Crawford (salah.merad@ons.gov.uk).

choice of those records was made using the selective editing procedure described in Hedlin (2003). However, selective editing cannot be directly used to identify records in error for specialized estimates, such as the estimate of the number of employees below the National Minimum Wage. It is important to track these errors, because any small change in the data will have an impact on the count of records declared to be below the NMW. It would be desirable to follow-up all records potentially in error in relation to low pay estimates. However, this is time consuming and cost prohibitive on account of the relatively large number of records that may be potentially in error. We propose a procedure that results in following-up only a subset of such records. As with the selective editing procedure, the method is constructed so as to avoid any significant bias.

5. The paper is structured as follows. In section II, we provide a brief summary of the analysis of the edit checks we constructed for records below the NMW for the 2003 NES data. In Section III, we present a new selective editing procedure that identifies records where two measures of hourly rate, the derived hourly rate and the stated hourly rate, differ significantly in relation to low pay estimates. Even though this procedure reduces the number of records to follow-up, this is not enough, as it exceeds available resources. In Section IV, we present a solution to this problem which is to adopt a hybrid editing procedure that follows up a smaller number of records potentially in error, and yet maintains data quality. In Section V, we describe our experience with the implementation of the procedure in NES 2004, and discuss the issues that we need to address for future instances of the survey.

II. ANALYSIS OF NES 2003 DATA

6. In our search for an effective editing method, we carried out analyses using pre-edited and post-edited NES data from the years 2003 and 2002. We assumed that the post-edited data were clean. The variables that are relevant to low pay statistics are:

(i) pay period, which can be weekly, two-weekly, four-weekly, monthly, or other;
(ii) basic average weekly hours worked in the pay period (overtime hours are excluded);
(iii) basic pay for the pay period;
(iv) incentive pay.

7. Other variables which we used in our analyses are: the Standard Occupation Classification (SOC), age, sex, the adult rate marker, which indicates if an employee is paid at a trainee or junior rate, and same job marker, which indicates if an employee has been in the same job for more than a year.

8. The NMW varies according to age and from year to year; before 2005 there was no NMW for employees below age 18. For 2005, a NMW for age group 16-17 was introduced. The table below shows the values of the NMW in different age groups for the years 2002 to 2005. In our analysis, we focus on the age group 22+.

	Age Group		
Period	**16-17**	**18-21**	**22+**
2002	N/A	£3.50	£4.10
2003	N/A	£3.60	£4.20
2004	N/A	£3.80	£4.50
2005	£3.00	£4.10	£4.85

Table 1. NMW by age group in 2002 to 2005.

9. Employees are declared as below the NMW according to their derived gross hourly rate, DGHR, which is defined as

$$DGHR = \frac{Derived\,Weekly\,Basic\,Pay + Derived\,Weekly\;\;Incentive\,Pay}{Basic\,Average\,Weekly\,Hours}.$$

We decided not to use incentive pay in the edits because it is not common and varies a lot from year to year. Instead, we use the basic derived hourly rate, DBHR, or just *Derived*, which is defined as

$$Derived = \frac{Derived\,Weekly\,Basic\,Pay}{Basic\,Average\,Weekly\,Hours}$$

The derived weekly basic pay is based on the basic pay and the pay period.

10. About 45% of employees reported a stated hourly rate (SHR), which we will refer to as *Stated*, in 2002. The question on hourly rate was removed in 2003, but was re-introduced in 2004. *Derived* is subject to more sources of error than *Stated*. They include scanning errors in all component variables, inaccurate average weekly hours for those on pay periods other than weekly, and basic pay reported for a week rather than for the pay period. Errors from the last source can be corrected automatically when the *Stated* is available.

11. We introduce the following notation. *Derived*(t), *Stated*(t), and NMW(t) are the derived basic hourly rate, stated hourly rate, and national minimum wage in year t, respectively. Records that are relevant to low pay statistics are those that satisfy either of two conditions.
The first is where the *Derived*(t)<NMW(t); the second is where *Derived*(t) $\geq$ NMW(t), but *Stated*(t)<NMW(t). Note that if *Stated* in not available, then the second condition is not applicable.

12. Through discussions with the low pay statistics team, we agreed that employees in the managerial and professional occupations were highly unlikely to be low earners. It was therefore decided that they should all be checked (edit 8 in Table 2). For the remaining occupations, where *Derived*(2003)<NMW(2003), *Derived*(2002) is available, and when employees have been in the same job for more than a year, we constructed four conditions which are based on *Derived*(2003) and *Derived*(2002) (edits 1 to 4 in Table 2). It was decided that only records satisfying the condition given by edit 2 would not be declared in error. It was thought highly plausible to have *Derived*(2003)<NMW(2003), when *Derived*(2002)<NMW(2002) and *Derived*(2003) $\geq$ *Derived*(2002).

13. Note that for records where *Derived*(2002)>NMW(2002) and *Derived*(2003)<NMW(2003) we constructed two separate conditions. One is for employees with a very low *Derived*(2003) (less than £3), given by edit 3, and one is for those with *Derived*(2003) that is not so low (£3 $\leq$ *Derived*(2003)<NMW(2003))**,** given by edit 4. The proportion of records failing edit 3 and found to be in error in 2003 after follow-up (hit rate) was considerably higher than that of records failing edit 4. The boundary that separates the two conditions was found empirically. Small perturbations in the value of the boundary led to very small changes to the hit rate values.

14. For employees with no previous *Derived* or who are not in the same job (edits 5 *to* 7), all records with *Derived*(2003)<NMW(2003) are considered to be potentially in error. No other reliable information could be used to refine the condition.

15. In 2003, only 141 out of 3264 records satisfied the condition that defines edit 2. This meant that about 95% of records in scope are potentially in error. Checking all these would be time consuming. We therefore needed an editing method that would reduce the amount of checking without introducing significant bias into the estimates.

Edit No.	One-digit SOC	Previous Year Available	Same Job	Condition	No. of records	Action
1.	4-9	Y	Y	*Derived*(2003)< NMW(2003), *Derived*(2002)<NMW(2002), *Derived*(2003)<*Derived*(2002)	45	Check everybody
2.	4-9	Y	Y	*Derived*(2003)< NMW(2003), *Derived*(2002)<NMW(2002), *Derived*(2003) ≥ *Derived*(2002)	141	Accept
3.	4-9	Y	Y	*Derived*(2003)<£3, *Derived*(2002) ≥ NMW(2002)	700	Check everybody
4.	4-9	Y	Y	£3 ≤ *Derived*(2003)<NMW(2003), *Derived*(2002) ≥ NMW(2002)	509	Preliminary edit/Hybrid editing
5.	4-9	N	N	*Derived*(2003)< NMW(2003)	825	
6.	4-9	Y	N	*Derived*(2003)< NMW(2003)	13	
7.	4-9	N	Y	*Derived*(2003)< NMW(2003)	684	
8.	1-3	-	-	*Derived*(2003)< NMW(2003)	347	Check everybody
				Total Number below NMW(2003)	3264	
				Total Number of failed records	3123	

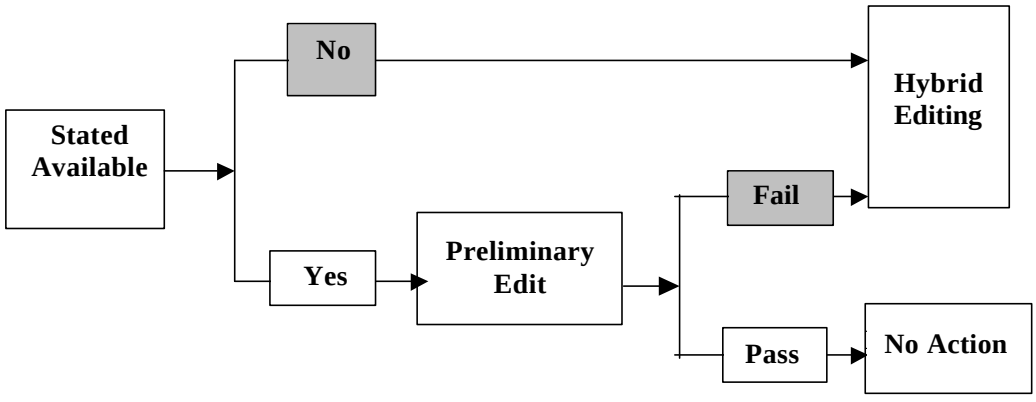
Table 2: Edit checks and counts for age group 22 and over in 2003.

16. We decided to follow-up all records that fail edits which indicate that the records are very likely to be in error (edits 3 and 8), and where the number of failed records is small (edit 1). The subset of failed records that contains such records is denoted by s_1. Records that satisfy the conditions associated with edits 4 to 7 make up subset s_2. We considered two methods for reducing the amount of checking in subset s_2; these methods can be used separately or can be combined.

(a) We apply a preliminary edit check that compares *Derived* to *Stated*, if the latter is available. Only records that fail this edit would go through the set of edits 4 to 7 in Table 2. The construction of the preliminary edit check is discussed in Section 3.

(b) We follow-up a random sample of the failed records and use imputation as a means to correct the remaining records; we call this hybrid editing. The sampling and imputation methods are discussed in Section 4.

17. The diagram below shows the editing process for records in subset s_2 when the two methods are combined. Note that records where *Derived*(t) ≥ NMW(t), but *Stated*(t)<NMW(t) can only go through the preliminary edit, and hence they are not covered by the following diagram.



III. PRELIMINARY EDIT BASED ON DIFFERENCE BETWEEN *STATED* AND *DERIVED*

18. The two variables *Derived* and *Stated* will differ by some amount. They represent the same value and by definition should be identical. Any difference between these variables implies that one or the other is in error. It is reasonable to assume that *Stated* is more accurate than *Derived* as the latter is subject to more sources of error.

19. How different should *Derived* and *Stated* be to follow-up the respondent? We obtain this difference by slightly modifying the selective editing scheme proposed by Hedlin (2003). Hedlin developed the scheme in the context of an estimation of a total parameter. Selective editing focuses on minimising the review of records potentially in error so as to obtain reasonable data aggregates such as totals. This objective is well supported by the fact that published studies of traditional editing processes indicate that many originally reported values are changed by insignificant amounts. In practice, a few errors are responsible for the majority of the total change. Records that are potentially in error and have the most impact on the estimate of a given parameter are followed-up if they are over a pre-set threshold.

20. Let s_{LP} be the set of records with *Stated* values and which are relevant to low pay, as defined in the previous section, and let n_{LP} be the size of this set. Let the *Stated* value of unit i be y_i (true) and its *Derived* be z_i (potentially in error). How do we decide whether z_i is significantly different from y_i?

21. Let $l_i = I_{\{Derived_i(t)<NMW(t)\}}$, $q_i = I_{\{Stated_i(t)<NMW(t)\}}$, and n_{Stated} be the number of records with *Stated* values in the whole NES sample. Let $\hat{p}_{Stated} = \frac{1}{n_{Stated}} \sum_{i=1}^{n_{LP}} q_i$ be an estimator based on *Stated* of the proportion of records below the NMW in the sub-population of employees paid on an hourly rate basis. The equivalent estimator based on *Derived* is given by $\hat{p}_{Derived} = \frac{1}{n_{Stated}} \sum_{i=1}^{n_{LP}} l_i$. If *Derived* and *Stated* values are quite different, $\hat{p}_{Derived}$ will be quite biased.

22. We follow a "selective-editing" like approach to isolate records that need following-up. First, we note that records where both *Stated* and *Derived* are below the NMW have no impact on the bias of $\hat{p}_{Derived}$. The absolute difference between q_i and l_i is equal to zero when both *Stated* and *Derived* are below the NMW, and is equal to 1 when one pay rate is below the NMW whereas the other is equal to or higher than the NMW. Let the difference between y_i and z_i be d_i. The score of record i, which indicates its potential impact on bias, is defined by

$$t_i = \left|(q_i - l_i)d_i\right|.$$

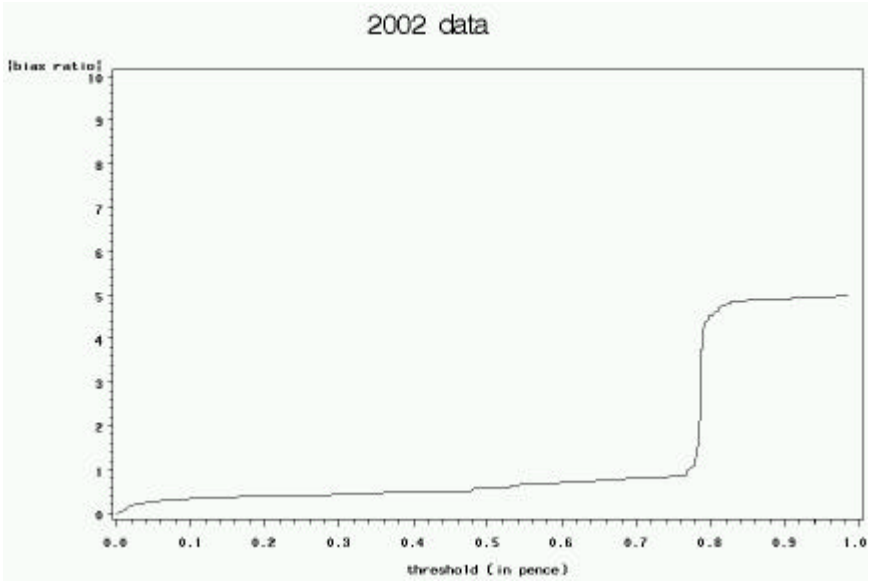
We rank the scores t_i from high to low. We denote the ranks generated in this fashion, and on the basis of decreasing magnitude of t_i, $r_1, r_2, \ldots, r_{n_{LP}}$: That is, $\left|t_{r_1}\right| \geq \left|t_{r_2}\right| \geq \ldots \geq \left|t_{r_{n_{LP}}}\right|$. We denote $\tilde{t}_i = t_{r_i}$ to simplify the notation; hence, $\tilde{t}_1 \geq \tilde{t}_2 \geq \ldots \geq \tilde{t}_{LP}$.

23. The bias of $\hat{p}_{Derived}$ can be reduced by following-up those units where the score t_i is above a given threshold. This threshold is determined as follows. The sample s_{LP} is divided into two parts, $s_{LP,1}$ and $s_{LP,2}$. Sub-sample $s_{LP,1}$ consists of the c units with the largest scores where we use $\tilde{y}_i$ (*Stated*). Sub-sample $s_{LP,2}$ consists of the remaining units where we use $\tilde{z}_i$ (*Derived*). All units in $s_{LP,1}$ will be followed-up, whereas none will be followed-up in $s_{LP,2}$.

24. Let $\hat{p}_{SEL}^{(c)}$ be an estimator of the proportion of employees below the NMW based on sub-samples $s_{LP,1}$ and $s_{LP,2}$. That is, $\hat{p}_{SEL}^{(c)} = \frac{1}{n_{Stated}}\left[\sum_{i=1}^{c}\tilde{q}_i + \sum_{i=c+1}^{n_{LP}}\tilde{l}_i\right]$. Note that $\hat{p}_{SEL}^{(0)} = \hat{p}_{Stated}$, while $\hat{p}_{SEL}^{(n_{LP})} = \hat{p}_{Derived}$. The use of $\hat{p}_{SEL}^{(c)}$ implies that we have rejected the "*c*" units where *Derived* values were $\tilde{z}_1,...,\tilde{z}_c$, and replaced them with *Stated* values $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_c$, respectively. We call $\hat{p}_{SEL}^{(c)}$ the contaminated estimator.

25. The (absolute) bias ratio of the contaminated estimator is given by $\frac{\left|\hat{p}_{Stated} - \hat{p}_{SEL}^{(c)}\right|}{s.e.\left(\hat{p}_{Stated}\right)}$, where the estimated standard error of $\hat{p}_{Stated}$, $s.e.\left(\hat{p}_{Stated}\right)$, is given by $\sqrt{\frac{\hat{p}_{Stated}\left(1-\hat{p}_{Stated}\right)}{n_{Stated}}}$. The (absolute) bias ratio is not smooth, but overall it decreases with increasing *c*. As a guide, it is required that the absolute bias ratio be less than 50% so that the 95% confidence intervals of the estimators are not distorted (Särndal et. al. 1997). Let $T_c = \tilde{t}_c$ be the threshold of the scores t_i, $i = 1,...,n_{LP}$. We want to determine the value of the threshold that yields a low (absolute) bias ratio.

26. Using data from NES 2002, we obtained a plot of the (absolute) bias ratio against threshold values, as shown in graph 1. We can see that for a threshold value of 0.1 pence, the (absolute) bias ratio is just over 30%, which is quite satisfactory. Moreover, the graph is quite flat for threshold values around 0.1 pence, and hence this threshold value can be considered as robust (Jones 2002). Even with a threshold value of 0.5 pence, the bias ratio is not too high. However, apart from 400 records where both *Stated* and *Derived* are below the NMW, only a handful of records would pass the preliminary edit with a threshold between 0.1 pence and 0.5 pence. We can see that there is a big jump in bias ratio when the threshold takes a value just over 0.8. On examining the scores, we found that there were 120 records, all with a monthly pay period, which have the same score and *Derived* value, and *Stated* at the NMW. We discussed this list of records with our clients, and agreed that they do not require follow-up. The relevant employers have computed monthly pay on the basis of 52 weeks in a year instead of 52.1, which we use to derive weekly pay from monthly pay. Excluding these records from the bias analysis produced a bias ratio with a smaller jump, but the threshold value of 0.1 pence, which corresponds to a 30% bias ratio, was unchanged.



.

Graph 1: (Absolute) bias ratio as a function of the threshold value.

IV. HYBRID EDITING

27. Another way to reduce the number of checks of potential errors is to edit some records and impute others; we call this method hybrid editing. This method is applied to records failing edits 4 to 7 where the hit rates are not very high (less than 50%). Because we do not follow-up every record potentially in error, an editing error will result. The expectation of this editing error, with respect to the selection of the random sample and the imputation method, gives the hybrid editing bias. We want this editing method to result in a non-significant bias. To be able to evaluate the method, we need to derive an expression of this bias.

A. Bias analysis

28. Let s_1 represent the subset of records that are fully followed-up, and s_2 the remaining set of records which have been through the preliminary edit, if applicable. Let $s_{2,1}$ be a random sample of records from s_2 that are followed up, and $s_{2,2} = s_2 - s_{2,1}$ be the set of records that are imputed. Also, let s_0 represent the set of records that do not fail low pay edits. We work out the bias analysis overall, but the results also hold at domain level. The un-weighted estimator of the proportion of employees below the NMW is given by

$$\hat{p} = \frac{\sum_{s_{,0}} d_i + \sum_{s_1} d_i^* + \sum_{s_{2,1}} d_i^* + \sum_{s_{2,2}} \tilde{d}_i}{n}$$

where *n* is the total number of respondents, $d_i = I_{\{Derived_i < NMW\}}$, $d_i^* = I_{\{Edited\ Derived_i < NMW\}}$, and $\tilde{d}_i = I_{\{\mathrm{Im}puted\ \ Derived_i < NMW\}}$. We require the hybrid editing to introduce little editing bias and variance.

29. The hybrid editing bias, which we denote by B_{edit}, is the bias that results from not fully editing the failed records. Its is given by

$$B_{edit} = \frac{1}{n}\left(E_{\ p,I}\left(\sum_{s_{2,1}} d_i^* + \sum_{s_{2,2}} \tilde{d}_i\ \right) - \sum_{s_2} d_i^*\right),$$

where the expectation is with respect to the random sampling scheme *p* and the imputation method *I*. Because $s_{2,1}$ and $s_{2,,2}$ are random samples from the subset s_2, it is easy to show that

$$B_{edit} = \frac{n_{2,2}}{n} E_{\ p}\left(E_I\left(\frac{\sum_{s_{2,2}} \tilde{d}_i}{n_{2,2}} - \frac{\sum_{s_{2,2}} d_i^*}{n_{2,2}} \middle| s_{2,2}\right)\right).$$

We require the imputation method to be such that this bias is small compared to the standard error of $\hat{p}$, say less than 30% of the standard error. The evaluation of the imputation method we used is discussed in Section 4d below. We first describe the imputation method.

B. Imputation method and definition of imputation classes

30. We consider the variables occupation class, sex, age, and adult rate marker to be good predictors of whether a record is truly below the NMW. Hence, using the edited records from subset s_2 as donors and matching on these variables should result in a small bias. An evaluation of the method is discussed below.

31. We used a donor method based on the Fellegi-Holt method, as implemented in the Statistics Canada software BANFF (Statistics Canada 2003). Basic pay and/or hours worked can be in error. We set both of these variables to be missing in the records to be imputed. We used the age group and the one-digit SOC to define the imputation classes. Within an imputation class we matched donors to recipients using the following variables : sex, age, and adult rate marker.

32. Note: We used sex and adult rate as matching variables rather than to define imputation classes so as to obtain imputation classes that are large enough. Although these variables are nominal categorical variables, they worked well as matching variables because they only take two possible values. The imputation produced a high percentage of perfect matches with both variables.

C. Sample size determination

33. We could use two criteria for determining how many records to follow-up. One criterion is based on available resources, whereas the other is based on the variance that results from sampling the subset of records to be followed up. We opted for the latter criterion.

34. As the donor method is a nearest neighbour method, most of the variance that results from not following-up all failed records is due to sampling. The hybrid method is applied to records that fail any edit from edits 4 to 7; the hit rates associated with these edits are different. To reduce this variance we stratify the sample s_2 by the type of edit that records fail. Note that, because only a small number of records failed edit 6 in NES 2003 data, records that fail either edit 5 or edit 6 form a single stratum.

35. In each stratum where a hybrid editing method is adopted, a random sample is selected for follow-up and validation. The sampling fraction is set such that the error size in the $100(1-a)\%$ confidence interval of the estimator of the proportion of actual errors in the stratum is bounded by a specified value e. The minimum sampling fraction is given by

$$f_h = \frac{z_{a/2}^2 S_h^2}{n_h e^2 + z_{a/2}^2 S_h^2},$$

where n_h is the number of records that fail Edit h, p_h is the proportion of actual errors (hit rate), and $S_h^2 = \frac{n_h}{n_h - 1} p_h (1 - p_h)$.

D. Evaluation of the imputation method

36. The hybrid editing bias ratio, BR_{edit}, is given by

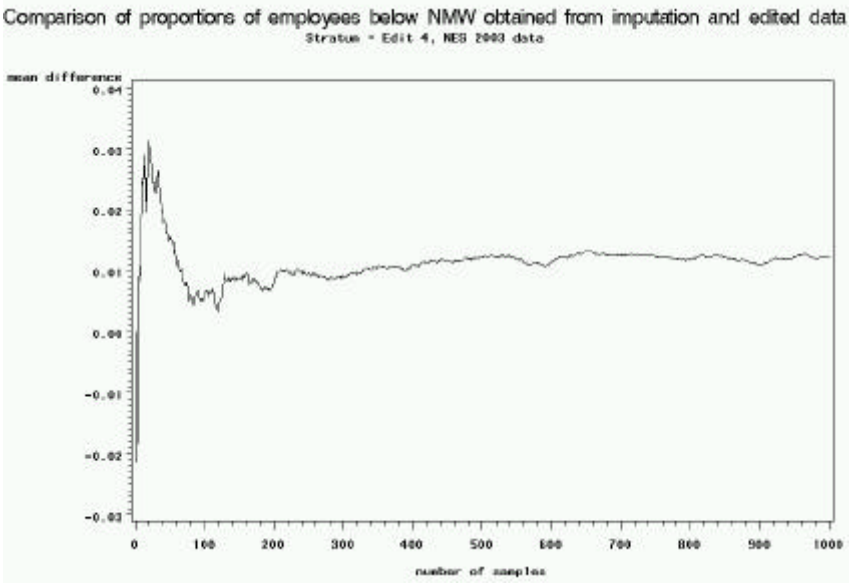
$$BR_{edit} = \frac{n_{22}}{n} \frac{E_p\left(E_I\left(\frac{\sum_{s_{2,2}} \tilde{d}_i}{n_{2,2}} - \frac{\sum_{s_{2,2}} d_i^*}{n_{2,2}} \Big| s_{2,2} \right) \right)}{s.e.(\hat{p})}$$

We used the minimum required sampling fractions given in Section 4c. These were obtained with the 2003 data, setting $e = 0.05$ and $a = 0.05$. This choice led to the ratios $\frac{n_{2,2}}{n}$ taking values around 0.01 overall and in most domains of interest (occupation groups, Males and Females). The overall estimate of the proportion of employees below the NMW is around 0.01, and the estimate of the standard error is around 0.00025.

37. To estimate the mean difference between the proportion $\frac{\sum_{s_{2,2}} \tilde{d}_i}{n_{2,2}}$, which is based on imputed values, and the proportion $\frac{\sum_{s_{2,2}} d_i^*}{n_{2,2}}$, which is based on edited values, with regard to sampling and imputation, we applied the method on 1000 random samples. In all strata where hybrid editing was applied, the (absolute) mean difference settled to values around 0.011 and 0.012 for means based on 500

samples or more. This is shown in Graph 2. This indicates that the (absolute) bias ratio is well under 50%, which is satisfactory.

38. At domain level (one-digit SOC and sex), the mean differences between the proportions settle around values very similar to the overall values. The ratios $\frac{n_{2,2}}{n}$ within each domain should also be similar to the overall value 0.01. The domain standard error estimates are likely to be slightly higher than the overall estimate, and hence the (absolute) bias ratios will be less than 50%, which is satisfactory.



Graph 2: Mean difference between 'imputed' and 'edited' proportions.

V. IMPLEMENTATION AND CONCLUDING REMARKS

39. The hourly rate question, which provides *Stated*, was amended for the 2004 questionnaire. It was therefore decided not to use the preliminary edit in NES 2004. The hybrid method was coded and integrated into the imputation system, which uses BANFF, and the results were used in the production of low pay statistics. Since the hybrid method is applied towards the end of the editing process, we were able to raise the sampling fractions to accommodate a lower number of failures than in our analysis. The reduction in the number of low pay failures was due, partly, to the fact that some of the would be failures were picked up in the early stages of the editing process using selective editing on the whole data set. Out of 2,212 low pay failures, 994, which represent only 0.6% of the total usable sample (169,000 records), were corrected via imputation. The bias ratio resulting from hybrid editing should then be comfortably less than 30%.

40. A repeat of the analysis in Section 3 using NES 2004 data led to a bias ratio similar to that found in the analysis based on NES 2002 data. Had the preliminary edit with a threshold of 0.1 pence been applied to NES 2004 data, it would have led to 490 records not being followed-up in the age group 22+. However, only 9 of these records have an impact on the bias. That is, in all 9 records, $Stated = NMW$ and $Derived < NMW$. In the bias analysis that we carried out in Section 3 we assumed that *Stated* is true, whereas *Derived* could be in error. This assumption is very likely to hold in general, but there are cases in which employers might think they are paying the NMW, but actually their employees receive less. This could be because of the way employers compute basic pay, as is the case in employees paid monthly. We managed to identify the problem with employees paid monthly, and agreed with our clients

not to follow-up such records. However, there may be others we do not know about. The true bias ratio from not following-up the records below a specified threshold, say 0.5 pence, could then be less than the estimated bias ratio.

41. In 2005, we plan to use the combination of the preliminary edit and the hybrid editing method. The bias ratio that will result from combining the two methods will be higher than that resulting from using each method separately. One way to keep the combined bias ratio under control is to fail the few records with a positive score in the preliminary edit; only these records can potentially increase the bias. However, following-up records that are not genuinely in error could lead to some respondents to revise their returned values artificially, so that they appear to be compliant with the NMW. This would then introduce negative bias. Balancing the positive bias, which could result from not editing records genuinely in error, and the negative bias, which could result from editing records that are actually clean, is a difficult problem. We will discuss with our clients the issues relating to the implementation of the combination of the two editing methods. We need to agree how to implement the preliminary edit in ASHE 2005.

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. (1997) "Model Assisted Survey Sampling". New York: Springer-Verlag.

Statistics Canada (2003) "Functional Description of the BANFF system for Edit and Imputation". Generalized System Methods Section, Business Survey Methods Division, December 2003.

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'.
