AUSTRIAN JOURNAL OF STATISTICS
Volume 38 (2009), Number 3, 171–182

Probability-Sampling Approach to Editing

Maiki Ilves[1] and Thomas Laitila[1,2]
[1]Department of Statistics, Örebro University, Sweden
[2]Research and Development Department, Statistics Sweden

Abstract: Editing for measurement errors is always part of data processing. In traditional editing, all data records are checked for errors and inconsistencies. In a new way of editing, only the subset with the most important erroneous responses is considered for editing. This approach is applied in selective editing procedures, which have been shown to save resources considerably. However, selective editing lacks a probabilistic basis and the properties of estimators cannot be established using standard methods. In particular, bias properties of the estimator are unknown except for level estimates based on historical data. This paper proposes combining selective editing with an editing procedure based on the traditional probability-sampling framework. The variance of a bias-corrected Horvitz-Thompson estimator is derived and a variance estimator is proposed. The results of a simulation study support the use of the combined editing procedure.

Zusammenfassung: Das Redigieren von Messfehlern ist immer Teil der Datenverarbeitung. Beim traditionellen Redigieren werden alle Datensätze auf Störungen und Inkonsistenzen überprüft. Bei einer neuen Art des Redigierens betrachtet man dafür nur die Teilmenge mit den wichtigsten fehlerhaften Antworten. Dieser Ansatz wird bei selektiven redigierenden Verfahren angewendet, und es zeigte sich dass dadurch beträchtliche Einsparungen erzielt werden konnten. Jedoch fehlt dem selektiven Redigieren die probabilistische Basis und Eigenschaften der Schätzer können nicht unter Verwendung von Standardmethoden hergeleitet werden. Insbesondere sind Bias Eigenschaften des Schätzers unbekannt, außer für Niveau Schätzer, die auf historische Daten basieren. In dieser Arbeit schlagen wir vor, selektives Redigieren mit einem redigierenden Verfahren zu kombinieren, das auf dem traditionellen System der Stichprobenauswahl beruht. Die Varianz eines bias-korrigierten Horvitz-Thompson Schätzers wird hergeleitet und ein Varianzschätzer wird vorgeschlagen. Die Resultate einer Simulationsstudie sprechen für die Verwendung des kombinierten redigierenden Verfahrens.

Keywords: Measurement Bias, Selective Editing, Two-phase Design.

1 Introduction

Accuracy is an important aspect on the quality of statistics and is included as a dimension of quality definitions used by statistical agencies.[1] The accuracy dimension is in turn divided into sampling and nonsampling errors, where measurement error is an example

[1]E.g. the quality definition used by EUROSTAT involves the dimensions Relevance, Accuracy, Timeliness and Punctuality, Accessibility and Clarity, Comparability and Coherence.

of the latter category. Measurement errors appear in surveys, censuses, and administrative data. Regardless of questionnaire testing, interviewer training, and built-in checks, measurement errors still occur. Data editing is a process by which possibly erroneous measurements obtained during data collection or processing are checked and corrected. The editing process serves three purposes: 1) to assess the data quality, 2) to improve the survey by identifying error sources, and 3) to find errors in the data and correct them.

Traditionally, editing aims to discover all errors and inconsistencies in the data and to correct them if necessary. This approach to editing is very laborious and costly, though it gives the impression of achieving high data quality. However, traditional microediting techniques are usually not justified as they can lead to overediting and biased results (Granquist and Kovar, 1997).

Selective editing is an alternate approach the goal of which is to focus on a subset of the erroneous responses without lowering the quality of survey estimates. Selective editing aims to identify suspicious responses, determine their impact on the final outcome through using a score function, and only edit responses with score values above a predetermined threshold.[2] The potential resource savings over traditional editing depend on the particular survey, but published results indicate that the gain can be large. For example, Latouche and Berthelot (1992) reported that, using Canadian Annual Retail Trade Survey data, almost no loss in estimate quality was noted by recontacting only one third of the suspicious units. Lawrence and McDavitt (1994) came to a similar conclusion when editing half of the suspicious responses.

The most emphasized advantage of selective editing over traditional editing is decreased workload for data editors and respondents. The cost savings can be substantial. Another possible advantage is improved timeliness. Due to these favorable properties, selective editing procedures have been implemented or are being considered for implementation by several national statistical agencies, for example, the UK Office of National Statistics (Hedlin, 2003), the Australian Bureau of Statistics (Lawrence and McDavitt, 1994), Statistics Netherlands (Hoogland, 2002), Statistics Sweden (Statistics Sweden, 2005), and Statistics Canada (Statistics Canada, 2003).

However, there are some problems and limitations in applying selective editing. The proposed editing approach lacks a basis in probabilistic theory and there has been no suggestion as to how to adjust the inference with regard to the editing procedure. Using a probabilistic sampling procedure, model-based approaches can likely be used for inference, for example, of the bias of estimators due to measurement errors in the unedited part of the data. Correlation between scores and errors could be used in estimating the error distribution in the unedited part. Such a model-based method has the disadvantage of requiring estimation based on extrapolations into the region of unedited score values.

This paper suggests that potential bias in estimates can be corrected using bias estimates obtained from a probability sample of unedited observations. Thus, observations are edited in two steps, selective editing being used in the first step, while a probability sample of observations is edited in the second. The measurement errors observed in the observations edited in the second step can be used in deriving a bias-corrected estima-

[2]Lawrence and McDavitt (1994) and Lawrence and McKenzie (2000) use the term "significance editing" instead of selective editing to indicate that the score function directly estimates the response effect on the survey estimates.

tor. This two-step procedure retains all the advantages of selective editing and in addition produces unbiased estimates.

This paper considers the Horvitz-Thompson estimator when selective editing is carried out and contributes a bias-corrected estimator. The variance of the corrected estimator is derived and a variance estimator is proposed. In an example, the bias-corrected Horvitz-Thompson estimator, its variance, and a variance estimator are derived for the case of a simple random sampling design with Poisson sampling of unedited units.

Selective editing, the suggested bias-corrected estimator, and the estimator properties are presented in Section 2. The third section considers the results of a simulation study of estimator properties. The paper ends with a discussion of the results and of proposals for future research.

2 Statistical Inference for Edited Data

Let us consider a population, $U = \{1, \ldots, N\}$, from which sample s_a of size n_a is drawn according to sampling design $p_a(\cdot)$. Denote true values by z_k and observed values by x_k. This section aims to derive an unbiased estimator of the population total of variable z, i.e., $t_z = \sum_{k=1}^{N} z_k$, in the case of measurement error in the observed sample units. For this purpose, the concept of selective editing and the theory of estimating measurement bias are combined into a two-step editing procedure. The first step is to carry out selective editing and the second is to edit observations randomly selected from the unedited part of the sample.

2.1 Selective editing

Selective editing procedures are mainly applied to quantitative data, and are considered to be most effective for highly skewed distributed variables where a small number of units account for much of the total estimate, as is frequently the case in business surveys, for example. Selective editing serves to correct the data and reduce the bias in the final estimates. In selective editing, suspicious responses are prioritized according to their influence on the survey estimate and only the most influential responses are edited. However, applying selective editing to a survey assumes considerable preliminary work.

First, to find out suspicious responses, comprehensive editing rules are needed. Editing rules can contain logical, consistency, and historical checks, and only responses failing these checks are further investigated. Errors not discovered by these editing rules will not be considered and their influence on the final estimates remains unknown. Therefore, selective editing requires comprehensive editing rules to work effectively.

Second, to prioritize suspicious responses, a score function needs to be defined and computed for all responses that failed the editing rules. The score function is a function of measured value and expected amended value. The score function should enable the identification of possibly erroneous responses having great influence on the survey estimates. The general form of the score function for unit k is

$$s_k = w_k |x_k - \hat{x}_k| \,,$$

where x_k is the observed value, $\hat{x}_k$ is the estimate of true value z_k, and w_k is the survey weight. It is possible to skip setting up editing rules by computing score values for all responses, but this places greater demands on the ability of the score function to discriminate between correct and erroneous responses.

In practice, several variables are measured in a survey, and score values, referred to as local scores, are calculated for each variable measurement obtained from a sampled unit. Instead of editing single variables separately, global score functions are constructed from the local scores. Editing decisions are based on the global scores and, if editing is decided on for a unit, all or a subset of variables are edited simultaneously. Thus, using global scores, the selective editing decisions are not based on the influence of a specific variable on a single total estimate.

Finally, the cutoff point between the most and least influential suspicious responses must be determined. Responses with scores above the cutoff point are selected for editing. It is assumed that correct responses are obtained during the editing. The cutoff point is usually selected through simulation using historical data, where survey estimates based on partly edited datasets are compared with estimates based on a fully edited dataset. The choice of cutoff point also determines the extent of bias remaining in the survey estimates. Note that bias estimates are obtained from simulations using prior data and are assumed to hold also in studies in which the editing rules are applied to new data.

The specific score function form and cutoff point are chosen on a survey-by-survey basis depending on available information.

In the literature, the effect of selective editing on estimates is assessed using simulation results. Lawrence and McKenzie (2000) mention two measures for evaluating the influence of selective editing on the survey estimates, absolute and relative pseudo-bias, defined as

$$\left|\frac{\hat{y}_q - \hat{y}_{100}}{\hat{y}_{100}}\right| \qquad \text{and} \qquad \left|\frac{\hat{y}_q - \hat{y}_{100}}{se(\hat{y}_{100})}\right| ,$$

respectively. Here $\hat{y}_q$ is the survey estimate based on the partly edited dataset, q denoting the proportion of responses considered and $\hat{y}_{100}$ the survey estimate based on the fully edited dataset. The term "pseudo-bias" is used because instead of the true quantity, t_z, the estimate $\hat{y}_{100}$ is used.

Theoretical expression of the estimator of the population total, and of its bias and precision under selective editing, is of interest. Because of measurement errors, true values are not obtained for all units, for some units $z_k \neq x_k$. After selective editing is carried out, the study variable, y_k, can be viewed as a composition of true and observed values

$$y_k = I_k^{edit} z_k + (1 - I_k^{edit}) x_k \,, \qquad k \in s_a \,,$$

where z_k is the accurate value obtained after editing unit k, x_k is the observed unedited value, and I_k^{edit} is an indicator function with a value of one for edited observations and zero otherwise. The variables z_k and x_k are here treated as nonrandom variables. Population U can be divided into two parts, U_1 and U_2, where U_1 includes units subjected to selective editing if selected for the sample and $U_2 = U - U_1$. However, this does not mean that the same units always respond with error or with the same amount of error; conditioning is done on the specific survey under consideration.

Now, the Horvitz-Thompson (HT) estimator of the total is

$$\hat{t}_y = \sum_{k=1}^{n_a} \frac{y_k}{\pi_{ak}} = \sum_{k \in s_a} \frac{y_k}{\pi_{ak}} , \tag{1}$$

where π_{ak} is the first-order inclusion probability for unit k. Hereafter the summation index is denoted in the form $k \in s_a$ to indicate that sum is taken over all units $k = \{1, \ldots, n_a\}$ belonging to the set s_a. The variance of the HT estimator is derived by using a general formula applicable to any without-replacement sampling design

$$\mathrm{var}(\hat{t}_y) = \sum_{k,l \in U} \sum \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} ,$$

where $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$ and π_{akl} being the second-order inclusion probability. The estimator (1) is not an unbiased estimator of the quantity of interest, t_z. The expected value of the estimator is

$$\mathrm{E}(\hat{t}_y) = \mathrm{E}\left[\sum_{k \in s_a} \frac{y_k}{\pi_{ak}}\right] = \sum_{k \in U_1} z_k + \sum_{k \in U_2} x_k ,$$

whereby the bias is obtained as

$$\mathrm{B}(\hat{t}_y) = \sum_{k \in U_2} e_k , \tag{2}$$

where $e_k = x_k - z_k$.

2.2 Bias-corrected HT estimator

For bias correction a two-step procedure is suggested here. Selective editing as described in 2.1 is carried out in the first step. To obtain unbiased estimates, the bias is estimated by subsampling units from the unedited part of the sample and determining the extent of the measurement error for all selected units. This constitutes the second step in the two-step procedure. The idea of estimating the remaining measurement bias through double sampling or two-phase sampling is described in the traditional editing context by Madow (1965), Lessler and Kalsbeek (1992), and Rao and Sitter (1997). However, in the present paper, bias correction is applied after selective editing and only to the set containing unedited units.

The editing process is interpreted as a two-phase sampling procedure in which the original sample is obtained in the first phase and the observations for editing are probability selected in the second. Using an estimator for two-phase sampling, the bias (2) can be estimated by

$$\hat{\mathrm{B}}(\hat{t}_y) = \sum_{k \in U_2} \frac{I_{ak} I_{k|s_{a2}} e_k}{\pi_{ak}\pi_{k|s_{a2}}} = \sum_{k \in s_2} \frac{e_k}{\pi_{ak}\pi_{k|s_{a2}}} , \tag{3}$$

where I_{ak} and $I_{k|s_{a2}}$ denote first- and second-phase sampling indicators, respectively. Similarly, π_{ak} and $\pi_{k|s_{a2}}$ denote the first- and second-phase first-order inclusion probabilities,

respectively. Estimator (3) is also called the $\pi^\star$-estimator (Särndal et al., 1992). Note that the second-phase sample contains units from U_2 only, i.e., units not selected in the first step of the procedure.

An unbiased estimator of t_z is now obtained by subtracting the estimated bias from the biased total estimate

$$\hat{t}_z = \sum_{k \in s_a} \frac{y_k}{\pi_{ak}} - \sum_{k \in s_2} \frac{e_k}{\pi_{ak}\pi_{k|s_a}}\,, \tag{4}$$

where

$$\sum_{k \in s_a} \frac{y_k}{\pi_{ak}} = \sum_{k \in s_{a1}} \frac{z_k}{\pi_{ak}} + \sum_{k \in s_{a2}} \frac{x_k}{\pi_{ak}}$$

with $s_a = s_{a1} \cup s_{a2}$ and $s_{a1} \subset U_1$, $s_{a2} \subset U_2$.

Since (4) is an unbiased estimator, its MSE can be written as

$$\mathrm{MSE}(\hat{t}_z) = \mathrm{var}(\hat{t}_y) + \mathrm{var}(\hat{\mathrm{B}}(\hat{t}_y)) - 2\mathrm{cov}(\hat{t}_y, \hat{\mathrm{B}}(\hat{t}_y))\,, \tag{5}$$

where

$$\mathrm{var}(\hat{t}_y) = \sum_{k,l \in U} \sum \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}}\,, \tag{6}$$

$$\mathrm{var}(\hat{\mathrm{B}}(\hat{t}_y)) = \sum_{k,l \in U_2} \sum \Delta_{akl} \frac{e_k}{\pi_{ak}} \frac{e_l}{\pi_{al}} + \mathrm{E}_a \left[\sum_{k,l \in U_2} \sum \Delta_{kl|s_{a2}} I_{ak} I_{al} \frac{e_k}{\pi_{ak}\pi_{k|s_{a2}}} \frac{e_l}{\pi_{al}\pi_{l|s_{a2}}} \right] \tag{7}$$

$$\mathrm{cov}(\hat{t}_y, \hat{\mathrm{B}}(\hat{t}_y)) = \sum_{k \in U} \sum_{l \in U_2} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{e_l}{\pi_{al}}\,, \tag{8}$$

and $\Delta_{kl|s_{a2}} = \pi_{kl|s_{a2}} - \pi_{k|s_{a2}}\pi_{l|s_{a2}}$ with $\pi_{kl|s_{a2}}$ being the second-order inclusion probability in the second phase.

Here (6) is a general variance expression for an estimator under any without-replacement sampling design, (7) is the variance of the total estimator under any without-replacement two-phase sampling design, and (8) is the covariance of two estimators (Särndal et al., 1992). Note that since $e_k = z_k - z_k = 0$, $k \in U_1$, summation over strata disappears in (7) and (8).

An unbiased estimator of (5) is

$$\hat{\mathrm{MSE}}(\hat{t}_z) = \hat{V}(\hat{t}_y) + \hat{V}(\hat{\mathrm{B}}(\hat{t}_y)) - 2\hat{C}(\hat{t}_y, \hat{\mathrm{B}}(\hat{t}_y))\,, \tag{9}$$

where

$$\hat{V}(\hat{t}_y) = \sum_{k,l \in s_a} \sum \frac{\Delta_{akl}}{\pi_{akl}} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}}\,, \tag{10}$$

$$\hat{V}(\hat{\mathrm{B}}(\hat{t}_y)) = \sum_{k,l \in s_2} \sum \frac{\Delta_{akl}}{\pi_{akl}\pi_{kl|s_{a2}}} \frac{e_k}{\pi_{ak}} \frac{e_l}{\pi_{al}} + \sum_{k,l \in s_2} \sum \frac{\Delta_{kl|s_{a2}}}{\pi_{kl|s_{a2}}} \frac{e_k}{\pi_{ak}\pi_{k|s_{a2}}} \frac{e_l}{\pi_{al}\pi_{l|s_{a2}}}\,,$$

$$\hat{C}(\hat{t}_y, \hat{\mathrm{B}}(\hat{t}_y)) = \sum_{k \in s_a} \sum_{l \in s_2} \frac{\Delta_{akl}}{\pi_{akl}} \frac{y_k}{\pi_{ak}} \frac{e_l}{\pi_{al}\pi_{l|s_{a2}}}\,. \tag{11}$$

Each term in (9) is an unbiased estimate of the corresponding term in (5).

Probability sampling of units for editing and estimating the bias completes the second step of the procedure.

3 One Example

In this example, a specific two-phase sampling design with simple random sampling in the first phase and Poisson sampling in the second is considered. The use of Poisson design in the second phase is advantageous in many ways. It allows units to be sampled simultaneously with data collection, and different inclusion probabilities can be assigned to the units to reflect the likelihood and influence of errors. In addition, the independent sampling of units simplifies the derivation of variance formulae.

For this example, the first-order inclusion probabilities and covariances are

$$\begin{aligned}
\pi_{ak} &= \frac{n_a}{N} = f_a\,, \\
\Delta_{akl} &= -f_a\frac{1-f_a}{N-1}\,, \qquad k \neq l \\
\Delta_{akk} &= f_a(1-f_a)\,, \\
\pi_{k|s_{a2}}\,, & \\
\Delta_{kl|s_{a2}} &= 0\,, \qquad k \neq l \\
\Delta_{kk|s_{a2}} &= \pi_{k|s_{a2}}(1-\pi_{k|s_{a2}})\,.
\end{aligned}$$

Now, the unbiased estimator of the total is

$$\hat{t}_z = \frac{N}{n_a}\left[\sum_{k\in s_a} y_k - \sum_{k\in s_2}\frac{e_k}{\pi_{k|s_a}}\right]$$

and its MSE is given by (5), where

$$\begin{aligned}
\operatorname{var}(\hat{t}_y) &= \frac{(1-f_a)N^2}{n_a}S^2_{yU}\,, \qquad (12)\\
\operatorname{var}(\hat{\mathrm{B}}(\hat{t}_y)) &= \frac{(1-f_a)N^2}{n_a(N-1)}\left[(N_2-1)S^2_{eU_2} + \left(\frac{1}{N_2}-\frac{1}{N}\right)\mathrm{B}^2(\hat{t}_y)\right] \\
&\quad + \frac{N}{n_a}\sum_{k\in U_2}\frac{1-\pi_{k|s_{a2}}}{\pi_{k|s_{a2}}}{e_k}^2\,, \\
\operatorname{cov}(\hat{t}_y, \hat{\mathrm{B}}(\hat{t}_y)) &= \frac{(1-f_a)N^2}{n_a(N-1)}\left[(N_2-1)(S^2_{xU_2} - S_{xzU_2})\right] \\
&\quad + \frac{(1-f_a)N^2}{n_a(N-1)}\left[\mathrm{B}(\hat{t}_y)\left(\frac{1}{N_2}-\frac{1}{N}\right)\sum_{k\in U_2} x_k - \frac{1}{N}\sum_{k\in U_1} z_k\right].
\end{aligned}$$

Here

$$S^2_{yU} = \frac{1}{N-1}\left(\sum_{k\in U} y_k - \frac{1}{N}\sum_{k\in U} y_k\right)^2$$

is the variance of the variable y in the population U, $S^2_{eU_2}$ and $S^2_{xU_2}$ is the variance of the measurement error e and the variable x, respectively, in the population U_2 and S_{xzU_2} is the covariance between x and z in U_2.

The unbiased estimator of MSE is given by (9), where

$$\begin{aligned}
\hat{\text{var}}(\hat{t}_y) &= \frac{(1-f_a)N^2}{n_a} S^2_{ys_a}\,, \qquad (13)\\
\hat{\text{var}}(\hat{\text{B}}(\hat{t}_y)) &= \frac{(1-f_a)N^2}{n_a(n_a-1)}\left[(n_2-1)S^2_{\breve{e}s_2} + \left(\frac{1}{n_2}-\frac{1}{n_a}\right)\left(\sum_{k\in s_2}\breve{e}_k\right)^2\right]\\
&\quad + \frac{(1-f_a)N^2}{n_a(N-n_a)}\sum_{k\in s_2}(1-\pi_{k|s_{a2}})\breve{e}_k^2\,,\\
\hat{\text{cov}}(\hat{t}_y, \hat{\text{B}}(\hat{t}_y)) &= \frac{(1-f_a)N^2}{{n_a}^2}\left[\sum_{k\in s_2} x_k\breve{e}_k - \frac{1}{n_a-1}\sum_{k\in s_a} y_k \sum_{s_2}\breve{e}_k\right],
\end{aligned}$$

where

$$\breve{e}_k = \frac{e_k}{\pi_{k|s_{a2}}} \qquad \text{and} \qquad S^2_{\breve{e}s_2} = \frac{1}{n_2-1}\left(\sum_{k\in s_2}\breve{e}_k - \frac{1}{n_2}\sum_{k\in s_2}\breve{e}_k\right)^2.$$

3.1 Simulation study

A simulation study is performed to compare the selective editing approach with the described two-step procedure. These two editing procedures can be compared by examining the performance of two estimators, $\hat{t}_y$ (selective editing) and $\hat{t}_z$ (the two-step procedure).

A population of size 10000 consisting of true values z and observed values x is generated as follows:

$$\begin{aligned}
z &\sim \text{Poisson}(5)\,,\\
x &= \begin{cases} z & \text{with probability } p_1 = 0.6\\ \text{Poisson}(2) & \text{with probability } p_2\\ \text{Poisson}(10) & \text{with probability } p_3\,. \end{cases} \qquad (14)
\end{aligned}$$

Three different cases of values p_2 and p_3 are considered. In case 1, $p_2 = p_3 = 0.2$, in case 2, $p_2 = 0.4$, $p_3 = 0$, and in case 3, $p_2 = 0$, $p_3 = 0.4$. Cases 2 and 3 correspond to the situations in which observed values are systematically smaller or larger, respectively, than true values.

For selective editing (SE), the following setup is used.

1. A global score s_k is computed for all units in the sample and the score values are used to distinguish possibly erroneous and influential responses. To simulate the use of global scores, the score function, $s_k(w) = w_k - \mu_w$, is constructed based on variable w correlated with the study variable, x. Variable w is generated as $w = x + v$, where $v \sim \text{Poisson}(\theta)$. It can be expected that the properties of selective editing with reference to the estimates of a single population total depend on the strength of the relationship between the local scores of the variable and the global scores. The correlation between the global score and the study variable is $\rho = \text{cor}(w, x) = (1 + \theta/\sigma_X^2)^{-1/2}$, where θ is the mean value of w and σ_X^2 denotes the variance of x. The mean, θ, describes the desired level of correlation between

the global score function and the variable x. Selective editing based on local score corresponds to the case of $\theta = 0$.

2. All responses not satisfying condition $|s_k(x)| \leq C$, where C is cutoff value, will be checked and, if necessary, corrected. The constant, C, is fixed before sampling and chosen so that the desired number of responses for editing is obtained.

The setup of two-step procedure (TSP) is following:

1. Selective editing as described above is carried out in the first step, except the larger cutoff value C is used in order to reduce the amount of responses to be checked.

2. For the second step, a random sample according to a Bernoulli sampling design is drawn and all selected units are examined for errors. The inclusion probability in the second phase corresponds to the proportion of units to be edited.

In two-step procedure, the value of C is varied to examine how the amount of editing in the second step effects the properties of the final estimates. Three different C values are considered and results are reported in Table 1.

Simulation results are obtained for a sample size of $n = 1000$ and for 1000 replications. Also, the population generated from model (14) is kept fixed over replications. To compare the two editing procedures, approximately the same number of responses is examined under both selective editing and the two-step procedure.

Table 1 gives the empirical bias and precision measures of the estimators $\hat{t}_y$ (SE) and $\hat{t}_z$ (TSP) in case 1 of model (14). Results are presented for different levels of correlation (ρ) between the global score and the study variable, and for different levels of selective and probability-based editing.

Table 1 shows unbiased estimates for the TSP estimator, as expected. The bias of the SE estimator is small but increases as the correlation ρ decreases. In terms of RMSE, the SE estimator is more efficient than the TSP estimator, with exception for only one case, $\rho = 0.33$ and all units chosen for editing are selected randomly. There is a general pattern that the TSP estimator has smallest RMSE when all edited units are selected at random. This estimator works, in terms of RMSE, equally well as the SE estimator for the two smallest correlation levels, $\rho = 0.5$ and $\rho = 0.33$.

The bias problem obtained by a pure selective editing procedure is underscored by the prediction intervals reported in Table 1. The prediction intervals represent the sampling distributions of the estimators. The intervals reported indicate sampling distributions with almost all probability masses located to the right (left) of the true population total ($t_z = 50032$) when $\rho < 1$ ($\rho = 1$).

The results of the simulations of cases 2 and 3 of model (14) are reported in Table 2. Here the results are quite different. The TSP estimator is still indicated to be unbiased, as expected, and the efficiency of the estimator in terms of RMSE is of the same level as observed in Table 1. However, the SE estimator is associated with very large biases and RMSE values. Biases are between 18% and 32% and the RMSE values are between 3 and 4 times larger then those of the TSP estimator. There is no clear pattern of the effect of the correlation ρ in cases 2 and 3.

Table 1: Empirical bias, precision measure (RMSE = Root Mean Square Error) and 95% prediction interval (PI) for estimators $\hat{t}_y$ (SE) and $\hat{t}_z$ (TSP) in case 1 ($p_2 = p_3 = 0.2$) for different levels of ρ. Three cases of probability editing are considered in the two-step procedure: C_1, C_2, C_3.[a]

		SE	TSP		
			C_1	C_2	C_3
$\rho = 1.0$					
	Edited[b]	**12%**+0%	**7%**+5%	**4%**+8%	**0%**+12%
	B($\hat{t}$)	2%	0%	0%	0%
	RMSE	1406	2964	2858	2852
	95% PI	48615±1480	49977±5808	49724±5603	49791±5589
$\rho = 0.7$					
	Edited[b]	**16%**+0%	**11%**+5%	**4%**+12%	**0%**+16%
	B($\hat{t}$)	2%	0%	0%	0%
	RMSE	1485	3791	2798	2525
	95% PI	51018±1669	49854±7430	49823±5485	49718±4948
$\rho = 0.5$					
	Edited[b]	**14%**+0%	**11%**+3%	**6%**+8%	**0%**+14%
	B($\hat{t}$)	4%	0%	0%	0%
	RMSE	2256	4740	3075	2509
	95% PI	51859±1813	49659±9291	49903±6027	49996±4919
$\rho = 0.3$					
	Edited[b]	**17%**+0%	**12%**+5%	**7%**+10%	**0%**+17%
	B($\hat{t}$)	6%	0%	0%	0%
	RMSE	2967	4416	3038	2345
	95% PI	52619±1825	50041±8655	49881±5955	49732±4596

[a] C_1, C_2, and C_3 were chosen to yield approximately 10%, 5%, and 0% of observations, respectively, to be edited in the first step.

[b] The first percentage in bold in the "Edited" row gives the proportion of observations examined in the first step and the second percentage in non-bold is the proportion examined in the second step of the procedure.

4 Discussion

This paper proposes that selective editing be replaced by a two-step procedure in which selective editing is carried out in the first step and a randomly selected set of observations is edited in the second. The purpose is to obtain an editing procedure based on statistical inference principles, which provide a means to control the properties of the estimator. The procedure is treated as a two-phase sampling design and a bias-corrected HT estimator is suggested. Its variance and a variance estimator are derived. In an example, expressions of estimators and variance are derived under an SI design for sample selection and a Poisson sampling of unedited sample units. The special case of Bernoulli sampling of unedited units is considered in the simulation study.

The results of the simulation study often favor the two-step procedure over selective editing. The results indicate that the bias-corrected estimator is more precise than the corresponding biased estimator based on pure selective editing. In fact, some results

Table 2: Empirical bias and RMSE for estimators $\hat{t}_y$ (SE) and $\hat{t}_z$ (TSP) in case 2 ($p_2 = 0.4$, $p_3 = 0$), and in case 3 ($p_2 = 0$, $p_3 = 0.4$) for different levels of ρ^b

		Case 2		Case 3	
		SE	TSP	SE	TSP
$\rho = 1.0$					
	Edited[b]	**14%**+0%	**4%**+10%	**11%**+0%	**3%**+8%
	B($\hat{t}$)	−18%	0%	24%	0%
	RMSE	9117	2380	12035	3801
$\rho = 0.7$					
	Edited[b]	**14%**+0%	**4%**+10%	**15%**+0%	**4%**+11%
	B($\hat{t}$)	−21%	0%	28%	0%
	RMSE	10301	2374	14061	3515
$\rho = 0.5$					
	Edited[b]	**15%**+0%	**4%**+11%	**19%**+0%	**4%**+15%
	B($\hat{t}$)	−20%	0%	29%	0%
	RMSE	10209	2399	14695	2951
$\rho = 0.3$					
	Edited[b]	**17%**+0%	**5%**+12%	**17%**+0%	**5%**+12%
	B($\hat{t}$)	−20%	0%	32%	0%
	RMSE	9948	2171	16057	3478

[b] See the footnote in Table 1.

favor the using probability sampling only when selecting observations for editing. That is, better estimator precision in terms of MSE is obtained when the selective editing stage is excluded from the two step procedure.

Implementing a two-step editing procedure does not call for more resources than pure selective editing does, because the same number of responses can be edited. Timeliness can also be preserved, as the random selection of observations in the second step can be made at the same time as the selective editing is being done. This is one advantage of using Poisson sampling in the second phase. Another advantage is that the two-phase approach provides a means for valid inference when data are used for secondary purposes. Without a random complement of observations, the effect of selective editing has to be judged based on the analysis of historical data. Finally, without historical data, the two-step procedure provides a way to reduce the editing of surveys.

Our paper is only a first study of the properties of a combined selective editing–probabilistic editing approach. The results presented are promising, but the two-step procedure needs further exploration. One problem is considering alternate estimators to the HT estimator, for example, bias correction of the generalized regression estimator. Another problem is the sampling design used for selecting observations for editing. A Bernoulli sampling scheme was used in the simulations in this paper, but we would expect even better properties if Poisson sampling were considered with inclusion probabilities proportional to the scores for selective editing.

Acknowledgements:

We would like to thank Dan Hedlin and an anonymous referee for constructive critics and useful suggestions.

References

Granquist, L., and Kovar, J. G. (1997). Editing of survey data: How much is enough? In L. E. Lyberg et al. (Eds.), *Survey Measurement and Process Quality* (p. 415-435). New York: Wiley.

Hedlin, D. (2003). Score functions to reduce business survey editing at the U.K. office for national statistics. *Journal of Official Statistics*, *19*, 177-199.

Hoogland, J. (2002). Selective editing by means of plausibility indicators. *Proceedings of Conference of European Statisticians*. (UNECE Work Session on Statistical Data Editing, Helsinki, Finland. Working Paper no. 33)

Latouche, M., and Berthelot, J.-M. (1992). Use of score function to prioritize and limit recontacts in editing 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.

Lessler, J. T., and Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: Wiley.

Madow, W. G. (1965). On some aspects of response error measurement. *Proceedings of the Social Statistics Section, ASA*, *16*, 182-192.

Rao, J. N. K., and Sitter, R. R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. In L. E. Lyberg et al. (Eds.), *Survey Measurement and Process Quality* (p. 753-768). New York: Wiley.

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

Statistics Canada. (2003). The use of a score function in a data collection context. *Proceedings of Conference of European Statisticians*. (UNECE Work Session on Statistical Data Editing, Madrid, Spain. Working Paper no. 28)

Statistics Sweden. (2005). A selective editing method considering both suspicion and potential impact, developed and applied to the swedish foreign trade statistics. *Proceedings of Conference of European Statisticians*. (UNECE Work Session on Statistical Data Editing, Ottawa, Canada. Working Paper no. 12)

Authors' addresses:

Maiki Ilves and Thomas Laitila
Department of Statistics
Örebro University
SE-701 82 Örebro, Sweden
E-mail: maiki.ilves@oru.se and thomas.laitila@oru.se