Journal of Official Statistics, Vol. 22, No. 3, 2006, pp. 565–584

Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey

Paul Knottnerus and Coen van Duin[1]

In the past few years, Statistics Netherlands has been implementing the repeated weighting estimator in its regular estimation process. This estimator ensures numerical consistency among tables estimated from different surveys. Especially when the tables have some variables in common, this approach appears to be very useful. After a concise summary of the repeated weighting procedure, this article gives the variance formulas for the repeated weighting estimator. It concludes with an example from the Dutch Labour Force Survey. The variance estimator for this example is discussed and the results of a simulation study testing the accuracy of this estimator are presented.

Key words: Combining registers and surveys; numerically consistent tables; recalibration; regression estimators; superresiduals; variance estimation.

1. Introduction

In classical survey estimation, each survey is carried out and processed independently of the other surveys and, consequently, the set of weights is held constant per survey. Such a unique set of weights for each survey makes it easy to compose various tables from the same survey. However, a serious drawback of the classical approach is that multidimensional tables from two or more surveys which have a variable in common may have different numerical values for the same variable, i.e., the tables need not be numerically consistent. In the last few years, Statistics Netherlands has been implementing an alternative estimation strategy, called repeated weighting (RW). The RW estimation strategy accommodates user demands to produce outputs that are numerically consistent. The underlying methodology is based on the seminal paper by Kroese and Renssen (1999). In a recent article by Houbiers (2004), considerations with respect to the RW estimation procedure plus its applicability in practice are described extensively. She also describes the social-statistical database (SSD) in which data from various surveys and registers are combined. For further details, see Houbiers (2004), Boonstra et al. (2003) and the references therein.

[1] Statistics Netherlands, Post Box 4000, 2270 JM Voorburg, The Netherlands. Emails: pkts@cbs.nl and cdin@cbs.nl, respectively. The views expressed in the article are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

Acknowledgments: The authors are grateful to Marianne Houbiers, Robbert Renssen, Harm-Jan Boonstra, Jan van den Brakel, and Rob van de Laar for valuable comments on earlier drafts. They also thank the Associate Editor and three anonymous referees for their helpful comments. This work was partially supported by the Eurostat project DACSEIS.

In essence, the RW estimation procedure amounts to an additional calibration step to adjust the standard regression weights. For some tables such a step might be necessary when margins of a table to be estimated are already estimated from a larger survey or the margins are known from a register. For more results on calibration estimators, see Deville and Särndal (1992). This article presents a derivation of the variance formulas for the RW estimator, as well as the results of a simulation study testing these formulas in practice.

The outline of the article is as follows. Section 2 gives a concise summary of the RW estimation procedure for a set of frequency tables and introduces some notation. In Section 3 we present a method for estimating the variance of the RW estimator. This method is based on the variance tree of so-called superresiduals described in Knottnerus (2001) and Boonstra et al. (2003). Section 4 gives a practical example with real data plus a number of useful recursions for estimating the variance. Section 5 describes the results of a simulation study. The simulations are carried out with data from the Dutch Labour Force Survey in order to get an insight into the performance of the variance estimators described in Section 3.

2. The Repeated Weighting (RW) Procedure

The main aim of the RW estimation procedure is to obtain a set of numerically consistent tables in cases where the tables are estimated from different sources. These sources may be either surveys or registers. Regarding a given reference period, a set of target tables is specified. Throughout this article we make four assumptions:

1. the reference period of the registers and the surveys is the same;
2. the registers and surveys refer to the same population;
3. variables with the same name have the same definition for all relevant registers and surveys;
4. the categorical variables have hierarchical classifications, i.e., each class of a more detailed classification is nested within one class of a less detailed classification.

If the first three assumptions are not fulfilled, the RW estimates are not meaningful, since we will have imposed numerical consistency on quantities that need not be numerically consistent. The last assumption is required for the RW procedure to work. The variables referred to in these assumptions are those appearing in the set of target tables that one wants to estimate. Therefore it is necessary to specify this set before one can check whether the requirements are met.

In this article we focus on RW estimates using the so-called *splitting up* procedure. This procedure is a practical way of dealing with the order problem, i.e., the problem that the estimation results depend on the ordering of the tables to be estimated (for further details, see Boonstra et al. 2003). For simplicity, we restrict ourselves here to the case of (multidimensional) frequency tables.

The proposed RW estimation procedure consists of three steps. First, the set of target tables is specified and ordered. Second, the tables are estimated by means of the regression estimator. Third, a reweighting step is performed in which the table estimates are consecutively adjusted in such a way that numerical consistency between the estimates is obtained. We discuss each of these steps in more detail.

2.1. Step 1. Specifying and Ordering the Tables

First, the set of target tables to be estimated is specified. Next, all margins of a target table are added to the set of tables to be estimated. A marginal table is obtained by (i) aggregating over one or more categorical variables of a multi-way table or (ii) using a less detailed classification of a categorical variable. For example, the two-way table $A \times B$ is a margin of the table $A \times B \times C$ and also of the table $A \times B^{(2)}$, where $B^{(2)}$ has a more detailed classification than B. The tables are then ordered in such a way that a margin of a multidimensional table always precedes that table.

2.2. Step 2. Regression Estimation of the Tables

In this step, each table is estimated by means of the regression estimator from the most appropriate data set. In general, this will be the largest survey or a combination of surveys, also called a block and denoted by S. Note that when all variables of a table are available from registers, the table can simply be counted from the register data. Before we introduce the regression estimator for a frequency table, we introduce some notation. Let x_i denote the vector of J auxiliary variables for the ith element in a regression from block S. Let t_x denote the corresponding vector of population totals of the auxiliary variables. Furthermore, let t_Y denote a *vectorized* multi-way frequency table of the (multiple) categorical variable Y with P mutually exclusive categories or cells. To indicate the actual categorical variables in the table in practice, Y will often be of the form $A \times B$ or $A \times B \times C$, etc; for the algebra of vectorizing multi-way tables, see Knottnerus (2003, pp. 371–372). The underlying vector y_i of the vectorized frequency table t_Y can be seen as a P-vector of zeroes except one unity value, indicating the appropriate class of the ith element with respect to the (multiple) categorical variable Y; note that $t_Y \equiv t_y \equiv \sum_{i \in U} y_i$, where U stands for the population. The regression estimator of a vectorized multi-way frequency table t_Y from block S can now be written in standard matrix notation as

$$\hat{t}_Y^{REG(S)} = \hat{t}_Y^{HT(S)} + \hat{B}'_{d,x}\left(t_x - \hat{t}_x^{HT(S)}\right)$$

$$= \sum_{i \in S} d_i^{(S)} y_i + \hat{B}'_{d,x}\left(t_x - \sum_{i \in S} d_i^{(S)} x_i\right)$$

$$\hat{B}_{d,x} = \left(\sum_{i \in S} d_i^{(S)} x_i x_i'\right)^{-1} \sum_{i \in S} d_i^{(S)} x_i y_i'$$

where $\hat{t}_Y^{HT(S)}$ and $\hat{t}_x^{HT(S)}$ are Horvitz-Thompson (HT) estimators from block S, and y_i', x_i' and B' are the transposes of y_i, x_i and B, respectively. The $d_i^{(S)}$ stand for the design weights in block S. That is, for a single sample we have straightforwardly $d_i^{(S)} = 1/\pi_{Si}$, where π_{Si} is the first-order inclusion probability for that sample. For a block consisting of, for instance, the union of two samples S_1 and S_2 we may choose

$$\hat{t}_Y^{HT(S)} \equiv \lambda_1 \hat{t}_Y^{HT(S_1)} + (1 - \lambda_1)\hat{t}_Y^{HT(S_2)} \equiv \sum_{i \in S} d_i^{(S)} y_i \qquad (1)$$

where λ_1 reflects the relative weight or reliability of S_1 in block S. Consequently,

$$d_i^{(S)} = \begin{cases} \lambda_1/\pi_{1i} & \text{if } i \in S_1 \text{ and } i \notin S_2 \\ (1-\lambda_1)/\pi_{2i} & \text{if } i \in S_2 \text{ and } i \notin S_1 \\ \lambda_1/\pi_{1i} + (1-\lambda_1)/\pi_{2i} & \text{if } i \in S_1 \cap S_2 \end{cases}$$

A simple manner for choosing an appropriate value for λ_1 is to set $\lambda_1 = n_1/(n_1+n_2)$ where n_1 and n_2 are the sample sizes of S_1 and S_2, respectively. In general, it can be shown that this choice of λ is optimal when S_1 and S_2 are two (independent) simple random samples with replacement, or when S_1 and S_2 are two mutually disjoint subsamples without replacement from a large SRS mother sample S without replacement; see Knottnerus (2003, p. 340). For examples illustrating that for unequal probability sampling the sampling error can be much more relevant than the sample size, see Section 5 and Houbiers et al. (2003).

In terms of weights the regression estimator can be written in the familiar form

$$\hat{t}_Y^{REG(S)} = \sum_{i \in S} w_i^{(S)} y_i$$

$$w_i^{(S)} = d_i^{(S)} \left\{ 1 + x_i' \left(\sum_{i \in S} d_i^{(S)} x_i x_i' \right)^{-1} \left(t_x - \hat{t}_x^{HT(S)} \right) \right\} = d_i^{(S)} g_i^{(S)} \tag{2}$$

A well-known property of the regression weights $w_i^{(S)}$ is that they satisfy the calibration equations

$$\sum_{i \in S} w_i^{(S)} x_i = t_x$$

Recall that the regression weights $w_i^{(S)}$ can be seen as the solution of the constrained minimization problem

$$\min_w \sum_{i \in S} d_i^{(S)} \left(\frac{w_i}{d_i^{(S)}} - 1 \right)^2 \quad \text{subject to} \quad \sum_{i \in S} w_i x_i = t_x$$

see Deville and Särndal (1992). In addition, throughout this article we assume that a constant term is included in the underlying regressions so that the residuals e_i from a regression on the whole population have a zero total, i.e., $t_e = 0$.

2.3. Step 3. The Reweighting Step

When for a certain table the regression weights $w_i^{(S)}$ lead to a margin that is numerically inconsistent with an estimate of that margin from a previously estimated table of the set, that table should be reweighted. Such an inconsistency will occur when, for instance, such a margin is observed in a block larger than S whereas that margin is not included in the vector x_i of auxiliaries or, in terms of calibration, the margin is not included in the calibration equations for the actual block S. By reweighting we mean an adjustment of

the regression weights $w_i^{(S)}$ for this specific table so that the margins of the reweighted table are in line with their estimates from a preceding table or their known counts from a register. Let m denote the vector variable of all linearly independent margins of the present, vectorized multi-way table t_Y. For a further discussion on margins when one of the variables in table t_Y has two or more hierarchical classifications, see Boonstra et al. (2003). Now the RW estimator of t_Y from block S is defined in a recursive way by

$$\begin{aligned} \hat{t}_Y^{RW} &= \hat{t}_Y^{REG(S)} + \hat{B}'_{w,m}\left(\hat{t}_m^{RW} - \hat{t}_m^{REG(S)}\right) \\ \hat{B}_{w,m} &= \left(\sum_{i\in S} w_i^{(S)} m_i m_i'\right)^{-1} \sum_{i\in S} w_i^{(S)} m_i y_i' \end{aligned} \tag{3}$$

For an example, see Section 4. The elements in $\hat{t}_m^{RW}$ are estimates from a preceding table or known counts from a register. Similarly to (2), we can write the RW estimator $\hat{t}_Y^{RW}$ in terms of weights. That is,

$$\hat{t}_Y^{RW} = \sum_{i\in S} r_i^{(Y)} y_i$$

$$r_i^{(Y)} = w_i^{(S)}\left\{1 + m_i'\left(\sum_{i\in S} w_i^{(S)} m_i m_i'\right)^{-1}\left(\hat{t}_m^{RW} - \hat{t}_m^{REG(S)}\right)\right\}$$

Hence, by construction the $r_i^{(Y)}$ satisfy the corresponding consistency requirements

$$\sum_{i\in S} r_i^{(Y)} m_i = \hat{t}_m^{RW}$$

This also holds true when the variables in m are linearly dependent, provided that the consistency constraints have a solution and the generalized inverse is used in the foregoing formulas; see Renssen and Martinus (2002).

In summary, repeated weighting can be seen as an additional calibration step for a new adjustment of the regression weights $w_i^{(S)}$ resulting in the final weights $r_i^{(Y)}$, which are consistent with the given margins of the present table. That is, these margins are already given by estimates from preceding tables or are already known from a register. Similarly to the second step, the $r_i^{(Y)}$ can be seen as the solution of the recalibration problem

$$\min_r \sum_{i\in S} w_i^{(S)}\left(\frac{r_i}{w_i^{(S)}} - 1\right)^2 \quad \text{subject to} \quad \sum_{i\in S} r_i m_i = \hat{t}_m^{RW}$$

Finally, it should be noted that the vector t_x in the regression estimator in Step 2 may include elements which are estimated for a preceding table from a larger block by either standard weighting or repeated weighting. However, because of its recursive definition the RW estimator always remains within the class of linear combinations of regression estimators from the underlying samples.

As an alternative to repeated weighting, one could simply include m_i in the auxiliary vector x_i to avoid numerical inconsistencies among the tables. Such an approach should work well when each sample size is sufficiently large and the number of variables in x_i and m_i is small. However, in the case of a large table the number of variables in x_i and m_i can be rather large and this approach will often become hard to apply because of a lack of observations in many cells. Repeated weighting reduces this problem by limiting the number of calibration equations involved in each particular table estimate.

When the samples are large enough, a fairly different approach is as follows. Assuming that there are q tables with common margins, define $t_Y = (t'_{Y1}, \ldots, t'_{Yq}, t'_x)'$. Let $Rt_Y = c$ be the set of linear consistency equations. This includes $t_x = t_{x0}$, where t_{x0} is the vector of known population totals. Let $\hat{t}_Y$ be the estimated vector resulting from the different surveys and let V_Y be its covariance matrix. Assuming normality, the optimal consistent solution according to least squares theory now becomes

$$\hat{t}_Y^{(cons)} = \hat{t}_Y + K(c - R\hat{t}_Y)$$

$$Cov(\hat{t}_Y^{(cons)}) = (I - KR)V_Y$$

$$K = V_Y R'(RV_Y R')^{-1}$$

For further details and examples, see Knottnerus (2003, Chapter 12) and the references given therein.

3. The Variance of the RW Estimator

In order to derive a formula for the variance of the RW estimator, we consider the situation with one register and two *independent* samples S_1 and S_2 without replacement of sizes n_1 and n_2, respectively. The first- and second-order inclusion probabilities are denoted by π_{ki} and π_{kij}, respectively ($k = 1, 2$). Let the vector x_i of auxiliaries be known from a register for all elements i in population U. Let the categorical variable Z be observed in S_1 and S_2, and let the categorical variable Y be observed in S_2. Suppose that for all initial estimates we use the regression estimator based on the auxiliaries in x, irrespective of the survey. It is obvious that, in general, the Z-margin from the estimated two-way table $\hat{t}_{Z\times Y}^{REG(S_2)}$ from S_2 is numerically inconsistent with the estimated table $\hat{t}_Z^{REG(S_{12})}$ from S_{12}, i.e., the union of S_1 and S_2. Therefore, according to the splitting up procedure the estimated table $\hat{t}_{Z\times Y}^{REG(S_2)}$ is to be reweighted with respect to its two margins. Applying (3), we get

$$\begin{aligned}\hat{t}_{Z\times Y}^{RW} &= \hat{t}_{Z\times Y}^{REG(S_2)} + \left(\hat{B}'_{w;Z}, \hat{B}'_{w;Y^-}\right)\begin{pmatrix} t_Z^{REG(S_{12})} - \hat{t}_Z^{REG(S_2)} \\ t_{Y^-}^{REG(S_2)} - t_{Y^-}^{REG(S_2)} \end{pmatrix} \\ &= \hat{t}_{Z\times Y}^{REG(S_2)} + \hat{B}'_{w;Z}\left(t_Z^{REG(S_{12})} - \hat{t}_Z^{REG(S_2)}\right) + 0\end{aligned} \tag{4}$$

where Y^- is obtained from Y by leaving out one of its categories; this removes a redundancy or, equivalently, a linear dependency among the margins in the underlying regression. Note that the matrix $\hat{B}_{w,m}$ [$= (\hat{B}'_{w;Z}, \hat{B}'_{w;Y^-})'$] stems from a weighted regression of the categorical variable $Z \times Y$ on the categorical variables Z and Y^- with weights $w_i^{(S_2)}$

or, more precisely, from the weighted regressions of the dichotomous variables in $y_i \otimes z_i$ on the dichotomous variables in z_i and y_i^-. Also note that $t_{Z\times Y} = \sum_{i\in U} y_i \otimes z_i$, where $\otimes$ is the Kronecker product and $t_{Z\times Y}$ is defined as the vector in which the columns of the two-way frequency table $Z \times Y$ are stacked one on top of the other. From (4) it can be seen that the RW estimator can be written as a linear combination of regression estimators from S_1 and S_2.

Before we derive the variance formulas for the RW estimator, it is recalled from standard sampling theory that a regression estimator for a frequency table from a random sample S of size n can be approximated by

$$\begin{aligned}\hat{t}_Y^{REG(S)} &= \hat{t}_Y^{HT(S)} + \hat{B}'_{d,x}\left(t_x - \hat{t}_x^{HT(S)}\right) \\ &= \hat{t}_Y^{HT(S)} + B'_x\left(t_x - \hat{t}_x^{HT(S)}\right) + (\hat{B}_{d,x} - B_x)'\left(t_x - \hat{t}_x^{HT(S)}\right) \qquad (5a) \\ &= \text{constant} + \hat{t}_e^{HT(S)} + O_p(N/n) \quad (\text{constant} = B'_x t_x = t_Y)\end{aligned}$$

$$e_i \equiv y_i - B'_x x_i$$

$$B_x = \left(\sum_{i\in U} x_i x'_i\right)^{-1} \sum_{i\in U} x_i y'_i \qquad (5b)$$

In (5b) we made the regularity assumption that all estimated population means and regression coefficients have variances of order $1/n$; see Knottnerus (2003, p. 119). Note that the so-called population residual e_i is a vector of ordinary variates; cf. the y_i. Subsequently, the covariance matrix of the regression estimator $\hat{t}_Y^{REG(S)}$ from (5a) can be approximated in the usual manner by

$$Cov\left(\hat{t}_Y^{REG(S)}\right) = \sum_{i,j\in U} (\pi_{ij} - \pi_i \pi_j) \frac{e_i e'_j}{\pi_i \pi_j}$$

Assuming that $1 << n << N$, we can borrow the variance estimator from the Hansen-Hurwitz (HH) estimator for estimating this variance; see Särndal et al. (1992, p. 422). That is, assuming that the sample size is much smaller than N, we may ignore the finite population correction. For the case at hand, this yields

$$\hat{Cov}\left(\hat{t}_Y^{REG(S)}\right) = \sum_{i\in S} \left(d_i^{(S)}\right)^2 \hat{e}_i \hat{e}'_i \qquad (6)$$

For a justification of this form, see (A3) in Appendix A. The estimated sample residuals $\hat{e}_i$ in (6) are based on the estimated regression matrices $\hat{B}_{d,x}$ from S.

Now it is rather straightforward to derive an approximation formula for the variance of an RW estimator. Similarly to (5b), we can approximate the RW estimator from (4) by

$$\hat{t}_{Z\times Y}^{RW} = t_{Z\times Y} + \hat{t}_{e(Z\times Y)}^{HT(S_2)} + B'_Z\left(\hat{t}_{e(Z)}^{HT(S_{12})} - \hat{t}_{e(Z)}^{HT(S_2)}\right) + O_p(N/n_2)$$

where $e_i(\cdot)$ is a vector of residuals for the ith element of population U from a regression of $(\cdot)$ on the variables in x. Decomposing $\hat{t}_{e(Z)}^{HT(S_{12})}$ into its two underlying components

according to (1) and neglecting higher-order terms, we get

$$\hat{t}_{Z\times Y}^{RW} = t_{Z\times Y} + \hat{t}_{e(Z\times Y)}^{HT(S_2)} + B_Z'\left\{\lambda_1 t_{e(Z)}^{HT(S_1)} + (1-\lambda_1)t_{e(Z)}^{HT(S_2)} - \hat{t}_{e(Z)}^{HT(S_2)}\right\}$$

$$\equiv t_{Z\times Y} + \hat{t}_{\varepsilon_1}^{HT(S_1)} + \hat{t}_{\varepsilon_2}^{HT(S_2)}$$

where the so-called superresiduals ε_{1i} and ε_{2i} for S_1 and S_2 are defined by

$$\varepsilon_{1i} = \lambda_1 B_Z' e_i(Z)$$

$$\varepsilon_{2i} = \lambda_2 B_Z' e_i(Z) + e_i(Z\times Y) - B_Z' e_i(Z) \quad (\lambda_2 = 1-\lambda_1)$$

respectively ($i = 1, \ldots, N$). These superresiduals play a crucial role for estimating the covariance matrix of an RW estimator. Using that S_1 and S_2 are independent, the covariance matrix of $\hat{t}_{Z\times Y}^{RW}$ can be approximated by

$$Cov\left(\hat{t}_{Z\times Y}^{RW}\right) = Cov\left(\hat{t}_{\varepsilon_1}^{HT(S_1)} + \hat{t}_{\varepsilon_2}^{HT(S_2)}\right) = \sum_{k=1}^{2}\sum_{i,j\in S_k}(\pi_{kij} - \pi_{ki}\pi_{kj})\frac{\varepsilon_{ki}\varepsilon_{ki}'}{\pi_{ki}\pi_{kj}}$$

Similarly to (6), this covariance can be estimated by

$$\hat{Cov}\left(\hat{t}_{Z\times Y}^{RW}\right) = \sum_{k=1}^{2}\sum_{i\in S_k}\left(d_i^{(S_k)}\right)^2 \hat{\varepsilon}_{ki}\hat{\varepsilon}'_{ki} \tag{7}$$

provided that $n_1, n_2 << N$. Note that the superresiduals ε_{ki} $(k = 1, 2)$ have zero totals as well. Furthermore, the estimated superresiduals from the samples are based on the estimated regression matrices $\hat{B}_{d,x}$ and $\hat{B}_{w;Z}$; note that these matrices depend on the actual block and the dependent variable of the underlying regression. For a discussion on the variance for RW estimators under two-phase sampling and a simple variance approximation of the RW estimator, which is useful in the context of one register in combination with one sample, see Boonstra et al. (2003). In the next section we will show how in more complicated situations the superresiduals can be calculated table by table in a recursive manner.

4. An Example from the Dutch Labour Force Survey

Houbiers (2004) describes the present state of the Social-Statistical Database (SSD) used by Statistics Netherlands. Furthermore, she explains how different surveys and registers can be combined in order to obtain reliable and consistent estimates from the SSD by means of repeated weighting. She also discusses a number of complications in the process of constructing the SSD and estimating consistent tables from it. In this section we will elaborate on her example of the Dutch Structure of Earnings Survey (SES), in which the Employment and Wages Survey (EWS) is combined with the Labour Force Survey (LFS); see Figure 1. In particular, we derive the variance estimators for the RW estimators, which were used to obtain numerically consistent estimates. A special feature of SES is that the population consists of *jobs*. The auxiliary variables *gender* (G) with 2 classes, *age* (A) with 5 classes, and *business class* (C) with 24 classes are known from a register for all jobs of

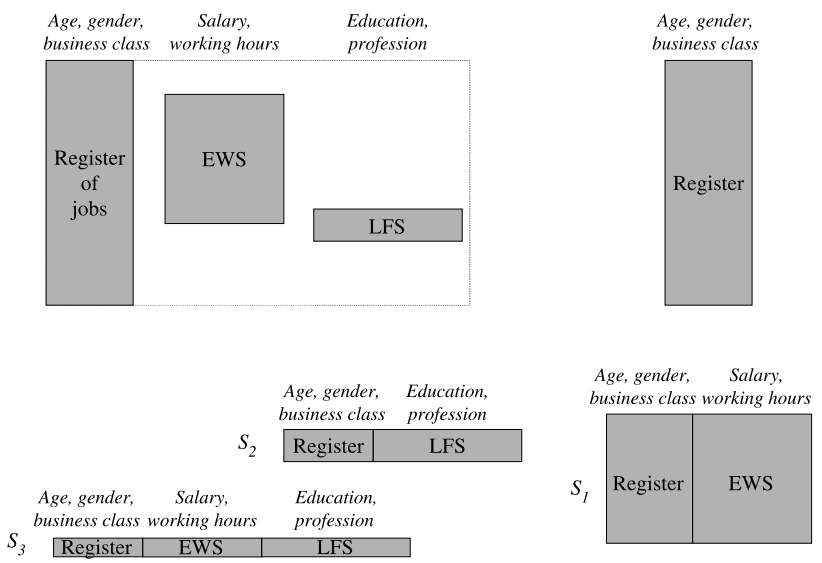

Fig. 1. Example of micro data from the SSD, and the construction of rectangular data blocks

the population. The study variable *working hour*s (H) with 2 classes is observed in the EWS (S_1) while the variable *education* or *learning* (L) with 7 classes is observed in the LFS (S_2).

Our target table is the three-way frequency table *gender* × *working hours* × *education* or, for short, $t_{G\times H\times L}$. This table with 28 ($= 2\times 2\times 7$) cells is to be estimated from the overlap of S_1 and S_2. S_1 is a sample of approximately 50% of the population of jobs while S_2 is an independent sample of only 1.56% of the population. The overall weighting scheme for S_1, S_2 as well as $S_3\,(\equiv S_1\cap S_2)$ consists of the dichotomous variables corresponding to the categorical variables *gender* × *age* and *business class* or, for short, $G\times A + C$. Taking into account the various number of classes, the vector x of linearly independent auxiliaries consists of $33\,(= 2\times 5 + 24 - 1$ dichotomous variables corresponding to the weighting scheme $(G\times A) + C$.

According to the splitting up procedure, we first estimate the one-way tables t_H and t_L from S_1 and S_2, respectively. Next, we estimate the two-way tables $t_{G\times H}$ from S_1, $t_{G\times L}$ from S_2, and $t_{H\times L}$ from the overlap of S_1 and S_2. Since G is included in the vector of auxiliaries, the estimated tables $t_{G\times H}^{REG(S_1)}$ and $t_{G\times L}^{REG(S_2)}$ are numerically consistent with the *gender* counts from the register. However, the table $t_{H\times L}^{REG(S_3)}$ from the overlap of S_1 and S_2 is to be reweighted since it is inconsistent with the estimated tables $\hat{t}_H^{REG(S_1)}$ and $\hat{t}_L^{REG(S_2)}$. Based on the $8\,(= 2 + 6)$ linearly independent margins of the reweighting scheme $H + L^-$, this yields

$$\hat{t}_{H\times L}^{RW} = \hat{t}_{H\times L}^{REG(S_3)} + \hat{B}'_{w;H}\left(\hat{t}_H^{RW} - \hat{t}_H^{REG(S_3)}\right) + \hat{B}'_{w;L^-}\left(\hat{t}_{L^-}^{RW} - \hat{t}_{L^-}^{REG(S_3)}\right)$$

$$\hat{t}_H^{RW} = \hat{t}_H^{REG(S_1)} \quad \text{and} \quad \hat{t}_{L^-}^{RW} = \hat{t}_{L^-}^{REG(S_2)}$$

Hence, neglecting the random character of the regression matrices $\hat{B}$, we get in terms of population residuals

$$\hat{t}^{RW}_{H\times L} = t_{H\times L} + \hat{t}^{HT(S_3)}_{e(H\times L)} + B'_H\left(\hat{t}^{HT(S_1)}_{e(H)} - \hat{t}^{HT(S_3)}_{e(H)}\right) + B'_{L^-}\left(\hat{t}^{HT(S_2)}_{e(L^-)} - \hat{t}^{HT(S_3)}_{e(L^-)}\right)$$

Now collecting the residuals for each S_k ($k = 1, 2, 3$), we get in terms of superresiduals

$$\begin{aligned}
\hat{t}^{RW}_{H\times L} &= t_{H\times L} + \hat{t}^{HT(S_1)}_{\varepsilon_{1i,H\times L}} + \hat{t}^{HT(S_2)}_{\varepsilon_{2i,H\times L}} + \hat{t}^{HT(S_3)}_{\varepsilon_{3i,H\times L}} \\
\varepsilon_{1i,H\times L} &= B'_H e_i(H) \\
\varepsilon_{2i,H\times L} &= B'_{L^-} e_i(L^-) \\
\varepsilon_{3i,H\times L} &= e_i(H\times L) - B'_H e_i(H) - B'_{L^-} e_i(L^-)
\end{aligned} \tag{8}$$

Likewise, for the RW estimator of the target table $G\times H\times L$ we have

$$\begin{aligned}
\hat{t}^{RW}_{G\times H\times L} = {} & \hat{t}^{REG(S_3)}_{G\times H\times L} + \hat{B}'_{w;G\times H}\left(\hat{t}^{RW}_{G\times H} - \hat{t}^{REG(S_3)}_{G\times H}\right) \\
& + \hat{B}'_{w;G\times L^-}\left(\hat{t}^{RW}_{G\times L^-} - \hat{t}^{REG(S_3)}_{G\times L^-}\right) \\
& + \hat{B}'_{w;H^-\times L^-}\left(\hat{t}^{RW}_{H^-\times L^-} - \hat{t}^{REG(S_3)}_{H^-\times L^-}\right) \\
\hat{t}^{RW}_{G\times H} = {} & \hat{t}^{REG(S_1)}_{G\times H} \quad \text{and} \quad \hat{t}^{RW}_{G\times L^-} = \hat{t}^{REG(S_2)}_{G\times L^-}
\end{aligned} \tag{9}$$

while $\hat{t}^{RW}_{H^-\times L^-}$ is given by (8). Recall that Houbiers (2004) does not include $H\times L$ in the reweighting scheme because, strictly speaking, it is not needed to achieve numerical consistency or to reduce the variance of the estimator. The weighting scheme she uses, $(G\times H) + (G\times L)$, is called the *minimal reweighting* scheme for this estimator; note that table $t_{H\times L}$ cannot be estimated from a larger survey than S_3. Here, we have chosen to use the more elaborate *splitting up* procedure because this allows us to demonstrate the use of recursions in the repeated weighting procedure.

In terms of population residuals the RW estimator of $t_{G\times H\times L}$ can be written as

$$\begin{aligned}
\hat{t}^{RW}_{G\times H\times L} = {} & \text{constant} + \hat{t}^{HT(S_3)}_{e(G\times H\times L)} + B'_{G\times H}\left(\hat{t}^{HT(S_1)}_{e(G\times H)} - \hat{t}^{HT(S_3)}_{e(G\times H)}\right) \\
& + B'_{G\times L^-}\left(\hat{t}^{HT(S_2)}_{e(G\times L^-)} - \hat{t}^{HT(S_3)}_{e(G\times L^-)}\right) \\
& + B'_{H^-\times L^-}\left(\hat{t}^{RW}_{H^-\times L^-} - \hat{t}^{HT(S_3)}_{e(H^-\times L^-)}\right)
\end{aligned} \tag{10}$$

Now it is not difficult to see, from (10) that the superresiduals for $\hat{t}^{RW}_{G\times H\times L}$ are related to the superresiduals for $\hat{t}^{RW}_{H^-\times L^-}$ as follows

$$\varepsilon_{1i,G\times H\times L} = B'_{H^-\times L^-}\varepsilon_{1i,H^-\times L^-} + B'_{G\times H}e_i(G\times H)$$

$$\varepsilon_{2i,G\times H\times L} = B'_{H^-\times L^-}\varepsilon_{2i,H^-\times L^-} + B'_{G\times L^-}e_i(G\times L^-)$$

$$\varepsilon_{3i,G\times H\times L} = B'_{H^-\times L^-}\varepsilon_{3i,H^-}\times L^- + e_i(G\times H\times L) - B'_{G\times H}e_i(G\times H)$$
$$- B'_{G\times L^-}e_i(G\times L^-) - B'_{H^-\times L^-}e_i(H^-\times L^-)$$

Similar recursions can be derived when one or more elements in t_x from the regression in the second step are RW estimates from a larger block; see Appendix B.

Since S_1 equals about half of the population, the variance of an estimator from S_1 for a population mean is of order $1/N$. Hence, the random character of the estimators from S_1 can be ignored since $n_2 << N$. Therefore,

$$Cov\left(\hat{t}^{RW}_{G\times H\times L}\right) = Cov\left(\hat{t}^{HT(S_2)}_{\varepsilon_{2,G\times H\times L}} + \hat{t}^{HT(S_3)}_{\varepsilon_{3,G\times H\times L}}\right)$$
$$= Cov\left(\sum_{i\in S_2} d_i^{(S_2)}\varepsilon_{2i,G\times H\times L} + \sum_{i\in S_3} d_i^{(S_3)}\varepsilon_{3i,G\times H\times L}\right) \tag{11}$$

Since S_3 equals approximately half of S_2, the covariance of two arbitrary HT estimators from S_2 and S_3 need not be negligible. In order to capture this problem, define S_4 as the difference of S_2 and S_3, i.e., S_4 is the set of records in the LFS which are not included in the EWS. Now we can rearrange the summations in (11) over S_3 and S_4, yielding

$$Cov\left(\hat{t}^{RW}_{G\times H\times L}\right) = Cov\left(\sum_{i\in S_3} u_{3i} + \sum_{i\in S_4} u_{4i}\right) \tag{12}$$

$$u_{3i} = d_i^{(S_2)}\varepsilon_{2i,G\times H\times L} + d_i^{(S_3)}\varepsilon_{3i,G\times H\times L}$$

$$u_{4i} = d_i^{(S_2)}\varepsilon_{2i,G\times H\times L}$$

Since S_3 and S_4 are relatively small compared to the population, the two sample totals on the right-hand side of (12) can be interpreted as independent HH estimators from S_3 and S_4, respectively. Conditioning on their sample sizes n_3 and n_4, the covariance matrix of the RW estimator of $t^{RW}_{S\times H\times L}$ can now be estimated by

$$\hat{Cov}\left(\hat{t}^{RW}_{S\times H\times L}\right) = \sum_{k=3}^{4}\left\{\frac{n_k}{n_k-1}\sum_{i\in S_k}\hat{u}_{ki}\hat{u}'_{ki} - \frac{1}{n_k-1}\left(\sum_{i\in S_k}\hat{u}_{ki}\right)\left(\sum_{i\in S_k}\hat{u}'_{ki}\right)\right\} \tag{13}$$

Here we used (A1) from Appendix A because the population totals of the underlying variables $\sum_{i\in U} u_{ki}/d_i^{(S_k)}$ (k = 3, 4) need not be zero. As before, the estimated sample variates $\hat{u}_{ki}$ are based on the estimated regression matrices $\hat{B}$ in the corresponding

formulas. For a discussion on the use of conditional sample sizes, see among others Holt and Smith (1979) and Knottnerus (2003, pp. 133–135).

Finally, some remarks are in order. All data from the EWS, the LFS and the register are extracted from the social-statistical database and gathered together in the database for the Structure on Earnings Survey. As stated before, the population in SES consists of *jobs*, whereas EWS is a business survey and LFS is a household survey. Due to these and other complicating factors, the proper inclusion probabilities for the jobs in $S_1, S_2,$ and S_3 were missing in SES. Hence the proper design weights $d_i^{(S_k)}$ are unknown ($k = 1, 2, 3$). For this reason we used the actual regression weights $w_i^{(S_k)}$in the formulas for the variance estimator instead of the $d_i^{(S_k)}$. This will hardly have any effect on the variance estimators because the RW estimator is a linear combination of regression estimators from various samples. Furthermore, using g-weights according to Särndal et al. (1992, p. 235) in the estimation procedure for variances in combination with the $d_i^{(S_k)}$ is equivalent to the direct use of the regression weights $w_i^{(S_k)}$ as can be seen from (6); note that for an arbitrary sample $d_i g_i e_i = w_i e_i$. However, when sample sizes are large, the use of the g-weights is of little interest. Hence the use of $w_i^{(S_k)}$ instead of the $d_i^{(S_k)}$ will hardly affect the variance estimator in the RW estimation procedure described here. Moreover, for smaller samples the use of the g-weights is recommended.

5. Simulation Results

In order to test the RW estimator under various conditions, a number of simulations were performed. In these simulations, samples were drawn from an artificial population of 6.4 million jobs, which was generated from the Dutch Structure of Earnings Survey. In Van Duin and Snijders (2003) results are presented for the bias and variance of the RW estimator of the quantitative table *average monthly wage* by $G \times H_5 \times L$, where H_5 represents a more detailed classification of the variable working hours than H, with 5 instead of 2 classes (hierarchically related to H). The bias and variance were obtained both for the case of minimal repeated weighting and for the splitting up procedure. It was found that, except for very small sample sizes, repeated weighting only resulted in very limited additional bias compared to the standard regression estimator. The estimator for the splitting up procedure was found to yield no larger bias than the one for minimal repeated weighting. The additional calibrations on previously estimated table margins resulted in a smaller variance for the RW estimator than for the standard regression estimator. The RW estimator for the splitting up procedure was found to have a somewhat larger variance than the one employing minimal repeated weighting. However, this difference was much smaller than the difference in variance between the standard regression estimator and the minimal RW estimator.

In the same set of simulations, we also looked at the performance of the variance estimators described in the preceding sections by comparing them with the "actual" variances as obtained from the simulations. For simplicity, we only considered minimal repeated weighting. We present here the results for three different situations for the frequency table $G \times H_5 \times L$. Because only minimal repeated weighting was used, the marginal table $H_5 \times L$ was not included in the table set.

Simulation 1. The data set consists of a register and two independent simple random samples, both of size 100,000. Survey 1 contains the variable H_5, survey 2 the variables H_5 and L. In the notation of Section 3, H_5 now plays the role of Z, and L that of Y. The RW estimator for the target table is given by (10), with the following adjustments

- we have H_5 instead of H (and hence 70 table cells instead of 28)
- the table $H_5 \times L$ is not estimated and is omitted from the weighting scheme of the target table
- estimates from S_1 in (10) are in this case from $S_1 \cup S_2$, estimates from S_2 or S_3 are now from S_2.

Since S_1 and S_2 are drawn with the same sampling scheme, they are given the same weight in $S_1 \cup S_2$: $\lambda_1 = \lambda_2 = 1/2$; see (1).

Simulation 2. The samples are drawn with nonconstant inclusion probabilities π_i using Poisson sampling. The sample sizes are approximately the same as in simulation 1: $E(\hat{n}_1) \approx E(\hat{n}_2) \approx 100,000$. Even though Poisson sampling is not a fixed-size scheme, the variance formulas derived in the previous sections are also applicable to this simulation. This is the case because the estimators that are considered in the simulation are calibrated on the population size, which corrects for the variance contribution from $\hat{n}$, provided that $\pi_i << 1$ and $n >> 1$. The inclusion weights for S_1 are taken from a much narrower distribution than those for S_2. The squared coefficients of variance for the weights in the two samples are given by

$$L_1 \equiv \overline{d^{(S_1)^2}}/\overline{d^{(S_1)}}^2 - 1 = 0.3, \quad L_2 \equiv \overline{d^{(S_2)^2}}/\overline{d^{(S_2)}}^2 - 1 = 4.6$$

Consequently, the quality of the two samples is not the same, even though they have the same average size. It is important to take the quality difference into account when determining the relative weights of the two samples in $S_1 \cup S_2$. Kish (1992) argues that, if the target variable and the weights are weakly correlated, the effect of nonconstant weights on the sampling variance can be taken into account by replacing the sample size n with an effective sample size $n_{\text{eff}} = n/(1 + L)$. We estimate the sampling variances for S_1 and S_2 through this procedure and fix the relative weights of these samples in $S_1 \cup S_2$ by requiring that the sampling variance for estimates from $S_1 \cup S_2$ is minimized (we use the fact that the covariance of estimators from S_1 and S_2 can be neglected, since these samples are small and drawn independently). This results in $\lambda_p = n_{\text{eff},p}/(n_{\text{eff},1} + n_{\text{eff},2})$, with $p = 1, 2$.

Simulation 3. This simulation mimics the Structure of Earnings Survey, which was discussed in Section 4. S_1 is drawn with Poisson sampling using the inverse EWS weights, and S_2 is drawn using weights with the same distribution as those of the LFS. (The clustering-effects resulting from the fact that the LFS is a household survey and not a survey of jobs were not addressed in this simulation.) S_1 contains 2.8 million elements (half the population), S_2 approximately 100,000. The target table is estimated from $S_3 = S_1 \cap S_2$, which contains about 50,000 elements. S_3 has a very large variance of the inclusion weights, which it inherits from S_1. As a result $n_{\text{eff},3}$ is only about 5,000. S_2 has an effective size that is only somewhat reduced: $n_{\text{eff},2} \approx 77,000$.

The target table estimator is given by (10), with H replaced by H_5 and with $H_5 \times L$ omitted from the weighting scheme. The variance estimator is given by (13) with the appropriate modifications.

For each simulation, the relative bias in the standard error estimator for each cell is determined from

$$\Delta \hat{Ser}_p = \frac{\frac{1}{R}\sum_r \sqrt{\hat{Var}\left(\hat{t}^{\mathrm{RW}p}_{G\times H_5\times L,r}\right)}}{\sqrt{\frac{1}{R-1}\sum_r \left(\hat{t}^{\mathrm{RW}p}_{G\times H_5\times L,r} - \frac{1}{R}\sum_r \hat{t}^{\mathrm{RW}p}_{G\times H_5\times L,r}\right)^2}} - 1 \tag{14}$$

where r denotes the simulation run, R the number of runs (600) and p the table cell ($p = 1, \ldots, P$). Due to the finite number of simulation runs, (14) is not the exact bias but an estimate. It has an approximate 95% margin of ± 0.06. This follows from the assumption that $\hat{t}^{\mathrm{RW}p}_{G\times H_5\times L,r}$ is approximately normal with variance σ^2_{RW}. Denoting the variance estimator in the denominator in (14) briefly by $\hat{V}^{sim}_{RW}$, it follows from standard statistical theory that for $R \to \infty$

$$\sqrt{R-1}\left(\frac{\hat{V}^{sim}_{RW}}{\sigma^2_{RW}} - 1\right) \to N(0,\ 2)$$

in distribution; see e.g., Knottnerus (2003, p. 298). Hence, by a Taylor series linearization

$$\sqrt{R-1}\left(\frac{\sigma_{RW}}{\sqrt{\hat{V}^{sim}_{RW}}} - 1\right) \to N\left(0, \frac{1}{2}\right) \tag{15}$$

in distribution, from which it follows that the 95% margin of $\Delta \hat{Ser}_p$ is approximately ± 0.06 [$= 1.96/\sqrt{2(R-1)}$]. Note that these margins are not affected by replacing σ_{RW} in the numerator of (14) by an estimator $\hat{\sigma}_{RW}$ with variance of order N^2/n^2R because the denominator has a variance of order N^2/nR. The latter result follows from a Taylor series expansion of (15) which yields

$$Var\left(\frac{\sqrt{\hat{V}^{sim}_{RW}}}{\sigma_{RW}}\right) = \frac{1}{2(R-1)} \quad \text{with} \quad \sigma^2_{RW} = O\left(\frac{N^2}{n}\right)$$

Hence, $Var\left(\sqrt{\hat{V}_{RW}^{sim}}\right) = O(N^2/nR)$. In addition, the variance of the estimator $\hat{\sigma}_{RW}$ in the numerator of (14) is of order N^2/n^2R because

$$Var(\hat{\sigma}_{RW}) = O\left\{\frac{1}{R}Var\left(\sqrt{\frac{N^2 s_{pr}^2}{n}}\right)\right\} = O\left\{\frac{N^2}{nR}Var(s_{pr})\right\} = O\left(\frac{N^2}{n^2R}\right)$$

$$s_{pr}^2 \equiv \frac{1}{n-1}\sum_{i\in S}(y_{ipr} - \bar{y}_{pr})^2 \quad (y_{ipr} = [l_i \otimes h_{5i} \otimes g_i]_{p,r})$$

Figure 2 shows $\Delta\hat{Ser}_p$ as a function of the effective cell size for the three simulations. Every data point corresponds to a cell of the target table for a particular simulation. The results clearly demonstrate the asymptotic nature of the variance estimators (7) and (13). Their derivation relies on a Taylor linearization procedure in which the sampling variance of the estimated regression coefficient B is neglected. This procedure works well for large (effective) cell sizes, but leads to an underestimation if the number of observations becomes too small. In simulation 1, the smallest cell size is 51 and no structural bias in the variance estimator is found. In simulations 2 and 3, effective cell sizes as low as 1 occur, which leads to a serious negative bias. (Although simulation 3 mimics the SES, the target table estimated here is much more detailed than the ones estimated in the SES.) The data points for the three simulations follow broadly the same curve. Apparently, the effect of nonconstant inclusion probabilities on $\Delta\hat{Ser}$ is taken into account well by replacing n with n_{eff} (for this particular estimator). Also, it does not seem to affect $\Delta\hat{Ser}$ whether the table is estimated from a union or an intersection of samples, and hence whether the variance estimator (7) or (13) is used. We do expect $\Delta\hat{Ser}$ to depend on the calibration scheme of the target table. The same scheme was used in all three simulations. If fewer calibrations were included, the value of n_{eff} where the linearization procedure breaks down could shift downward.

The negative bias at small cell sizes is not specific to the RW variance estimator. In fact, the linearized variance estimator for the regression estimator (2) suffers from the same

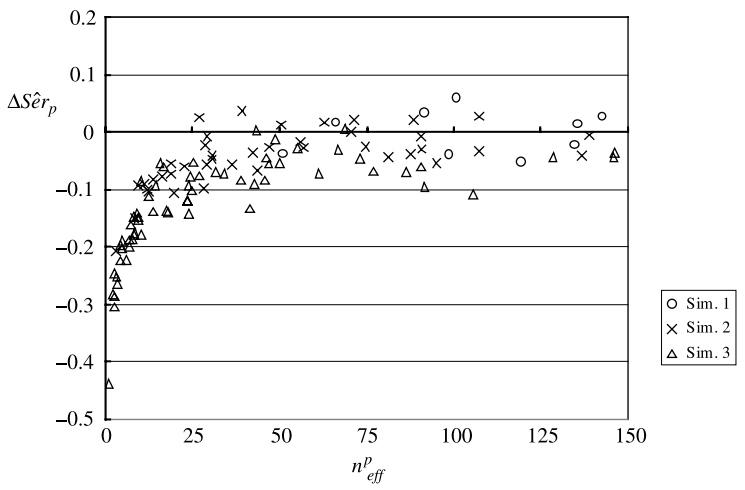

Fig. 2. Relative bias in the standard error estimator for cell-counts of the target table $G \times H_5 \times L$ in the three simulations

problem. Because the RW estimator involves a large number of calibrations, the negative bias for small samples is likely worse than for (2). However, Boonstra (2004) argues that this can be compensated by the smaller variance of the RW variance estimator and that, consequently, the mean squared errors of the variance estimator for RW and for the regression estimator are often comparable.

6. Discussion and Further Research

Technically speaking, repeated weighting amounts to a further cosmetic adjustment of the commonly used regression weights w_i resulting in new final weights r_i. This possible cosmetic adjustment resembles the adjustment in the regression or calibration estimator used by many National Statistical Institutes, where the regression weights w_i can be seen as an adjustment of the starting weights $(1/\pi_i)$. The consequence is that the final weights may vary from table to table.

Currently, Statistics Netherlands is implementing the repeated weighting (RW) estimator in its regular estimation process to obtain consistency among tables. Apart from the assumptions mentioned in Section 2 it also is important for applying this RW estimation procedure to have an appropriate metadata system underlying the micro databases from the surveys and available registers. For instance, Statistics Netherlands has a software tool, called VRD, for the collection of tables related to a given target table. Such a tool is necessary when there are many multidimensional tables to be estimated with variables with many different (hierarchical) classifications or variables like income in either categorical or quantitative form.

This article focuses on classification variables in order to avoid a number of complications caused by the use of quantitative variables. A complicating factor for the latter type of variables is that a quantitative variable, such as *income*, can be used both as a classification and as a quantification variable. A problem that may arise in this context is the consistency of a table on total income per income class. That is, the mean income of a low income class must be lower that the mean income of higher income classes. This problem may arise when the number of persons in an income class is estimated independently; see Renssen et al. (2001). Another problem is the treatment of edits of the form: i) the number of persons in a certain region with a driver's licence cannot exceed the number of persons who are 18 or older in that region, or ii) costs plus profits must be equal to the turnover of all enterprises. Van de Laar (2004) points out how this kind of edits can be incorporated in the RW estimation strategy. Although the RW estimation procedure becomes somewhat more complicated, the same variance formulas can be applied.

Different simulations indicate that compared to the variance estimator of the regression estimator the (negative) bias of the variance estimator of the RW estimator slightly increases but both the variance and the mean squared error of the RW variance estimator decrease provided that the effective sample sizes are not too small. In practice the numerical differences between the *splitting up* and *minimal weighting* procedures are small. The former has only a somewhat larger variance than the latter. However, this difference is smaller than the difference in variance between the standard regression estimator and the minimal RW estimator provided that the cell size is large enough. Further research is needed to investigate various methods for improving the variance formulas in the case of small cell sizes.

Appendix A: Variance Estimators for the Hansen-Hurwitz (HH) Estimator

Consider a sample S of n independent drawings $y_1, \ldots, y_n$ with replacement from a population U of N numbers $Y_1, \ldots, Y_N$. For each drawing the drawing probabilities are $p_1, \ldots, p_N$. The Hansen-Hurwitz estimator of population total t_Y and its variance are given by

$$\hat{t}_Y^{HH} = \frac{1}{n}\sum_{i\in S}\frac{y_i}{p_i} = \sum_{i\in S} d_i y_i \left(d_i = \frac{1}{np_i}\right)$$

$$\mathrm{Var}\left(\hat{t}_Y^{HH}\right) = \frac{1}{n}\sum_{i\in U} p_i\left(\frac{Y_i}{p_i} - t_Y\right)^2$$

respectively. The variance can be estimated unbiasedly by

$$\begin{aligned} \hat{Var}\left(\hat{t}_Y^{HH}\right) &= \frac{1}{n(n-1)}\sum_{i\in S}\left(\frac{y_i}{p_i} - \frac{1}{n}\sum_{i\in S}\frac{y_i}{p_i}\right)^2 \\ &= \frac{1}{(n-1)}\left\{\frac{1}{n}\sum_{i\in S}\left(\frac{y_i}{p_i}\right)^2 - \left(\frac{1}{n}\sum_{i\in S}\frac{y_i}{p_i}\right)^2\right\} \\ &= \frac{n}{n-1}\sum_{i\in S}(d_i y_i)^2 - \frac{1}{n-1}\left(\sum_{i\in S} d_i y_i\right)^2 \end{aligned}$$

In other words, writing the HH-estimator as $\hat{t}_Y^{HH} = \sum_{i\in S} a_i$ with $a_i = d_i y_i$, its variance can be estimated by

$$\hat{Var}\left(\hat{t}_Y^{HH}\right) = \frac{n}{n-1}\sum_{i\in S} a_i^2 - \frac{1}{n-1}\left(\sum_{i\in S} a_i\right)^2$$

When y_i is a random P-vector corresponding to a categorical variable Y, the corresponding covariance matrix of $\hat{t}_Y^{HH}$ becomes

$$\hat{Cov}\left(\hat{t}_Y^{HH}\right) = \frac{n}{n-1}\sum_{i\in S} a_i a_i' - \frac{1}{n-1}\left(\sum_{i\in S} a_i \sum_{i\in S} a_i'\right) \tag{A1}$$

We use (A1) in Sections 3 and 4. If we have prior knowledge that $t_Y = 0$, an unbiased estimator of the covariance matrix is

$$\hat{Cov}\left(\hat{t}_Y^{HH}\right) = \sum_{i\in S} d_i^2 y_i y_i' \tag{A2}$$

This is of interest for the estimator of the covariance matrix of a vector regression estimator with a constant among the auxiliaries in x_i so that

$$\sum_{i\in U} e_i = \sum_{i\in U}(y_i - B_x' x_i) = 0$$

$$B_x = \left(\sum_{i\in U} x_i x_i'\right)^{-1} \sum_{i\in U} x_i y_i'$$

Note that $\text{Col}(E) \perp \text{Col}(X)$ and hence, $t_e = 0$ provided $\iota_N \in \text{Col}(X)$, where ι_N is an N-vector of ones, E is the $N \times P$ matrix of residuals, X the $N \times J$ matrix of auxiliaries, and Col(.) stands for the column space spanned by the columns of (.); see Knottnerus (2003, p. 25). Also note that in practice B is to be replaced by its estimator $\hat{B}_{d,x}$, leading to

$$\hat{Cov}\left(\hat{t}_Y^{REG}\right) = \sum_{i\in S} d_i^2 \hat{e}_i \hat{e}_i' \quad (\hat{e}_i = y_i - \hat{B}_{d,x}' x_i) \tag{A3}$$

Appendix B: Recursions in the Case of Estimated Totals Among the Auxiliaries

In this appendix, we consider the situation described in Section 3 with two independent samples and a register. In addition, we assume that, quite generally, some elements in the vector t_x, used for the regressions in block S_2, are RW estimates from S_{12}. We write this preceding RW estimator in terms of superresiduals as

$$\hat{t}_x^{RW} = t_x + \sum_{i\in S_1} d_i^{(S_1)} \varepsilon_{1i,x} + \sum_{i\in S_2} d_i^{(S_2)} \varepsilon_{2i,x}$$

Neglecting the random character of the estimated matrices $\hat{B}$ in the remainder, the regression estimator for an arbitrary frequency table t_V from S_2 can be written as

$$\begin{aligned} \hat{t}_V^{REG(S_2)} &= \hat{t}_V^{HT(S_2)} + B_{V;x}'\left(\hat{t}_x^{RW} - \hat{t}_x^{HT(S_2)}\right) \\ &= t_V + \hat{t}_{e(V)}^{HT(S_2)} + B_{V;x}'\left(\hat{t}_{\varepsilon_{1,x}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,x}}^{HT(S_2)}\right) \\ B_{V;x} &= \left(\sum_{i\in U} x_i x_i'\right)^{-1} \sum_{i\in U} x_i v_i' \end{aligned} \tag{B1}$$

Consider now the RW estimator for t_Y from S_2 with the initial regression estimators being based on $\hat{t}_x^{RW}$. That is,

$$\hat{t}_Y^{RW} = \hat{t}_Y^{REG(S_2)} + B_{Y;m}'\left(\hat{t}_m^{RW} - \hat{t}_m^{REG(S_2)}\right) \tag{B2}$$

where according to (B1) the regression estimators are given by

$$\hat{t}_Y^{REG(S_2)} = t_Y + \hat{t}_{e(Y)}^{HT(S_2)} + B'_{Y;x}\left(\hat{t}_{\varepsilon_{1,x}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,x}}^{HT(S_2)}\right)$$

$$\hat{t}_m^{REG(S_2)} = t_m + \hat{t}_{e(m)}^{HT(S_2)} + B'_{m;x}\left(\hat{t}_{\varepsilon_{1,x}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,x}}^{HT(S_2)}\right) \tag{B3}$$

Furthermore, by construction, the estimate $\hat{t}_m^{RW}$ from the preceding tables can be written as

$$\hat{t}_m^{RW} = t_m + \hat{t}_{\varepsilon_{1,m}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,m}}^{HT(S_2)} \tag{B4}$$

Substituting (B3) and (B4) into (B2) gives

$$\begin{aligned}\hat{t}_Y^{RW} =& t_Y + \hat{t}_{e(Y)}^{HT(S_2)} + B'_{Y;x}\left(\hat{t}_{\varepsilon_{1,x}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,x}}^{HT(S_2)}\right) \\ &+ B'_{Y,m}\left\{\hat{t}_{\varepsilon_{1,m}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,m}}^{HT(S_2)} - \hat{t}_{e(m)}^{HT(S_2)} - B'_{m;x}\left(\hat{t}_{\varepsilon_{1,x}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,x}}^{HT(S_2)}\right)\right\}\end{aligned} \tag{B5}$$

On the other hand writing $\hat{t}_Y^{RW}$ in terms of superresiduals

$$\hat{t}_Y^{RW} = t_Y + \hat{t}_{\varepsilon_{1,Y}}^{HT(S_1)} + \hat{t}_{\varepsilon_{2,Y}}^{HT(S_2)} \tag{B6}$$

it follows from comparing (B5) and (B6) that the superresiduals for Y from the reweighting step obey the recursions

$$\varepsilon_{1i,Y} = (B'_{Y;x} - B'_{Y;m}B'_{m;x})\varepsilon_{1i,x} + B'_{Y;m}\varepsilon_{1i,m}$$

$$\varepsilon_{2i,Y} = (B'_{Y;x} - B'_{Y;m}B'_{m;x})\varepsilon_{2i,x} + B'_{Y;m}\varepsilon_{2i,m} + e_i(Y) - B'_{Y;m}e_i(m)$$

7. References

Boonstra H.J.H. (2004). DACSEIS deliverable 7.3: A Simulation Study of Repeated Weighting Estimation. Heerlen: Statistics Netherlands.

Boonstra, H.J.H., van den Brakel, J.A., Knottnerus, P., Nieuwenbroek, N.J., and Renssen, R.H. (2003). DACSEIS deliverable 7.2: A Strategy to Obtain Consistency Among Tables of Survey Estimates. Heerlen: Statistics Netherlands.

Deville, J.C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Duin, C. van and Snijders, V. (2003). Simulation Studies of Repeated Weighting. Discussion Paper, Voorburg: Statistics Netherlands.

Holt, D. and Smith, T.M.F. (1979). Poststratification. Journal of the Royal Statistical Society, Series A, 142, 33–46.

Houbiers, M. (2004). Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics, 20, 55–75.

Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H., and Snijders, V. (2003). Estimating Consistent Table Sets: Position Paper on Repeated Weighting. Discussion Paper, Voorburg: Statistics Netherlands.

Kish, L. (1992). Weighting for Unequal P_i. Journal of Official Statistics, 8, 183–200.
Knottnerus, P. (2001). Variances in Repeated Weighting. Voorburg: Statistics Netherlands. [In Dutch]
Knottnerus, P. (2003). Sample Survey Theory: Some Pythagorean Perspectives. New York: Springer-Verlag.
Kroese, A.H. and Renssen, R.H. (1999). Weighting and Imputation at Statistics Netherlands. Proceedings of the IASS Conference on Small Area Estimation, Riga, 109–120.
Renssen, R.H., Kroese, A.H., and Willeboordse, A. (2001). Aligning Estimates by Repeated Weighting. Heerlen: Statistics Netherlands.
Renssen, R.H. and Martinus G.H. (2002). On the Use of the Generalized Inverse in Sampling Theory. Survey Methodology, 28, 209–212.
Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.
Van de Laar, R.W.A. (2004). Edit Rules and the Strategy of Consistent Table Estimation. Discussion Paper, Voorburg: Statistics Netherlands.

Received October 2004
Revised October 2005