Discussion paper

A note on efficient audit sample selection

Laura Boeschoten
Sander Scholtus
Arnout van Delden

August 2023

Content

Summary

Statistical output, e.g. totals or means of a target variable, are often published for subpopulations that are defined by categorical domain variables (such as categories of educational level, categories of economic activity, and so on). When these domain variables are constructed as a statistical register by combining various sources prone to errors, it is important to check the quality of these variables. A way to do this is to perform an audit on a sample of that population, that is representative with respect to the domain variable. When a possibly non-random sample (a non-probability sample) of units has already been audited earlier, it would be most efficient to re-use as many of these already audited units as possible. In order to achieve a sample that is representative with respect to the domain variable and containing as many previously audited units as possible, standard sampling techniques are not sufficient. In this paper, a method is introduced that selects an audit sample which re-uses previously audited cases by considering the selection of an audit sample that is representative with respect to domain variables as a constrained minimization problem. Furthermore, the performance of this method is evaluated by means of a simulation study, and the method is applied to draw an audit sample of establishments to evaluate the quality of the establishment register that is used to produce statistics on energy consumption per economic activity.

Keywords

Audit sample, validation sample, gold standard, non-probability sample, deviance, minimization procedure

1. Introduction

Since the introduction of the sustainable development goals (UN General Assembly, 2015), the interest in statistics regarding energy consumption has increased substantially, including many statistics which have never been produced before. In the Netherlands, there has recently been interest in statistics on energy consumption per economic sector, which Statistics Netherlands aims to calculate by summing up the energy consumption of all establishments that operate within an economic sector. In order to produce such statistics, a statistical register is required that contains all establishments in the Netherlands, their energy consumption and the economic sector in which they operate.

Table 1: Illustration of how energy connections, addresses and establishments can be related

Address	Energy connection	Establishment
Address 1	Connection 1 (gas)	Establishment 1
Address 1	Connection 2 (electricity)	Establishment 1
Address 2	Connection 3 (gas)	Establishment 2
Address 2	Connection 3 (gas)	Establishment 3
Address 2	Connection 4 (electricity)	Establishment 2
Address 2	Connection 4 (electricity)	Establishment 3

To construct such a statistical register, currently multiple (incomplete) administrative sources are combined on the unit level, where a unit is an energy connection. Note that one address can host multiple energy connections and multiple establishments, as illustrated in Table 1. These different statistical registers contain in addition information regarding the economic activity of the establishment, and the observed economic activity of an establishment can vary over the different administrative sources, as they are prone to error. The economic activity is classified according to the NACE rev. 2 codes (Eurostat, 2008) and is referred to as the ‘domain variable’ in the remainder of this manuscript.

As the domain variable can contain errors (i.e. is error-prone), it is important to evaluate the quality of this classification variable in the newly constructed statistical register, which can for example be evaluated by means of an audit. In general, it is desired that such an audit takes place on a sample of the population register that is representative with respect to the domain variables, such as the economic sector, so that the quality can be assessed for the different sectors. An audit performed on a sample representative with respect to the domain variable allows us to estimate the quality of the statistical output created using this domain variable. In addition, an audit can help us to estimate the amount of misclassification present in the various administrative sources which contain this variable.

Throughout the year, a subset of units is typically already audited for other purposes, and this subset is not representative with respect to the domain variable, economic activity. Given that auditing an establishment is a burdensome and expensive exercise, it is desired to re-use as many as possible of the previously audited units in this new audit sample that is representative with respect to the distribution of economic sectors. To achieve this goal, we propose to approach the selection of an audit sample as a constrained minimization problem. In this paper, we propose a framework where an audit sample can be selected, representative with respect to domain variables, that re-uses as many earlier audited units as possible. In Section 2, notation and the proposed framework are introduced. In Section 3, it is illustrated how the framework can be used to re-use previously audited units. In Section 4, a simulation study is used to investigate the performance of the method under different conditions. Section 5 illustrates how the proposed method can be applied in practice using ready-made R scripts. Finally, section 6 concludes.

Besides the field of energy statistics, auditing is a widely used method for quality improvement and is applied in many fields of research. Within the field of official statistics, audits have also been used to estimate the quality of administrative sources or of surveys by including an audit sample in a structural equation model (Scholtus et al., 2015; Sobel & Arminger, 1986). Furthermore, audits are used in the field of clinical research to evaluate the performance of diagnostic tests in comparison to a 'gold standard' test, see for example Chataway et al. (2004). In addition, organizations are often required by law to perform financial audits (TheWorldBank, 2019), see for example Derks et al. (2019) and Elder et al. (2013). When evaluating prediction models, validation samples are used to evaluate their performance, see for example Hernandez et al. (2014). Here, researchers sometimes deal with the issue of 'sample selection bias' (Zadrozny, 2004), see for example Klingwort et al. (2021). As researchers in these respective fields typically also deal with all sorts of challenges related to their audit sample, we hope that our solution can also be of relevance for them.

2. Background

We are interested in auditing a sample that is representative with respect to the target population on a domain variable. When drawing the audit sample, an error-prone measure of this domain variable is available for the complete target population. More specifically, it is the goal of the audit to estimate the error probabilities of this domain variable.

In what follows, all variables are assumed to be categorical. The target population is denoted as U, with $U = 1, 2, \dots, N$. The true variable of interest is denoted by W, with W_g denoting the value of this variable for element g in U. The aim is to estimate certain parameters of the distribution of W, for instance, its frequency distribution in the target population. It is assumed that W is observed only inside the audit sample. For all units in the target population U, an error-prone version of the variable of interest is measured, which is denoted by X. The values of X in the target population are denoted by $X_1, X_2, \dots, X_N$. In addition to estimating the distribution of W, a secondary aim may be to estimate the association between W and X, e.g., the error probabilities $\Pr\left(X_g = x \middle| W_g = w\right)$. Knowledge of these error probabilities makes it possible to estimate the distribution of W based on the target population, not just the audit sample, which is more efficient. In addition, these error probabilities may be of interest in their own right, as quality measures of the error-prone indicator X. In addition to X, we suppose that one or more covariates are available in the observed sample, collectively denoted as Y, with values $Y_1, Y_2, \dots, Y_N$. We assume that the joint distribution of Y is known (or previously estimated) for the target population and that covariate values are observed without measurement error. Let Z denote the selection indicator of the audit sample within U, where Z_g is the value for element g with $g = 1, 2, \dots, N$: $Z_g = 1$ when the element is included in the sample and $Z_g = 0$ otherwise. The expected value of Z_g with respect to the sample design is denoted by

$$\pi_g = E(Z_g) = \Pr(Z_g = 1). \tag{1}$$

The audit sample size, denoted by n, is equal to the sum of the values of Z:

$$n = \sum_{g=1}^{N} Z_g. \tag{2}$$

In order to make appropriate inferences with respect to the target population, it is important that the audit sample can be used to obtain (approximately) unbiased estimates of target population parameters. If this is possible, we will refer to the sample as representative for the purpose at hand. For any audit sample that is drawn by a random mechanism, in such a way that the probabilities π_g from Equation (1) are known and positive for all units in the population, it is known from design-based sampling theory how to obtain an unbiased estimator; see, e.g., Cochran (1977) or Särndal et al. (1992). Here, we will focus on the more complicated situation that (initially) an audit sample is available for which the probabilities π_g are not known or not useful. This may occur, for instance, because we are given a sample of previously audited units that we did not draw ourselves, or because a non-random selection mechanism was used (e.g., selecting only the largest units in the population). The term *non-probability sampling* is often used for this situation (Baker et al., 2013). In general, a non-probability sample is not representative, since there is no available method that guarantees unbiased estimation of population parameters. There is a growing literature on methods that attempt to make valid inference from a given non-probability sample, sometimes by combining it with a probability sample; see, e.g., Elliott and Valliant (2017) and Rao (2021) for an overview. Here, we propose a different approach and try to adapt the initial audit sample, possibly by both adding and removing cases, to make it more amenable to design-based estimation.

As a theoretical starting point to investigate whether the audit sample is representative with respect to the domain variables and covariates of the target population, the joint distribution of the variables (W, X, Y, Z) in the target population is of interest. Here W and X represent the true domain variable and its error-prone observed version, Z indicates whether a unit is included in the audit sample and Y represents a covariate. We make the following simplifying assumption:

Assumption A. There is no direct association between the true domain variable W and the audit inclusion indicator Z, once X and Y are accounted for. That is to say,

$$\Pr(Z_g = z \big| X_g = x, Y_g = y, W_g = w) = \Pr(Z_g = z \big| X_g = x, Y_g = y),$$

for all possible values (w, x, y, z).

For any audit sample that we draw ourselves, we can ensure that this assumption is satisfied. However, if we are given a non-probability sample of previously audited units, the assumption may require that the right covariates are included in Y. The importance of this assumption which will be seen below is that it allows the representativeness of the audit sample to be investigated by analyzing the joint distribution of (X, Y, Z) instead of (W, X, Y, Z) (see the end of Section 2.1). In practice, the former distribution is known, whereas inference about the latter distribution becomes possible only once a representative audit sample has been obtained.

It does not matter if the distribution of Z depends on X, as long as differences in probabilities to be included in the audit are completely related to the covariates in Y for which the distribution in the target population is known:

$$\Pr(Z_g = 1 \big| X_g = x, Y_g = y) = \Pr(Z_g = 1 \big| Y_g = y). \tag{3}$$

From a missing data perspective, it can be said that if the distribution of Z depends on X, this is not problematic as long as the audit exclusion pattern is Missing At Random (MAR) (Rubin, 1976). However, if the distribution of Z depends on X and this dependence is not explained by Y, this can be problematic, in particular when the error probabilities $\Pr(X_g = x \big| W_g = w)$ are of interest. Then, it can be said that the audit exclusion pattern is Missing Not At Random (MNAR) (Rubin, 1976). Therefore, our goal is to select a final audit sample in such a way that Equation (3) holds; we refer to this as the *MAR requirement of representativeness*.

2.1 Deviance as a criterion for representativeness

We propose to test whether a given audit sample meets the MAR requirement of representativeness (3) by analyzing the joint distribution (X, Y, Z) in the observed target population using a non-saturated log-linear model $(XY)(YZ)$. Note that the difference between the non-saturated model and a saturated model is that the direct association (XZ) and the three-way interaction (XYZ) are excluded. In the non-saturated model it is assumed that X and Z are conditionally independent given Y, i.e. the observed domain characteristic and the audit inclusion probability are conditionally independent given the covariates under consideration. Therefore, we refer to it as the *conditional independence (CI) model*.

We remark that the CI model is equivalent to condition (3). In terms of probabilities the CI model assumes that

$$\Pr\left(Z_g = z, X_g = x \middle| Y_g = y\right) = \Pr\left(Z_g = z \middle| Y_g = y\right) \Pr\left(X_g = x \middle| Y_g = y\right).$$

Furthermore, it holds in general that, for all (x, y) with $\Pr\left(X_g = x, Y_g = y\right) > 0$,

$$\begin{aligned}\Pr\left(Z_g = z \middle| X_g = x, Y_g = y\right) &= \frac{\Pr\left(Z_g = z, X_g = x, Y_g = y\right)}{\Pr\left(X_g = x, Y_g = y\right)} \\ &= \frac{\Pr\left(Z_g = z, X_g = x \middle| Y_g = y\right)}{\Pr\left(X_g = x \middle| Y_g = y\right)}.\end{aligned}$$

Under the CI assumption, the latter expression reduces to

$$\Pr\left(Z_g = z \middle| X_g = x, Y_g = y\right) = \Pr\left(Z_g = z \middle| Y_g = y\right).$$

It follows that the CI model implies condition (3) and vice versa. Hence, a test of the CI model is also a test of condition (3).

The actual number of observations in cell $(X = i, Y = j, Z = k)$ is denoted as n_{ijk} and the number of observations estimated by the CI model is denoted as $\hat{n}_{ijk}$. The fit of the CI model can be measured by the likelihood ratio test statistic or deviance (D) comparing it to the saturated model where $\hat{n}_{ijk} = n_{ijk}$ (Agresti, 2013):

$$D \;=\; 2\sum_{i,j,k} n_{ijk} \log n_{ijk} - 2\sum_{i,j,k} n_{ijk} \log \hat{n}_{ijk}, \qquad (4)$$

where $D \geq 0$. Throughout this paper, we use the convention that $0 \log 0 = 0$.

A larger value for D indicates a stronger conditional dependence between X and Z given Y in the observed data-set, which means that the problem of not having a representative audit sample is more substantive. This suggests that an audit sample should be selected for which D is 'sufficiently small'. For the CI model, the estimated value for $\hat{n}_{ijk}$ can be calculated directly, without the use of an iterative algorithm, as

$$\hat{n}_{ijk} = \frac{n_{ij+}\, n_{+jk}}{n_{+j+}}, \tag{5}$$

see for example Agresti (2013) or Bishop et al. (1975). Here, $n_{ij+} = \sum_k n_{ijk}$, $n_{+jk} = \sum_i n_{ijk}$ and $n_{+j+} = \sum_{i,k} n_{ijk}$.

The definition of D can be expressed alternatively by incorporating (5):

$$\begin{aligned} D &= 2\sum_{i,j,k} n_{ijk} \log n_{ijk} - 2\sum_{i,j} n_{ij+} \log n_{ij+} - 2\sum_{j,k} n_{+jk} \log n_{+jk} \\ &\qquad + 2\sum_{j} n_{+j+} \log n_{+j+} \\ &= C + 2\sum_{i,j,k} n_{ijk} \log n_{ijk} - 2\sum_{j,k} n_{+jk} \log n_{+jk}, \end{aligned} \tag{6}$$

where

$$C = 2\sum_{j} n_{+j+} \log n_{+j+} - 2\sum_{i,j} n_{ij+} \log n_{ij+} \tag{7}$$

is a constant term which depends only on the distribution of (X, Y) and therefore will be the same for every possible choice of audit sample. For the second term in the first line of Equation (6) we used that $\sum_{i,j,k} n_{ijk} \log n_{ij+} = \sum_{i,j} n_{ij+} \log n_{ij+}$; for the third and fourth term alike.

Recall that the full (unobserved) distribution of interest is (W, X, Y, Z). We will now show that under Assumption A (there is no direct association between W and Z), the selectivity of the audit sample can be analysed using the deviance defined in Equation (6), which is based on the observed distribution (X, Y, Z).

Let n'_{hijk} denote the number of observations in cell $(W = h, X = i, Y = j, Z = k)$, with n_{ijk} = n'_{ijk} = $\sum_h n'_{hijk}$. Under Assumption A, the maximal hierarchical log linear model for n'_{hijk} is $(WXY)(XYZ)$. For this model, the predicted values n'_{hijk} can also be obtained directly:

$$\hat{n}'_{hijk} = \frac{n'_{hij+}\, n'_{+ijk}}{n'_{+ij+}} = \frac{n'_{hij+}\, n_{ijk}}{n_{ij+}};$$

see Bishop et al. (1975).

Hence, the deviance that compares this model to the saturated model with $\hat{n}'_{hijk} = n_{hijk}$ can be written analogously to Equation (6) as

$$\begin{aligned} D_{W,max} = 2\sum_{h,i,j,k} n'_{hijk} \log n'_{hijk} - 2\sum_{h,i,j} n'_{hij+} \log n'_{hij+} \\ - 2\sum_{i,j,k} n_{ijk} \log n_{ijk} + 2\sum_{i,j} n_{ij+} \log n_{ij+}. \end{aligned}$$

The maximal model contains a direct association between X and Z, pointing to the fact that the audit sample is not representative with respect to the domain variable. The analogue of the CI model for (W, X, Y, Z) is given by the log linear model $(WXY)(YZ)$. For this model, the predicted values satisfy:

$$\hat{n}'_{hijk} = \frac{n'_{hij+}\, n'_{++jk}}{n'_{++j+}} = \frac{n'_{hij+}\, n_{+jk}}{n_{+j+}};$$

see Bishop et al. (1975).

Hence, the deviance that compares this CI model to the saturated model for (W, X, Y, Z) can be written as

$$\begin{aligned} D_W = 2\sum_{h,i,j,k} n'_{hijk} \log n'_{hijk} - 2\sum_{h,i,j} n'_{hij+} \log n'_{hij+} \\ - 2\sum_{j,k} n_{+jk} \log n_{+jk} + 2\sum_{j} n_{+j+} \log n_{+j+}\,. \end{aligned}$$

However, under Assumption A the saturated model is too large, and it makes more sense to compare the fit of the CI model to that of the above maximal model. This can be done using the conditional deviance for nested models (Bishop et al., 1975), which in this case is given by:

$$\begin{aligned} D_W - D_{W,max} = 2\sum_{i,j,k} n_{ijk} \log n_{ijk} - 2\sum_{i,j} n_{ij+} \log n_{ij+} \\ - 2\sum_{j,k} n_{+jk} \log n_{+jk} + 2\sum_{j} n_{+j+} \log n_{+j+}. \end{aligned}$$

Clearly, $D_W - D_{W,max} = D$ in Equation (6). Hence, under Assumption A, to analyze the fit of the CI model compared to the maximal model for the full distribution (W, X, Y, Z), it suffices to analyse the fit of the model $(XY)(YZ)$ for the observed distribution (X, Y, Z).

3. Method

In this section, we introduce how an audit sample can be selected that is representative with respect to domain variables by using the deviance defined in Equation (6) as a criterion for representativeness. The proposed method has three main advantages. First, the auditor can determine a maximum number of units to be audited. Here, the proposed method will provide an audit sample that is as representative as possible given that maximum number. Second, if some units have already been audited or the auditor is planning to include specific units in the audit sample, the method attempts to include as many of these units into the final audit sample as possible, given the constraint that this final sample should be sufficiently representative. Third, the number of units to be audited can also be limited by setting a deviance-value as a boundary.

3.1 Basic optimization problem

Let n_{ijk} denote the number of observations in cell $(X = i, Y = j, Z = k)$ prior to applying the method in this section. If previously audited units are available, these have $Z = 1$ and so it holds that $n_{ij1} > 0$ for certain cells. If the selection of the audit sample starts with a 'blank canvas', then $n_{ij1} = 0$ and $n_{ij0} = n_{ij+}$ for all i and j. We consider the general situation where additional units may be selected for auditing (moved from $Z = 0$ to $Z = 1$) and previously selected units may be excluded (moved from $Z = 1$ to $Z = 0$). When applying the method, the user should specify a maximum number of additional units to include in the audit sample (M_+) and a maximum number of previously audited units to exclude (M_-). Special cases are obtained by setting one of these bounds to zero. Moreover, in the case of no previously audited units, it automatically holds that $M_- = 0$ and M_+ indicates the maximal size of the audit sample. After applying the method, the adjusted number of units in cell $(X = i, Y = j, Z = k)$ is denoted as m_{ijk}. We write

$$m_{ij1} = n_{ij1} + \delta_{ij}^{+} - \delta_{ij}^{-}, \qquad (8)$$

where δ_{ij}^{+} and δ_{ij}^{-} indicate the number of additional units to audit and the number of previously audited units to exclude with $(X = i, Y = j)$, respectively. Note that during this procedure, values in the marginal table (X, Y) are not adjusted, only units are transported from $Z = 0$ to $Z = 1$ and vice versa. Hence, $m_{ij+} = n_{ij+}$ for all i and j.

Now, the goal is to sample the (additional) units from different cells of table (X, Y, Z) in such a way that D defined in Equation (6) is minimized. This is a minimization problem for which the target function can be written as:

$$D(m) = C + 2 \sum_{i,j,k} m_{ijk} \log m_{ijk} - 2 \sum_{j,k} m_{+jk} \log m_{+jk}, \qquad (9)$$

since C remains equal to Equation (7), as $m_{ij+} = n_{ij+}$ and $m_{+j+} = n_{+j+}$. Therefore, C can be ignored when minimizing $D(m)$.

When minimizing $D(m)$, a number of constraints apply. First, the user-specified bounds on the number of additional units to include (M_+) and the number of units to exclude (M_-) lead to the following constraints:

$$\sum_{i,j} \delta_{ij}^{+} \leq M_+; \tag{10}$$

$$\sum_{i,j} \delta_{ij}^{-} \leq M_-. \tag{11}$$

Second, for each combination $(X = i, Y = j)$, the number of units in- and excluded in the audit sample should add up to the known marginal value:

$$m_{ij1} + m_{ij0} = n_{ij+}, \qquad \begin{array}{l} i = 1, \dots, I; \\ j = 1, \dots, J, \end{array} \tag{12}$$

where I is the number of categories in X and J is the number of categories in Y. Together with m_{ij1} defined in Equation (8), this restriction implies that

$$m_{ij0} = n_{ij0} - \delta_{ij}^{+} + \delta_{ij}^{-}. \tag{13}$$

Third, for each combination $(X = i, Y = j)$, there exist bounds on δ_{ij}^{+} and δ_{ij}^{-} based on the initial counts n_{ijk}, since no more units can be moved from $Z = 0$ to $Z = 1$ or vice versa than are initially available:

$$0 \leq \delta_{ij}^{+} \leq n_{ij0}, \qquad \begin{array}{l} i = 1, \dots, I; \\ j = 1, \dots, J; \end{array} \tag{14}$$

$$0 \leq \delta_{ij}^{-} \leq n_{ij1}, \qquad \begin{array}{l} i = 1, \dots, I; \\ j = 1, \dots, J. \end{array} \tag{15}$$

In combination with the other restrictions, these bounds imply that $m_{ijk} \geq 0$ for every value. In practice, it is desirable to let at most one of δ_{ij}^{+} and δ_{ij}^{-} be non-zero for each combination of i and j; we will return to this point in Section 3.3.

The minimization procedure can now be written as follows:

$$\min \left\{ C + 2 \sum_{i,j,k} m_{ijk} \, \log m_{ijk} - 2 \sum_{j,k} m_{+jk} \, \log m_{+jk} \right\} \tag{16}$$

under constraints

$$\begin{gathered} m_{ij1} = n_{ij1} + \delta_{ij}^{+} - \delta_{ij}^{-}; \\ m_{ij0} = n_{ij0} - \delta_{ij}^{+} + \delta_{ij}^{-}; \\ \sum_{ij} \delta_{ij}^{+} \leq M_+; \end{gathered}$$

$$\sum_{ij} \delta_{ij}^{-} \leq M_{-};$$
$$0 \leq \delta_{ij}^{+} \leq n_{ij0};$$
$$0 \leq \delta_{ij}^{-} \leq n_{ij1}.$$

This is a non-linear minimization problem with linear constraints which can be solved by standard software, for instance using the R function ‘constrOptim’ (R Core Team, 2023). Note that optimization (16) is carried out with respect to the variables δ_{ij}^{+} and δ_{ij}^{-}. The target function depends on these variables through Equation (8) and Equation (13) according to the first two lines of constraints. To find the optimal value efficiently, it is useful to have the gradient of the objective function $D(m)$, i.e., the vector of partial derivatives with respect to the free parameters δ_{ij}^{+} and δ_{ij}^{-}. For the elements of this gradient, we find (using Equations (8) and (13) and some standard calculus):

$$\frac{\partial D(m)}{\partial \delta_{ij}^{+}} = 2\left(\log m_{ij1} - \log m_{ij0} - \log m_{+j1} + \log m_{+j0}\right); \tag{17}$$
$$\frac{\partial D(m)}{\partial \delta_{ij}^{-}} = 2\left(\log m_{ij0} - \log m_{ij1} - \log m_{+j0} + \log m_{+j1}\right); \tag{18}$$

The above optimization problem can be extended in a natural way to a situation in which different costs apply to adding or removing units to the sample in different strata. In some applications these costs might represent the actual financial costs involved in auditing a unit, but in other applications they could reflect user preferences for auditing more or fewer units in certain strata. Within the stratum $(X = i, Y = j)$, let $c_{ij}^{+} > 0$ and $c_{ij}^{-} > 0$ denote the costs of, respectively, adding one unit to or removing one unit from the audit sample. The bounds M_{+} and M_{-} are now used to denote the total budgets that are available for, respectively, adding units to the sample and removing units from the sample. Instead of constraints (10) and (11), we now use the following constraints:

$$\sum_{i,j} c_{ij}^{+}\delta_{ij}^{+} \leq M_{+};$$
$$\sum_{i,j} c_{ij}^{-}\delta_{ij}^{-} \leq M_{-}.$$

The rest of minimization problem (16) remains the same as before.

3.2 A procedure for optimizing the audit sample

In practice, it turns out that the objective function defined in Equation (16) is difficult to minimize, because it has many local minima. To ensure that a global minimum is found, multiple starting values have to be tried in the optimization algorithm.

Algorithm 1: Optimize the audit sample

```
input:  The number of observations $n_{ijk}$ $(i = 1,\dots,I;\ j = 1,\dots,J;\ k = 0,1)$.
1:      Initialize $D_{best}$ (the best deviance found so far) as the value of
        $D$ for the original counts $n_{ijk}$.
2:      Determine the bounds $M_+$ and $M_-$.
3:      Determine $N_{\text{attempts}}$, the maximal number of attempts to search for
        the best solution.
4:      for 1:$N_{\text{attempts}}$ do
5:      Select random starting values for the solution:
        - Draw a value $\widetilde{M}_+$ from a uniform distribution on [$l_+M_+$, $u_+M_+$],
          where $l_+$ and $u_+$ are chosen constants with 0 ≤ $l_+$ < $u_+$ ≤ 1.
        - Assign random starting values $\delta_{ij}^+$ inside their feasible
          intervals such that $\sum_{i,j} \delta_{ij}^+ = \widetilde{M}_+$.
        - Draw a value $\widetilde{M}_-$ from a uniform distribution on [$l_-M_-$, $u_-M_-$],
          where $l_-$ and $u_-$ are chosen constants with 0 ≤ $l_-$ < $u_-$ ≤ 1.
        - Assign random starting values $\delta_{ij}^-$ inside their feasible
          intervals such that $\sum_{i,j} \delta_{ij}^- = \widetilde{M}_-$.
6:      Run the optimization algorithm to solve (16) with these
        starting values.
7:      Compute $D(m)$ for the current solution. If $D(m) < D_{best}$, then set
        $D_{best} := D(m)$ and store the current solution. Otherwise, discard
        the current solution.
8:      end for
9:      If $D_{best}$ is still considered too large, return to step 2 and
        change the bounds $M_+$ and/or $M_-$.
output: The number of additional units to audit $\delta_{ij}^+$ and remove from the
        audit sample $\delta_{ij}^-$ $(i = 1,\dots,I;\ j = 1,\dots,J)$ found for the best solution,
        as well as the value of $D_{best}$.
```

For solving Equation (16), we propose the practical procedure found under Algorithm 1. In step 5 of this algorithm, the constants l_+, u_+, l_- and u_- can be used (optionally) to ensure that random starting values are generated for which $\sum_{i,j} \delta_{ij}^+$ and $\sum_{i,j} \delta_{ij}^-$ lie within more limited ranges than their full feasible intervals $[0, M_+]$ and $[0, M_-]$.[1] We found that this can help in practice to reduce the number of attempts that are needed before the optimal solution is found, because often this optimal solution has $\sum_{i,j} \delta_{ij}^+$ and $\sum_{i,j} \delta_{ij}^-$ closer to the upper limits of their feasible intervals than to 0.

[1] Note that step 5 of the algorithm takes into account the requirement of the R function 'constrOptim' that the starting values have to satisfy all constraints of the optimization problem. It would also be possible to use a different optimization method that does not require the starting values to represent a feasible solution. In that case, another useful choice of starting values may be to set, for each j, m_{ij1}/m_{+j1} and m_{ij0}/m_{+j0} both proportional to the same distribution, say the observed distribution n_{ij0}/n_{+j0} in the original data. This choice is based on the observation that $D(m) = 0$ if m_{ijk} satisfies the CI model, which occurs in particular when $\Pr(X_g = x | Y_g = y, Z_g = 1) = \Pr(X_g = x | Y_g = y, Z_g = 0)$ for all (x, y). (Thanks to Jeroen Pannekoek for this remark.)

To help decide when the best deviance value D_{best} should be considered 'too large' in step 9, it may be noted that under the CI model $(XY)(YZ)$ the deviance asymptotically follows a chi-square distribution with $J \times (I-1)$ degrees of freedom (Agresti, 2013). Hence, the $(1-\alpha) \times 100\%$ percentile $\chi^2_{J(I-1)}(1-\alpha)$ of this distribution (e.g., with $\alpha = .05$) could be used as a cut-off point: we accept the current solution in step 9 when $D_{best} \leq \chi^2_{J(I-1)}(1-\alpha)$. (A possible drawback of using this criterion will be discussed in Section 5.)

Once the optimal values δ^+_{ij} and δ^-_{ij} have been obtained, a representative audit sample can be obtained by applying the following two steps (in parallel) to each stratum $(X = i, Y = j)$:

- If $\delta^+_{ij} > 0$, draw a simple random sample without replacement of size δ^+_{ij} from the n_{ij0} units in this stratum with $Z = 0$. These units are selected for additional auditing, so moved to $Z = 1$.
- If $\delta^-_{ij} > 0$, draw a simple random sample without replacement of size δ^-_{ij} from the (original) n_{ij1} units in this stratum with $Z = 1$. These previously audited units are removed from the audit sample, so moved to $Z = 0$.

Once these steps have been run, the final audit sample consists of all units with $Z = 1$.

3.3 Minimizing the deviance while maximizing the number of re-used cases

The approach introduced in Section 3.2 uses deviance as a primary criterion to select an audit sample, where the sample is selected with the smallest value for deviance. Recycling units that have been in the initial audit sample is considered as a secondary criterion. However, in practice a deviance that is as small as possible is not always essential, and recycling as many cases as possible can be more important. Depending on sample size and number of categories in Y a critical value for deviance can be determined and all selected samples with a deviance smaller than the critical value can be considered as being representative with respect to Y.

More formally stated, the specification of the optimization problem in (16) allows that the optimal solution contains combinations (i, j) where $\delta^+_{ij} > 0$ and $\delta^-_{ij} > 0$ simultaneously. This corresponds to a solution where in the same stratum some units are added to the audit sample while other, previously audited units, are removed from the audit sample. This seems inefficient from a practical point of view, since we are (partly) removing units for which W_g is already known and replacing them by units from the same stratum for which W_g still has to be obtained by auditing.

It is easy to correct this after the solution has been found, by moving to an equivalent solution with

$$\delta^+_{ij,alt} = \max\{0, \delta^+_{ij} - \delta^-_{ij}\},$$
$$\delta^-_{ij,alt} = \max\{0, \delta^-_{ij} - \delta^+_{ij}\}.$$

Clearly, $\delta^+_{ij,alt}$ − $\delta^-_{ij,alt}$ = δ^+_{ij} − δ^-_{ij}, so this corresponds to the same solution in terms of m_{ijk}, but now it holds that $\delta^+_{ij,alt}\delta^-_{ij,alt} = 0$ in all cases, as desired. However, it would be preferable to avoid this type of inefficient solution altogether.

A possible solution is to replace the objective function $D(m)$ of the minimization problem defined in Equation (16) by one of the following alternatives:

$$F_1(m) = D(m) + \lambda \sum_{i,j} \left(\delta^+_{ij} + \delta^-_{ij}\right); \tag{19}$$

$$F_2(m) = D(m) + \exp\left\{-\frac{D(m)}{\kappa}\right\} \sum_{i,j} \left(\delta^+_{ij} + \delta^-_{ij}\right). \tag{20}$$

where λ is a small positive constant (e.g., $\lambda = 0.01$) and κ is a positive constant such that $D(m)/\kappa$ is larger than, say, 10 for any candidate solution with a deviance value for which the CI model would be rejected (as discussed in Section 3.2).

Both alternative objective functions penalize candidate solutions with large total numbers of additional units to audit and previously audited units to remove (i.e., large values of $\sum_{i,j}(\delta^+_{ij} + \delta^-_{ij})$. This should make any candidate solution with $\delta^+_{ij}\delta^-_{ij} > 0$ unattractive, because the value of $F_1(m)$ or $F_2(m)$ can be reduced by replacing $\left(\delta^+_{ij}, \delta^-_{ij}\right)$ by $\left(\delta^+_{ij,alt}, \delta^-_{ij,alt}\right)$ as above. The exponential term in $F_2(m)$ is supposed to ensure that the penalty term becomes relevant only for candidate solutions with small (i.e., acceptable) values of $D(m)$, since achieving an acceptable deviance value remains our primary objective. For large values of $D(m)$, it holds that $F_2(m) \approx D(m)$. The same can be achieved with $F_1(m)$, provided that the constant λ is chosen with some care.

The problem of minimizing (16) with the target function replaced by (19) or (20) can be solved by the same procedure as outlined in Section 3.2. To obtain the gradient values for $F_1(m)$, we simply add a term λ to (17) and (18). For $F_2(m)$ we obtain:

$$\frac{\partial F_2(m)}{\partial x} = \frac{\partial D(m)}{\partial x} + \exp\left\{-\frac{D(m)}{\kappa}\right\}\left\{1 - \frac{1}{\kappa}\frac{\partial D(m)}{\partial x}\sum_{i,j}\left(\delta^+_{ij} + \delta^-_{ij}\right)\right\},$$

where $x = \delta^+_{ij}$ or $x = \delta^-_{ij}$.

3.4 Inference based on the final audit sample

Once the final audit sample has been selected as described at the end of Section 3.2 and observed values W_g have been obtained for all units in this sample, the next step is to use these data to estimate one or more parameters of the target population. For ease of exposition, suppose first that the target parameter is the number of units in the population with $W = w$: $\theta^W_w = \sum_{g=1}^{N} I\left(W_g = w\right)$.

The CI model $(WXY)(YZ)$ is equivalent to a logistic regression model for Z that, besides the constant term, uses only Y as a predictor; cf. Agresti (2013, Section 9.5.1). Hence, under the assumption that the CI model holds for the final audit sample, the probability π_g that unit g is included in this sample depends only on the value of Y_g. That is to say, $\pi_g = n_y/N_y$ for all units with $Y_g = y$, where n_y and N_y denote the number of units with $Y_g = y$ in the audit sample and the target population, respectively. According to standard design-based sampling theory, an unbiased estimator of θ_w^W is therefore given by $\hat{\theta}_w^W = \sum_{g=1}^{N} d_g Z_g I\left(W_g = w\right)$, with the sampling weight $d_g = 1/\pi_g = N_y/n_y$ for all units with $Y_g = y$. In this context, this is a model-based estimator, in the sense that its unbiasedness relies on the CI model — and in particular Assumption A — being true.

In general, the sample size n_y per stratum will be a random variable. However, it follows from the above logistic regression model for Z that the distribution of n_y does not depend on the target parameter θ_w^W: the sample sizes n_y are *ancillary statistics* with respect to these target parameters. As discussed by Holt and Smith (1979), inference should be done conditionally on ancillary statistics. In particular, it is preferable to evaluate the variance of an estimated parameter conditional on any ancillary statistics. Hence, in our case the variance of $\hat{\theta}_w^W$ should be evaluated conditional on the realized sample sizes n_y. Since under the CI model all units in the same stratum based on Y have the same (final) inclusion probability, therefore, when the realized sample sizes n_y are treated as fixed, it makes sense (in the absence of more information) to consider the design of the audit sample as equivalent to a stratified simple random sample. We therefore propose to approximate the variance of θ_w^W by its design-based variance under stratified simple random sampling; see, e.g., Cochran (1977), Särndal et al. (1992) or the examples in the next section.

This approach to inference extends straightforwardly to other target parameters that can be written as a continuously differentiable function of one or more population totals based on W and X. Replacing each population total in this function by a weighted estimate using the sampling weights d_g then leads to an approximately unbiased estimator. The design-based variance of this estimator under stratified simple random sampling can be approximated using Taylor linearization. We refer to Sarndal et al. (1992) for more details.

4. Simulation study

4.1 Simulation approach for deviance and bias

To empirically evaluate the performance of the audit sample selection procedure, we conducted a simulation study using R (R Core Team, 2023). The code used for the simulation study is available here and a small illustrative example of the code applied to one situation is available here.

The main simulation study consists of three sets of conditions. For each set, theoretical populations with varying relationships between W, X, Y and Z are generated, where X and Y have three categories and Z has two categories by design. The first set investigates how the performance of the proposed procedure might be affected by the strength of the relationship between X and Z before the start of the optimization procedure, varying from no relationship to a strong relationship. The second set investigates how the performance might be affected by the strength of the relationship between W and X, varying from a perfect relationship (meaning that observed variable X is a perfect measurement of the outcomes after audit W) to an imperfect and asymmetrical relationship (meaning that observed variable X contains measurement error and that the probability of an incorrect score differs for different scores of the audit variable W). The third set investigates how the performance of the procedure might be affected by the strength of the relationship between W and Y, varying from a strong relationship to a weak relationship. Note that for each set, the conditions are ordered from most desired situation to least desired situation. Each set contains four conditions resulting in a total of twelve simulation conditions, which can all be found in Table 2. Note that when the different relationships between X and Z are investigated, the first conditions listed in Table 2 for (W, X) and (W, Y) are selected and similar approaches are used when investigating (W, X) and (W, Y) to investigate the main effects of these relationships. In the last part of the study, the interactions between the selected relations are investigated by taking the most desired and least desired condition for each relation, and investigating their combinations in a full factorial design.

In all conditions, a joint probability density is generated by multiplying the probabilities listed in Table 2. The probabilities generated here follow the log-linear model $(WX)(WY)(XZ)$. This model is contained within the previously discussed maximal model $(WXY)(XYZ)$ and it contains a term (XZ) that cannot be found in the CI model $(WXY)(YZ)$. Because of this term, it is expected that this model results in a large value for the deviance and it would therefore be beneficial to adjust the audit sample via the proposed minimization procedure. For each of the described conditions, 1000 data-sets of size N = 10 000 are sampled from the generated joint probability density. The expected size of the audit sample is 300 in all conditions, as follows also from the proportions described in Table 2.

Table 2: Overview of the bivariate relationships used to generate data for the simulation study

Simulation condition number												
	X	*Z*			*W*	*X*				*Y*		
		0	1			1	2	3		1	2	3
1	1	.323	.010		1	.333	.000	.000		.267	.033	.033
	2	.323	.010		2	.000	.333	.000		.033	.267	.033
	3	.323	.010		3	.000	.000	.333		.033	.033	.267
2	1	.323	.012		1	.267	.033	.033		.200	.067	.067
	2	.323	.010		2	.033	.267	.033		.067	.200	.067
	3	.323	.008		3	.033	.033	.267		.067	.067	.200

3	1	.323	.015		1	.333	.000	.000		.300	.017	.017
	2	.323	.010		2	.017	.300	.017		.033	.267	.033
	3	.323	.005		3	.033	.033	.267		.050	.050	.233
4	1	.323	.018		1	.300	.017	.017		.267	.033	.033
	2	.323	.010		2	.033	.267	.033		.067	.200	.067
	3	.323	.002		3	.050	.050	.233		.100	.100	.133

Per condition, the procedure as described in Section 3.2 is applied on each generated data-set. The bounds M_+ and M_- that should be determined at step 2 of Algorithm 1 are set at 100 and 10 respectively and the maximal number of attempts for the optimization procedure to search for a solution at step 3 is set at $N_{attempts} = 200$. Note that step 9 (adjusting the bounds M_+ and M_- if the obtained value for the deviance is considered too large) is left out to appropriately compare the results under different conditions. The settings described here remain consistent for all simulation conditions.

We considered two types of target parameters to estimate in this simulation study: the true proportion of units in the population with $W_g = w$ for each category w,

$$P_w^W = \frac{1}{N} \sum_{g=1}^{N} I\left(W_g = w\right), \tag{21}$$

and the proportion of units in the population with true category $W_g = w$ that are observed in category $X_g = x$ (measurement error probabilities):

$$P_{x|w}^{X|W} = \frac{\sum_{g=1}^{N} I\left(W_g = w, X_g = x\right)}{\sum_{g=1}^{N} I\left(W_g = w\right)}. \tag{22}$$

Under the assumption that the CI model holds, P_w^W defined in Equation (21) can be estimated without bias from the audit sample by

$$\hat{P}_w^W = \sum_y P_y^Y \frac{\sum_{g=1}^{n} I\left(W_g = w, Y_g = y\right)}{\sum_{g=1}^{n} I(Y = y)} \equiv \sum_y P_y^Y p_{w|y}^{W|Y}. \tag{23}$$

where P_y^Y is the (known) proportion of units with $Y_g = y$ in the population and $p_{w|y}^{W|Y}$ denotes the (unweighted) observed proportion of cases with $Y_g = y$ in the audit sample that also have $W_g = w$. (Here, for convenience it is assumed that the audit sample consists of the first n units.) Similarly, $P_{x|w}^{X|W}$ defined in Equation (22) can be estimated consistently from the audit sample by:

$$\hat{P}_{x|w}^{X|W}=\frac{\sum_y P_y^Y \frac{\sum_{g=1}^n I\left(W_g=w,X_g=x,Y_g=y\right)}{\sum_{g=1}^n I(Y=y)}}{\sum_y P_y^Y \frac{\sum_{g=1}^n I\left(W_g=w,Y_g=y\right)}{\sum_{g=1}^n I(Y=y)}} \equiv \frac{\sum_y P_y^Y p_{wx|y}^{WX|Y}}{\sum_y P_y^Y p_{w|y}^{W|Y}}. \qquad (24)$$

Equations (23) and (24) are examples of the weighted estimation approach discussed in Section 3.4. In particular, note that (24) is a special case of the so-called combined ratio estimator (Cochran, 1977). We investigate the bias of $\hat{P}_w^W$, $B(\hat{P}_w^W)=E(\hat{P}_w^W-P_w^W)$, and of $\hat{P}_{x|w}^{X|W}$, $B(\hat{P}_{x|w}^{X|W})=E(\hat{P}_{x|w}^{X|W}-P_{x|w}^{X|W})$, while selecting only $Z=1$ both before and after the procedure and compare these results.

4.2 Simulation approach for variance

As discussed in Section 3.4, under the assumption that the CI model holds, we propose to estimate the variances of $\hat{P}_w^W$ and $\hat{P}_{x|w}^{X|W}$ by treating the audit sample as a stratified simple random sample with Y as a stratifying variable. For $\hat{P}_w^W$ this yields the following variance estimator, assuming for simplicity that the sampling fraction in each stratum is small enough so that finite population corrections can be neglected:

$$\widehat{\mathrm{var}}(\hat{P}_w^W)=\sum_y \frac{(P_y^Y)^2}{n_y} p_{w|y}^{W|Y}(1-p_{w|y}^{W|Y}), \qquad (25)$$

where n_y denotes the number of units with $Y_g=y$ in the audit sample. Similarly, under the same assumptions the approximate variance of $\hat{P}_{x|w}^{X|W}$ can be derived from expression (6.51) in Cochran (1977). Assuming that $NP_y^Y \gg 1$ for all y, we obtain the following variance estimator:

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{P}_{x|w}^{X|W})=\frac{1}{(\hat{P}_w^W)^2}\sum_y \frac{(P_y^Y)^2}{n_y}\Big\{&p_{wx|y}^{WX|Y}\left(1-p_{wx|y}^{WX|Y}\right)\\ &+\left(\hat{P}_{x|w}^{X|W}\right)^2 p_{w|y}^{W|Y}\left(1-p_{w|y}^{W|Y}\right)\\ &-2\hat{P}_{x|w}^{X|W} p_{wx|y}^{WX|Y}\left(1-p_{w|y}^{W|Y}\right)\Big\}.\end{aligned} \qquad (26)$$

Note that these design-based variances treat the target population of size N as fixed. To evaluate the performance of these variance estimators, a separate set of simulations was run. Combining the least desired condition of (W,X) in Table 2 (i.e., the fourth condition) with each of the four conditions of (W,Y), four fixed target populations of size N = 10 000 were generated. From each of these populations, 1000 initial audit samples were generated according to the least desired condition of (X,Z) in Table 2. Again, the procedure from Section 3.2 was applied to each audit sample, with the same settings as before. For each final audit sample, we computed $\hat{P}_w^W$ and $\hat{P}_{x|w}^{X|W}$ as well as the associated variance estimates from (25) and (26). This allowed us to compare the estimated variances with the empirical variances of $\hat{P}_w^W$ and $\hat{P}_{x|w}^{X|W}$ across 1000 simulation rounds.

4.3 Results

4.3.1 Deviance

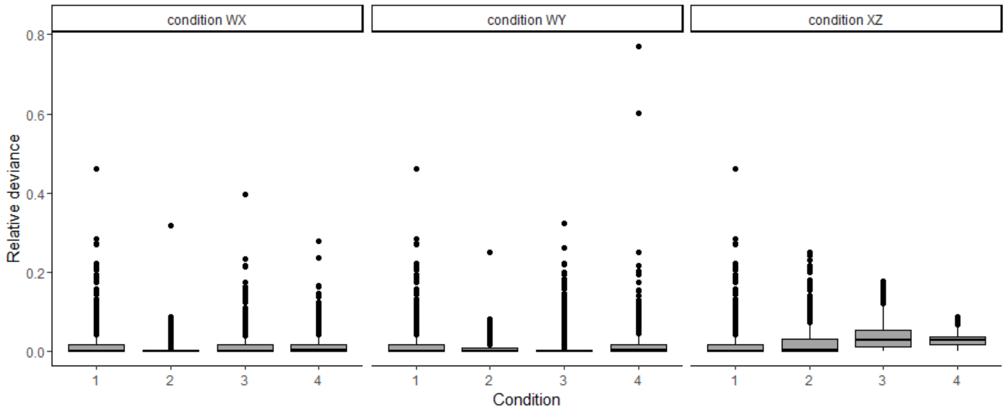

Figure 1: Results in terms of relative deviance (see text). Boxplots demonstrate the spread of the 1000 relative deviance values obtained per simulation condition. The three panels represent the three sets of bivariate relationships specified in the study, and the four columns per panel illustrate the four alternative strengths of those relationships specified (see Table 2).

Figure 1 illustrates the different values obtained for relative deviance under the different simulation conditions. By relative deviance we mean the deviance after applying the optimization procedure as a proportion of the deviance before applying the procedure, so a value close to zero means that a new audit sample is drawn that substantially improves the representativity of the audit. In the left panel, the results are shown for the four different WX relationships under which the initial sample was drawn. As representativity is defined with respect to Y, not with respect to W, differences in relative deviance between these WX conditions are not expected and this is confirmed here. In the middle panel, the results are shown for the four different WY relationships under which the initial sample was drawn. The (lack of) differences between these boxplots illustrates that the proposed method is able to perform in situations where the relationship between domain variables and the variable of interest in the audit sample is strong and possibly also unbalanced. In the right panel, results are shown for the four different XZ relationships under which the initial audit sample was drawn. Here, a small increase in average relative deviance can be detected when inclusion in the initial audit sample relates more strongly to scores on X, while the spread of the results becomes drastically smaller. This is likely to be caused by the fact that the deviance of the initial model was more substantive.

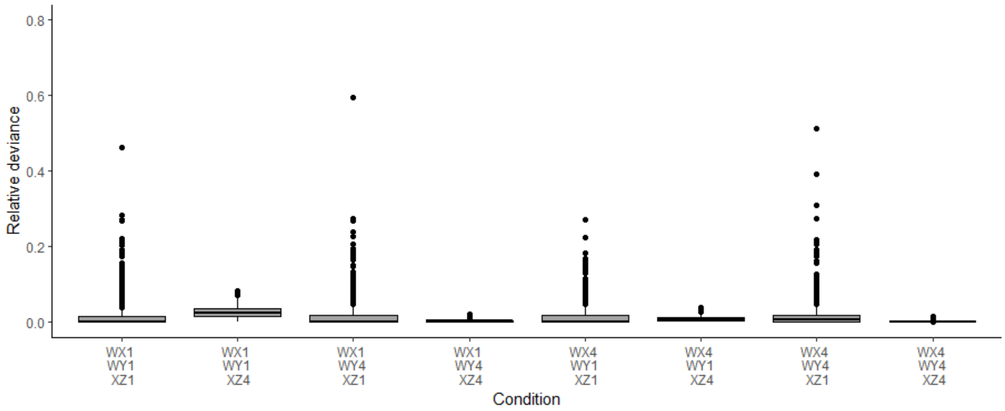

Figure 2: Results in terms of relative deviance (see text). Boxplots demonstrate the spread of the 1000 relative deviance values obtained per simulation condition for eight different combinations of different strengths of relationships between the variables (WX), (WY) and (XZ) (see Table 2).

Figure 2 illustrates the different values obtained for relative deviance under interactions between the different simulation conditions. As indicated by the boxplot labels, for each boxplot, the starting dataset is generated under a different combination of conditions, for which the complete overview can be found in Table 2. These boxplots illustrate that the scores for relative deviance are stable over different data-generating conditions. It is particularly noteworthy that a strong relationship for (XZ) apparently has the most substantial influence on scores for relative deviance.

4.3.2 Bias in W

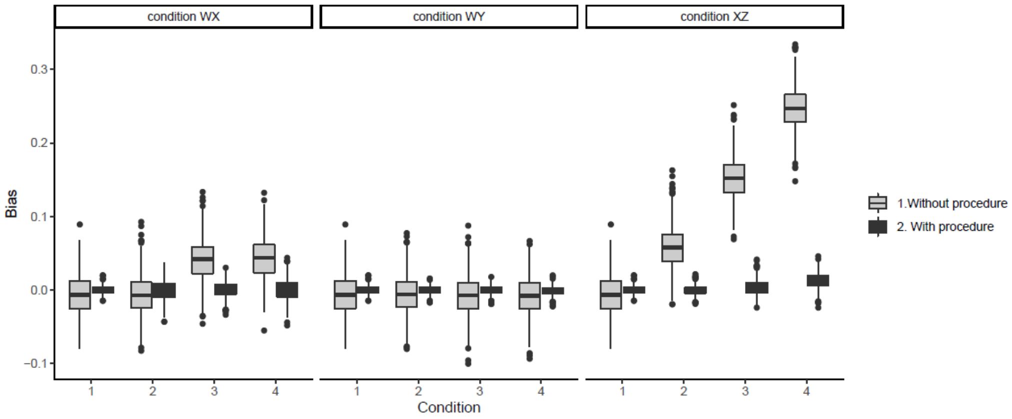

Figure 3: Boxplots of the bias distribution of the estimated proportions of W, based on 1000 replicates per simulation condition. The three panels represent the three sets of bivariate relationships specified in the study, and the four columns per panel illustrate the four strengths of those relationships (see Table 2).

Figure 3 illustrates the different values obtained for bias under the different simulation conditions compared to when the initial audit sample would have been used directly. Only the results for $W = 1$ are shown as the results for $W = 2$ and $W = 3$ behaved very similarly. In the left panel, the results are shown for the four different WX relationships under which the initial sample was drawn. In the first condition $W = X$. As X is the observed target variable of interest and W its true version, it is expected that no bias is present in results for W if $W = X$. In conditions 2-4, $W \neq X$ and small increases in spread can be detected as a consequence. However, it can also be seen that if the procedure is not applied in such cases, bias is induced in the results for W. In the middle panel, the results are shown for the four different WY conditions. Similar to the results found for the deviance, it can be concluded that the relationship between the variable of interest and domain variables does not affect the bias in W. However, bias is also not detected before the procedure was applied, so a correction procedure is not required in such situations. In the right panel, results are shown for the four different XZ conditions. Here, a small increase in bias can be detected when inclusion in the initial audit sample relates more strongly to scores on X, i.e. the procedure finds it more difficult to obtain a sample that is unbiased with respect to W if inclusion in the initial sample becomes more unbalanced with respect to X. However, not applying the procedure in such situation results in very substantive amounts of bias in W.

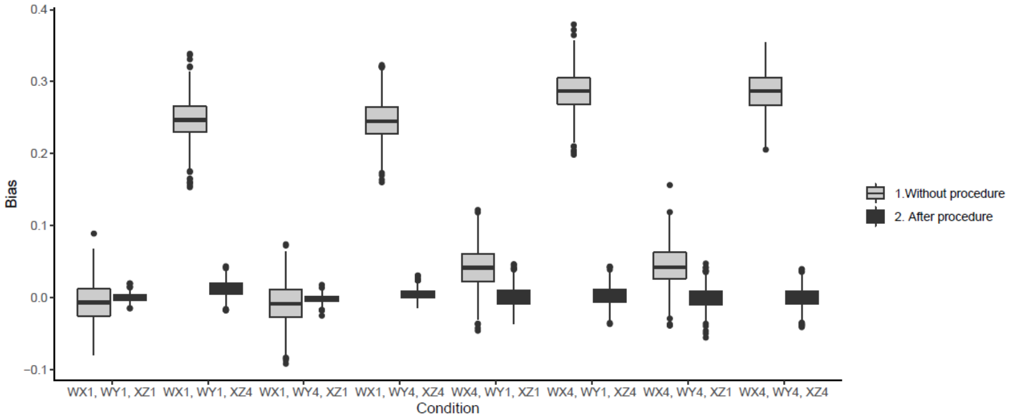

Figure 4: Boxplots of the distribution of the bias of the estimated proportions of W, based on 1000 replicates per simulation condition for eight combinations of different strengths of relationships between the variables (WX), (WY) and (XZ) (see Table 2)

Figure 4 illustrates the different values of bias in W after applying the method under interactions between the different simulation conditions. As indicated by the boxplot labels, for each boxplot, the data is generated under a different combination of conditions. These boxplots illustrate that although the bias present in estimates of W can be caused by both a weaker relationship in (WX) or a stronger relationship in (XZ), the combination of these conditions do not result in a multiplication in the effect in terms of bias after applying the procedure. Furthermore, deviations in (WX) result in a wider spread while deviations from CI between Z and X result in more systematic bias. In addition, these boxplots again illustrate that in situations of an imbalance in XZ, applying a procedure to obtain a representative audit sample is essential.

4.3.3 Bias in XW

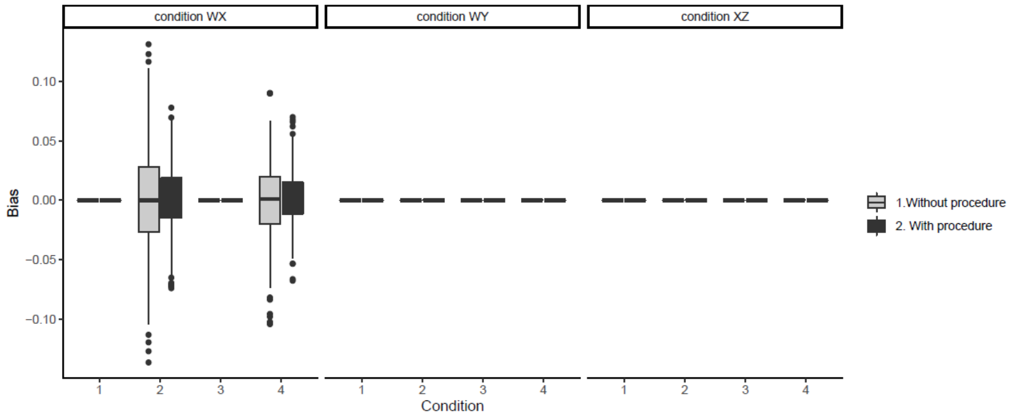

Figure 5: Boxplots of the distribution of the bias of the estimated proportions of X conditional on W, based on 1000 replicates per simulation condition. The three panels represent the three sets of bivariate relationships specified in the study, and the four columns per panel illustrate the four alternative strengths of those relationships specified (see Table 2).

Figure 5 illustrates the different values obtained for bias in $X|W$. Only the results for $(X = 1|W = 1)$ are shown as the results for other proportions were very similar. In the left panel, the results are shown for the four different WX relationships. Similarly as when evaluating the bias for W, X is the observed target variable of interest and W its true version. It is therefore expected that no bias is present in results for $X|W$ if $W = X$. In conditions 2 and 4, $W \neq X$ and small increases in bias can be detected as a consequence, while in condition 3 there is also no bias present due to the fact that here $W = X$ for $X = 1$. Similarly, no bias is present in $X|W$ under the different simulation conditions for WY and XZ, as under these conditions $W = X$ as well. In cases of observed bias, the spread of the bias is smaller when the procedure is applied compared to when the procedure is not applied.

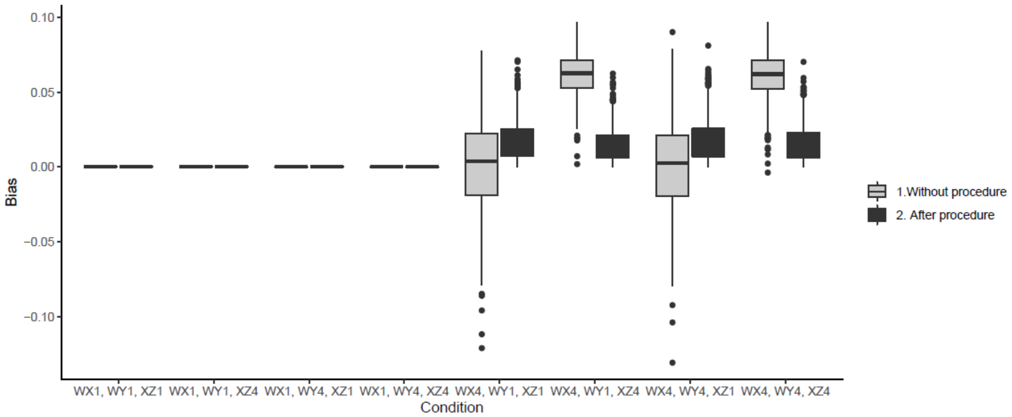

Figure 6: Boxplots of the distribution of the bias of the estimated proportions of X conditional on W before and after applying the procedure, based on 1000 replicates per simulation condition for eight combinations of different strengths of relationships between the variables (WX), (WY) and (XZ) (see Table 2).

In Figure 6 it is illustrated that if $W \neq X$, generally more bias is present, although the amount of bias remains limited if the audit sampling procedure is applied. Furthermore, it is shown that combinations of measurement error in W $(W \neq X)$, different strengths of WY and selectivity in the initial audit sample, XZ, do not cause for more substantive bias in $X|W$ than already caused by WX. Again, in such cases the bias is more substantive if the audit sample procedure has not been applied.

4.3.4 Variance estimation

Table 3: Estimated true standard deviation (true sd) and se-sd ratio based on 1000 simulations for the estimated proportions of W. Rows refer to different simulation conditions, columns to different categories of W. (* = no optimization.)

	True SD			SE-SD ratio		
Condition	w=1	w=2	w=3	w=1	w=2	w=3
WX4,WY1,XZ4	.012	.013	.013	1.24	1.23	1.18
WX4,WY2,XZ4	.012	.014	.014	1.49	1.36	1.34
WX4,WY3,XZ4	.011	.014	.014	1.26	1.14	1.19
WX4,WY4,XZ4	.012	.014	.014	1.41	1.36	1.34
WX4,WY1,XZ1*	.019	.020	.019	1.03	0.99	1.03

Table 4: Results in terms of se-sd ratio based on 1000 simulations for the estimated proportions of X conditional on W. Rows refer to different simulation conditions, columns to different combinations of categories of W and X. (* = no optimization.)

	w=1			w=2			w=3		
Condition	x=1	x=2	x=3	x=1	x=2	x=3	x=1	x=2	x=3
WX4,WY1,XZ4	1.18	1.12	1.17	1.26	1.29	1.33	1.31	1.43	1.46
WX4,WY2,XZ4	1.12	1.08	1.06	1.14	1.20	1.20	1.25	1.21	1.37
WX4,WY3,XZ4	1.19	1.19	1.16	1.20	1.28	1.34	1.22	1.30	1.35
WX4,WY4,XZ4	1.12	1.14	1.11	1.19	1.22	1.15	1.17	1.18	1.23
WX4,WY1,XZ1*	0.99	0.98	0.95	1.02	1.02	0.97	1.00	1.01	1.02

Table 3 and Table 4 summarize the results of the separate simulation study mentioned in Section 4.2, to evaluate the performance of the variance estimators defined in Equation (25) and Equation (26) for a fixed target population. Each se-sd ratio in these tables represents the ratio of the average standard error (se) of the 1000 simulations and the empirical standard deviation (sd) across 1000 simulations, for a particular condition and target parameter P_w^W (Table 3) or $P_{x|w}^{X|W}$ (Table 4). Ideally, this se-sd ratio would be equal to 1.

For the estimated P_w^W the empirical sd is also shown in Table 3. It is seen that the sd values were similar for all four conditions where the initial audit sample was not representative with respect to domain variables. For the estimated $P_{x|w}^{X|W}$ the empirical sd values are omitted here to save space; again, these values were similar across all four conditions.

It is seen that for all conditions where the initial audit sample was not representative with respect to domain variables, the variance estimators as defined in Equation (25) and Equation (26) tended to overestimate the true variance; the overestimation in terms of standard error varied between 6% and 49%, which in practice could be considered a moderate bias. As a benchmark, we also included a condition where (XZ) is given by the first matrix in Table 2, so that the initial audit sample already satisfied the CI model. In this case we did not apply the optimization procedure to alter the initial sample. Here it is seen that both variance estimators performed much better (last line in Tables 3 and 4). Thus, for a truly random sample from a population that satisfies the CI model, our variance estimators are approximately correct, and the overestimation of the variance for the other conditions is due to the optimization procedure. Apparently, by minimizing the deviance, the distribution of possible audit samples obtained by this procedure is more restricted than a stratified simple random sampling design, and this is not reflected by Equation (25) and Equation (26).

Within the conditions where the audit sample was not representative with respect to W, it is seen that the degree of overestimation in both tables differs between, on the one hand, the first and third condition and, on the other hand, the second and fourth condition of (WY). From Table 2, it is seen that this distinction corresponds to associations between W and Y that are relatively strong and relatively weak, respectively. For the estimated proportions $\hat{P}_w^W$, the overestimation by variance estimator (25) in Table 3 appears to be smaller when the association between W and Y is stronger. Surprisingly, for the estimated error probabilities $\hat{P}_{x|w}^{X|W}$ and variance estimator (26), the opposite effect is seen in Table 4.

5. Application

We used the proposed framework to select an audit sample that can be used for the production of statistics on energy consumption. The combined statistical register that we used for this application consists of 2 037 088 units (as illustrated in Table 1) in the Netherlands in 2019. 174 214 of these units were either previously audited manually, or the NACE codes were equal over three or more separate registers and therefore considered to be correct. (Note that the vast majority of audited units belonged to the second group; only a small minority was audited manually.) However, these audited units were not distributed in a representative manner with respect to the domain variable X and covariate Y, where domain variable X consists of 21 economic sectors (listed with the first digit NACE codes) and covariate Y consists of six categories:

- SG: Small gas consumption
- MG: Middle gas consumption
- LG: Large gas consumption
- LV: Low voltage
- HV: High voltage
- OPC: other profiles

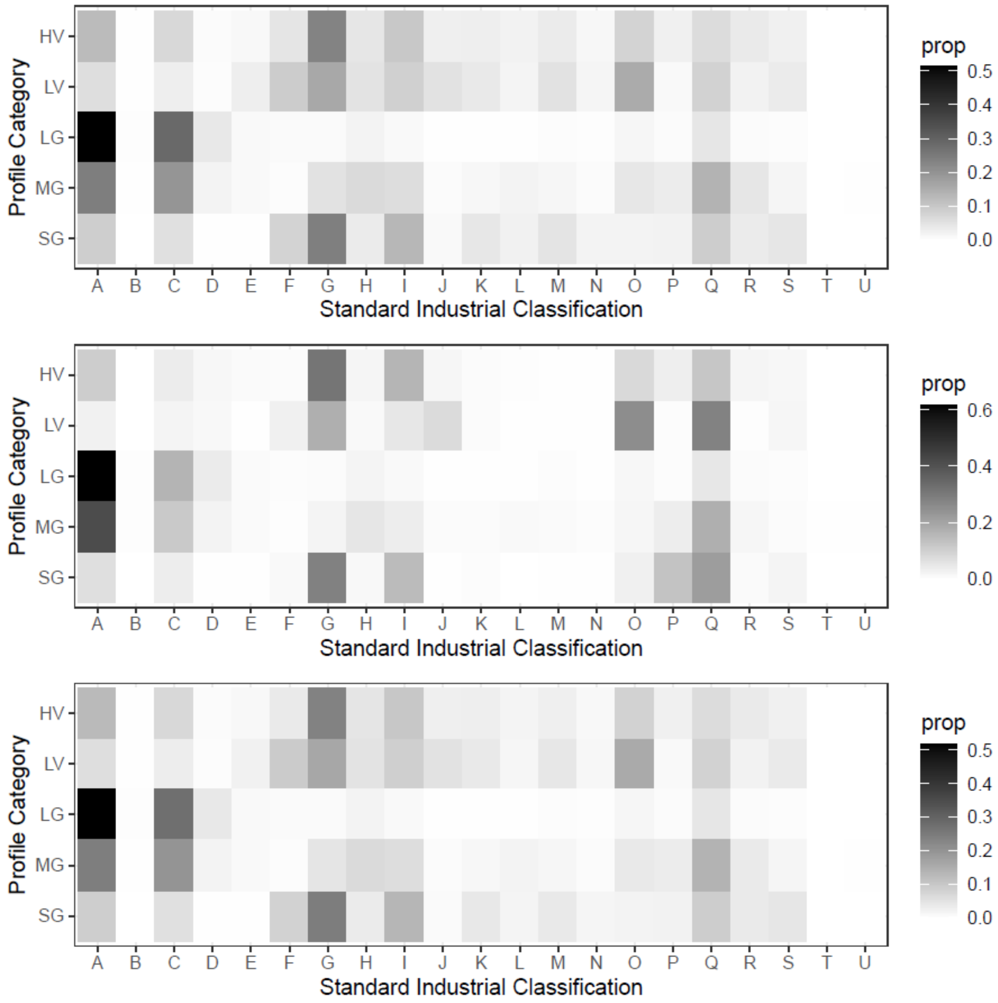

Figure 7: Heatmap of the proportion of establishments per economic sector for different profile categories: for the combined administrative data set (upper panel), for the original audit sample (middle panel) and for the audit sample adjusted with the proposed procedure (lower panel). Note that we excluded the category 'OPC' as it highly affected the interpretability of the heatmap.

The upper heat map of Figure 7 illustrates the observed proportions of economic sectors X per profile category Y. The purpose of the intended audit is to investigate the quality of these economic sector codes. In this heatmap we for example see that in the category of large gas consumers, most units are in the agricultural sector ("A") followed by manufacturing ("C"). In the category of small gas consumers, most units are in the sector wholesale and retail ("G").

Of all units shown in the upper heat map of Figure 7, a subset has previously been audited, and the final economic sector codes for these units are known and can be found in the middle heat map of Figure 7. By comparing the two heat maps, it can be clearly seen that given certain profiles certain sectors have been more thoroughly audited than others. For example, units in the economic sector "human health and social work activities" ("Q") have been audited relatively thoroughly, while audits for sectors such as water and waste ("E"), construction ("F"), finance ("K"), real estate ("L"), science ("M") and administration ("N") are relatively scarce.

To obtain an audit sample that is more representative in terms of the economic sectors and with respect to the covariate, we applied the methodology introduced in this manuscript. We first applied the methodology discussed in Section 3. Here, we applied the method multiple times using audit sample sizes 500, 1000, 1500 and 2000 (M_+), and for each sample size we varied the number of cases to exclude from the initial sample (M_-). The size of M_- was specified as a factor with respect to M_+, and we used the values of 50, 100 and 120 for this factor, as there were a lot of units already audited. Here, we investigated the resulting deviance values and concluded that we would probably be able to obtain a representative audit sample of a size between 1000 and 1500 and we concluded that we would need to increase the factor for removing cases. Therefore, we started with an audit sample of 1500 and were able to reduce the sample size to 1200. We tried to reduce the sample size even further to 1100 but this did not result in a representative sample by using our deviance criterion.

Now that we knew that we were able to obtain a representative sample by sampling approximately 1200 new cases, we applied the adjusted procedure for the objective function F_2 as described in Section 3.3 to ensure that all additionally sampled cases were indeed not in the initial audit. To assess the deviance value, as suggested in Section 3.2, we compared it to a chi-square distribution with $J \times (I-1)$ degrees of freedom, where I and J are the number of categories of X and Y, respectively, Here, with $I = 21$ and $J = 6$, we used a chi-square distribution with 120 degrees of freedom. Choosing $\alpha = .05$, we found a value of 147 as an approximate cut-off point. Thus, if we could draw an audit sample with a deviance lower than 147, we would find this acceptable in terms of representativity.

Additionally, we tried to reduce the sample size even further, but this was not successful. Finally, we selected a sample of 1200 and we removed 144000 of the units from the initial audit sample (multiplication factor of 120), with a deviance of 117.27. It should be noted that a cut-off point based on the chi-square distribution is not ideal as a criterion for accepting or rejecting an audit sample. It is known that, for a given lack of fit of the CI model in terms of $\Pr(X_g = x, Y_g = y, Z_g = z)$, the expected value of the deviance increases with the number of observations in the data (see, e.g., Agresti, 2013, Section 16.3.5). Hence for large populations, the deviance is sensitive to minor deviations from the CI model and a cut-off value of 147 might be unnecessarily restrictive.

The last heat map in Figure 7 (lower panel) illustrates the selected audit sample. This audit sample is a combination of units from the initial audit (middle heat map), and additionally selected units for audit (1200 units). In this heat map, it can be seen that certain economic sectors that were underrepresented in the initial audit are now more prominently present, such as water and waste, construction, finance and administration (“F”, “J”, “K” and “N”). Economic sectors real estate and science (“L” and “M”) are still underrepresented. However, overall it can be concluded that the distribution of the final audit sample is more similar to the distribution in the combined administrative data set compared to the initial audit sample, while the final audit comprises a smaller selection of units (31 414).

6. Conclusion

In this paper, we introduced a method that can be used to efficiently adjust an audit sample to make it more representative for a target population with respect to domain variables, when the initial audit sample was not. More specifically, our method uses the joint distribution between the domain variable of interest, covariates and an indicator for initial audit inclusion. Furthermore, the method assumes that there is no direct association between the true domain variable of interest and the initial audit inclusion pattern, conditional on the observed variable of interest and the covariates. Here, we can analyze the previously mentioned joint distribution using a conditional independence model with respect to the domain variables. The fit in terms of deviance of this model can be compared to that of the saturated model. Our method then numerically searches a new solution that minimizes the deviance by including new cases and excluding already audited cases.

The simulation study illustrated that particularly when there is measurement error in the observed domain variable, biased estimates of target parameters by domain can occur if the audit sample selection procedure is not applied to obtain a more representative audit sample. In addition, the study illustrated that without applying the procedure, biased estimates of target parameters per domain are obtained in situations where inclusion in the initial audit is related to the scores of the domain variable of interest. This conclusion is particularly relevant, as there is in practice often a relation between the scores on the target variable and initial audit inclusion. With financial audits, the largest companies are for example often audited by default. Furthermore, we conclude from our simulation study that the variance estimator based on stratified simple random sampling is often overestimating the true variance, due to the fact that by minimizing the deviance, the distribution of possible audit samples is more restricted than the distribution of stratified simple random samples would be.

The reason why we developed this procedure was that we wanted to perform an audit on establishments to investigate whether classification errors occurred in economic sector. This domain variable is used to publish statistics on energy consumption stratified by economic sector. In this application a non-probability audit sample was already available, but this sample was not representative in the sense that not all economic sectors had the same inclusion probability. Therefore we developed, investigated and applied the method introduced in this paper. By applying this method, we were able to select an audit sample that was more representative with respect to the economic sectors, that utilized as many of the cases that were already audited as possible and we had control over the size of the new audit sample. As a result, we are now able to estimate the quality of the statistical output created using the domain variable economic sector. In addition, we are now able to estimate the proportion of misclassifications present in the various administrative sources that measure economic sector.

We expect that the method is also interesting for other applications. For instance, assume that we have trained and tested a supervised machine learning model to derive a new variable. For instance, Tollenaar et al. (2018) describes that they were interested to predict whether potential crimes in police records concern cybercrime or not. Unfortunately, the annotated set that was used to train and test the model was not representative with respect to domain variables, because keywords were used to search for cases likely to be cybercrime rather than taking a randomized sample. Consequently, the model prediction error in the test set might not be representative for the true prediction error in the population. The police records also contain covariates (Y) that relate to the target variable, such as crime type and whether the case has been declared by a victim or not. One could then use the predicted cybercrime in all police records by the originally trained machine learning model as a proxy (X) for the true cyber variable (W) which is available only in the set that was not representative with respect to domain variables. Next, one could apply the procedure to select additional units to generate a test set that is more representative for the target population. The additionally selected units could then be manually annotated to obtain W. Finally, one could use this adjusted test set to obtain a more reliable estimate of the prediction error of the model.

In the current paper we illustrated that the method can be used in different ways. For example, an audit sample can be selected that truly has the lowest deviance, or a trade-off can be made between obtaining an acceptable value for the deviance and maximizing the amount of cases from the initial audit to be re-used. When applying the method in practice, the auditor should be aware of the practical implications when selecting a certain amount of cases to leave out or additional cases to include, particularly because including more cases in the audit will improve the probability to meet the deviance criterion on the one hand, but are also combined with increasing costs to perform the audit in practice on the other hand.

Further research should give more insight in how the method performs when different implementations of the method are chosen. In our application, we compared the deviance to a cut-off point from a chi-square distribution as a simple criterion to decide whether an audit sample was sufficiently representative, but it may be useful to develop more refined criteria that, for instance, also take the number of available observations into account. In addition, more practical applications should provide insight into what further improvements are interesting for researchers. For example, are users of the method typically interested in selecting a certain audit sample of a maximum size that is within their budget? Or is this often more flexible and are users more interested in an optimal combination of smallest sample size and appropriate representativity? Finally, the results in terms of variance invite for a more thorough investigation if the audit samples drawn using this method will also be used to draw conclusions in terms of variance of estimates.

Besides further investigating the method as proposed, it would also be interesting to see if and how the method can be adjusted to handle other challenging audit situations. A first example of such a situation is when the aim is not to match the distribution of certain domain variables, but to oversample certain subgroups. A second example of such a situation is when adding and/or dropping of specific units is a requirement, meaning that the deltas are both positive.

While our proposed method can be applied in a variety of situations, for some applications other, more specialized methods may be preferable. In the special case where an audit sample is to be selected ‘from scratch’, with no previously audited units available, the method of Falorsi and Righi (2015) may provide a more direct solution to optimize the design of the audit sample. A further comparison of these two approaches may be interesting. Future research could also focus on comparing our proposed method to other available approaches for handling non-probability samples, such as pseudo-weighting and superpopulation modeling (Elliott & Valliant, 2017; Rao, 2021).

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors would like to thank Lotte Mensink for many interesting discussions on this topic and for checking the R code of the simulation study, and Jeroen Pannekoek for reviewing a previous version of this paper and providing several useful comments.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling*. Journal of Survey Statics and Methodology* 1, 90–143.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and applications*. MIT Press.

Chataway, J., Davies, N., Farmer, S., Howard, R., Thompson, E., & Ward, K. (2004). Herpes simplex encephalitis: An audit of the use of laboratory diagnostic tests. *Qjm*, 97(6), 325 330.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). John Wiley & Sons.

Derks, K., de Swart, J., Wagenmakers, E.-J., Wille, J., et al. (2019). *Jasp for audit:*

Bayesian tools for the auditing practice.

Elder, R. J., Akresh, A. D., Glover, S. M., Higgs, J. L., & Liljegren, J. (2013). Audit sampling research: A synthesis and implications for future research. *Auditing: A Journal of Practice & Theory*, 32(sp1), 99–129.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249–264.

Eurostat (2008). Ramon - reference and management of nomenclatures. https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LSTNOMDTL&StrNom=NACEREV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC

Falorsi, P. D., & Righi, P. (2015). Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. *Survey Methodology*, 41(1), 215–236.

Hernández, B., Parnell, A., & Pennington, S. R. (2014). Why have so few proteomic biomarkers “survived” validation? (Sample size and independent validation considerations). *Proteomics*, 14(13-14), 1587–1592.

Holt, D., & Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142(1), 33–46.

Klingwort, J., Burger, J., & Buelens, B. (2021). Inferring network traffic from sensors without a sampling design. *Working Paper* 2021-02, Statistics Netherlands, The Hague.

R Core Team (2023). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. https://www.R-project.org/

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya: The Indian Journal of Statistics*, 83-B(1), 242–272.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

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

Scholtus, S., Bakker, B. F. M., & van Delden, A. (2015). Modelling measurement error to estimate bias in administrative and survey variables. *Discussion Paper* 2015-17, Statistics Netherlands, The Hague

Sobel, M. E., & Arminger, G. (1986). Platonic and operational true scores in covariance structure analysis: An invited comment on Bielby’s “Arbitrary metrics in multiple indicator models of latent variables”. *Sociological Methods & Research*, 15(1-2), 44–58.

TheWorldBank (2019). *Listed domestic companies.* https://data.worldbank.org/indicator/CM.MKT.LDOM.NO

Tollenaar, N., Rokven, J., Macro, D., Beerthuizen, M., & van der Laan, A. M. (2018). Predictieve textmining in politieregistraties: cyber- en gedigitaliseerde criminaliteit (tech. rep.). *Rapport van het Wetenschappelijk Onderzoek- en Documentatiecentrum, Ministerie van Justitie en Veiligheid*.

United Nations (2015). Resolution adopted by the General Assembly on 25 September 2015. *Transforming our world: The 2030 agenda for sustainable development.* https://www.refworld.org/docid/57b6e3e44.html

Zadrozny, B. (2004, July). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning* (p. 114).

Appendix

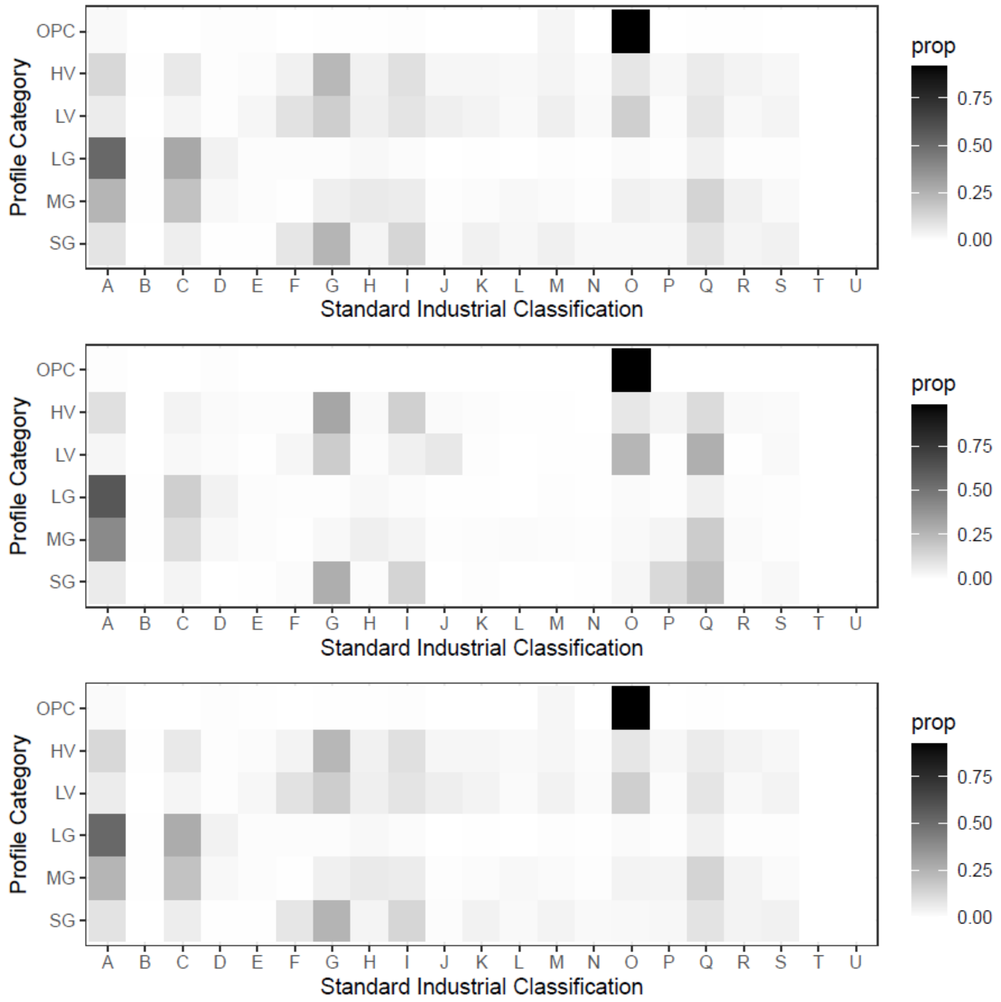

Figure 8: Alternative version of the heatmap of the proportion of establishments per economic sector for different profile categories: for the combined administrative data set (upper panel), for the original audit sample (middle panel) and for the audit sample adjusted with the proposed procedure (lower panel). In this alternative version we included the category 'OPC'. Here it can be seen that this highly affects the interpretability of the heatmap.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: http://www.cbs.nl/information