Estimating consistent table sets: position paper on repeated weighting

Discussion paper 03005

M. Houbiers, P. Knottnerus, A.H. Kroese, R.H. Renssen, and V. Snijders

The authors would like to thank Coen van Duin for allowing to use results from simulation studies regarding variances and biases of the estimation method. Chris Skinner is thanked for various comments and suggestions on the method of repeated weighting.

The views expressed in this paper are those of the author
and do not necessarily reflect the policies of Statistics Netherlands

Statistics Netherlands

Voorburg/Heerlen, August 2003

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2002–2003	= 2002 to 2003 inclusive
2002/2003	= average of 2002 up to and including 2003
2002/’03	= crop year, financial year, school year etc. beginning in 2002 and ending in 2003

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10
Production code: 6008303005

Statistics Netherlands

ESTIMATING CONSISTENT TABLE SETS: POSITION PAPER ON REPEATED WEIGHTING

Summary: Statistics Netherlands is presently making a considerable effort in combining data from administrative sources with, mainly household, survey data. By making efficient use of register data, Statistics Netherlands intends to improve the accuracy of its statistical information, and, at the same time, to decrease the response burden on households. The resulting large micro dataset with combined data is called the 'Social Statistical Database' (SSD); estimates related to social statistics are obtained from this SSD. Preferably, these estimates should be numerically consistent, although they might be obtained from different sources. At Statistics Netherlands, a new estimation method has been developed which, under certain conditions, ensures numerically consistent table sets. This method is called 'repeated weighting', and is based on a repeated use of the regression estimator. In the present paper we describe this new estimation method. We will also derive approximate variance formulas for the estimates. Simulation studies show that the method leads to estimates with lower variances than usual estimation methods, due to a better use of auxiliary information. As an illustration, we describe two cases in which the method is applied to produce numerically consistent table sets from the SSD.

Keywords: Benchmarking, combining registers and surveys, consistent estimates, record linkage, regression estimator, simulations, variance estimation.

1. Introduction

In recent years, Statistics Netherlands has focused on an increased use of registration data instead of survey data in the production process of statistical information. This shift is mainly provoked by growing users' demands for detailed, coherent, timely, relevant, and authoritative statistical information. This statistical information should provide an insight into complex relationships between different aspects of social and economic life, see Bakker (2002) and Statistics Netherlands (2000).

These demands can, in principle, be met by sending out comprehensive surveys, covering as many variables as possible, and preferably with sample sizes large enough to allow for detailed information on small subgroups in the population. However, there are some serious limitations to this approach. First of all, nonresponse rates in The Netherlands are generally quite high, especially in the case of household surveys. Since nonresponse is usually selective, it may lead to severe biases, resulting in a questionable quality of the estimates. In addition, there is a strong political pressure to minimize the response burden on households and enterprises. Finally, there is a continuous demand to cut down staff and costs at Statistics Netherlands. In short, these extensive surveys offer in reality no feasible way to satisfy the users' demands.

An alternative approach is to make efficiently use of data from administrative registrations. Due to the improved capacity of computers, the size of data sets is, after all, no longer an obstacle in the statistical process. Registrations offer an efficient and – also important – relatively cheap source of statistical information: large numbers of records can be obtained in one swoop, without too much effort. Examples of administrative registrations are the population registration, and social security and tax data. Although registrations may in general not contain all variables that one is interested in, they contain information on more or less all elements in the population, or a certain subpopulation. Therefore, they do allow for detailed information on certain small subgroups in the population. Furthermore, by linking registrations, cross links between different fields can be generated, thus generating possibilities for completely new output. Information that is not present in registrations still needs to be collected via additional surveys. But since a large part of the required information is already gathered from registrations, the response burden on enterprises and households in surveys will be lowered substantially. The registration data offer, in addition, much more auxiliary information to correct the surveys for nonresponse and to reduce the variances of estimates, thus improving the accuracy of estimates.

The use of register data in combination with survey data is widely recognized by National Statistical Institutes (NSI's) as a way to improve the quality of estimates. An investigation among European NSI's with respect to the use of auxiliary information from the available registers for the Labor Force Survey shows, however, that the majority of countries do not use registers due to law and privacy reasons, linking key problems, the complete absence of (suitable) population registers, or bias and frame errors in the registers, see Knottnerus and Wiegert (2002). The NSI's that do use register data, use (post)stratification, the regression estimator, calibration and raking methods, and sometimes imputation to correct for nonresponse, to reduce the bias, increase the accuracy of estimates, and to secure (some) consistency between estimates from various sources, see, for instance, Thomson and Kleive Holmøy (1998).

Statistics Netherlands started a few years ago to exploit the potential of administrative sources by contructing a large database in which all available data (from registrations and surveys) related to social statistics are linked and stored, see Statistics Netherlands (2000). This database, which eventually will contain all relevant micro data on persons, families, households, jobs, social benefits, and living quarters in The Netherlands, is called the 'Social Statistical Database' (SSD).

Output related to social statistics will be obtained from the SSD and published, for instance, via StatLine on the internet. Since estimates from the SSD may be based on data from different sources, an estimate for a certain statistical quantity may be numerically inconsistent with other estimates for the same variable. These inconsistencies become much more visible due to the large-scale dissemination of statistical information via internet, and may cause confusion among the users of statistical information. Statistics Netherlands' figures should, after all, be unquestionable. Statistics Netherlands tries to track down and remove such inconsistencies whenever possible. An important issue at Statistics Netherlands is evidently to prevent inconsistencies in estimates in the first place. Therefore, a major goal has been to develop an estimation method that guarantees – as much as possible – that estimates are numerically consistent with each other. With the development of the method of 'repeated weighting', see Kroese and Renssen (1999, 2000) and Renssen et al. (2001), Statistics Netherlands has to a large extent succeeded in this mission.

Repeated weighting is a method that can be used to estimate well-defined, relatively simple, table sets. The method uses the calibration properties of the regression estimator to enforce numerical consistency between all estimates in the table set, even though each table may be estimated from different micro data. Roughly speaking, repeated weighting works as follows. Each table is estimated using as many records as possible: depending on the variables of interest, the table may be counted from register data, or estimated from survey data from one or more surveys. For each table it is determined which margins the present table has in common with the tables in the set that are already estimated. The table is subsequently estimated by calibrating on these common margins. This ensures that the table estimate is numerically consistent with all earlier estimated tables in the set. At the same time, the estimate may be more accurate, because the margins may be estimated from larger data sets or even counted from register data, and as a result serve as auxiliary information for the table estimate.

The technique of repeated weighting is designed to prevent as much as possible numerical inconsistencies due to survey errors. Although the method may yield estimates with lower variances, it is in the first place applied for cosmetic purposes. It is, however, not applicable in every situation. The method is not suitable when fast estimates, or estimates on several subpopulations are required, or when numerous edit rules between variables define extra constraints on the estimates. But for estimates for which timeliness is not, and numerical consistency ís important, the method is a useful tool to produce estimates.

The estimation strategy envisioned by Statistics Netherlands is as follows. Suppose one wants to estimate a set of tables which all relate to the same target population. From all the data available in the SSD, the part that is relevant for this tableset is then isolated from the SSD: that is, a rectangular dataset with all elements of the target population as rows and all variables in the tables as columns is extracted from the SSD. This rectangular dataset may contain a lot of empty cells. Depending on the particular purposes of some publication, a suitable estimation strategy is chosen to obtain estimates from the extracted data set. One can think of imputation (common in, for instance, household statistics), usual weighting methods (for fast production of provisional figures), or repeated weighting (when numerical consistency is important).

In this paper we describe the current best practices of repeated weighting. The paper is organized as follows. In Section 2, we focus on the social statistical database and its present status. In Section 3, we explain the estimation strategy of repeated weighting in a mathematical framework. We derive approximate variance formulas for the repeated weighting estimator. We also mention limitations of the method. In Section 4, we give two examples of table sets that were estimated from the social statistical database, using repeated weighting. The first example concerns the Structure of Earnings Survey, the second table set involves the search behavior on the labor market of persons on social welfare. In Section 5, we present some results from a simulation study in which biases, variances and estimated variances of repeated weighting estimators are tested. In Section 6, we summarize and conclude.

2. Social statistical database

As explained in the Introduction, Statistics Netherlands is constructing a Social Statistical Database (SSD), in which data from administrative sources and surveys are linked. These linked data sets provide a very rich source of information on social statistical topics. In this section, we give a general overview of the data which are presently available in the SSD, and roughly explain how Statistics Netherlands intends to obtain estimates from it.

2.1 Sources of data

For the construction of the SSD, several registrations are linked to each other as well as to survey data sets. The registrations that are available at present at Statistics Netherlands include the Municipal Base Administration (MBA), the jobs register (both for employers and for employees), and the social welfare payments register.[1] The first registration contains information on age, sex, ethnicity, place of birth, place of residence, marital status etc., for persons in The Netherlands, except for illegals. The second registration contains information such as size class and activity code of the corresponding business on all jobs in The Netherlands. Via a unique key based

[1] The jobs and social welfare payments registers are constructed at Statistics Netherlands. They are based on other data sources from, e.g., the tax offices, employee insurance registrations, and social welfare agencies. Clearly, the jobs and social welfare registers are no administrative registrations in the usual sense. They are in fact ‘integration data sets’; it takes a while before these data sets become available, so they are not up-to-date. Despite of that, we address them as ‘registers’ in this paper.

on the social security number, these jobs can be linked to persons,[2] or persons can be linked to jobs, depending on the population one is interested in. The third registration contains information on social welfare payments (such as type of social welfare, amount, and duration of payment). This registration can also via the social security numbers be linked to persons- and the jobs registers. All three registrations are volume registrations, which means that they contain longitudinal information about all elements in the population during a certain time period. Therefore, they can be linked at any day of the year, thus creating a linked register at a certain reference date. By linking the registrations at two or more days of the year, and subsequently averaging, an (approximate) average register is obtained. Depending on the estimates one is interested in, an average register or a register at a certain reference date is used as the backbone in the estimation process. Many cross tabulations can be counted from the combined registers.

In addition to linking the registrations to each other, survey data are linked to the register data. In principle, all household and person surveys are already linked to the MBA (via the unique key mentioned earlier), so these data sets can without much effort be added to the SSD. Examples of sample surveys that at present are linked in the SSD, are the Employment and Wages Survey (EWS), the Labor Force Survey (LFS), and the Integrated System on Social Surveys (ISSS, the Dutch equivalent of the Living Conditions Survey). The EWS is a large two-stage survey among businesses and contains information on, for example, wages and hours of employment of employees. The LFS is a household survey and contains variables such as profession, education, and search behavior on the labor market. The ISSS is a person survey, and contains information related to, for instance, education and health. The register and survey data presently available in the SSD are used in the applications in Section 4.

In order to obtain unbiased estimates, these surveys must relate to the same time period as the register data. In particular, the survey data must be linked to the corresponding records in the registrations at the survey date. This is especially true for variables that change rapidly with time, such as search behavior on the labor market. Variables that are relatively fixed, such as educational level or profession, can be linked ‘around’ survey date, that is, they can be linked to the register data at a certain desired reference date not too far from the survey date, as if they were

[2] Ideally, the records in the registers and surveys are equipped with some unique key so that they can be linked at the micro level. In practice, such a unique key must often be derived from certain identifiers. In The Netherlands, most people have a social security number. The social security number can, with a check on date of birth and gender, be used as a unique key to link records. For people or records without a social security number, the identifiers date of birth, gender, postal code, and number of the house (at a certain point in time), are used to link records. However, this combination is in a small number of cases not unique as, for instance, for identical twins living at the same address. Still, the fraction of exact matches is close to one hundred percent. The fraction of mismatches and missed matches is small (less than one percent) and assumed not to affect the estimates.

collected at this reference date. When calibrating the surveys on register totals, one should use a register that relates to the same time period as the surveys. Thus, in the first case, an average of the register over the time period of the survey is required. In the second case, the survey is assumed to be collected at the reference date, and a cross section of the linked registers at this particular reference date can be used.

2.2 Estimates from the SSD

As described in the previous section, the SSD contains several registrations and survey data sets. These data sets each relate to specific variables, and from these data estimates on topics related to social statistics can be obtained. The SSD contains far more data than is relevant for some specific table set that one wants to publish. Therefore, one first of all should isolate and link the data that are useful for that particular table set: all elements (records) that belong to the target population, and for each element all available variables one is interested in. An important idea of the SSD is to use for each variable of interest all data that is available in the SSD. Some variables are available for all elements in the target population since they are present in one of the registers, others only for a fraction of the population because these variables were observed in one or more surveys. Note that for any table set related to a fixed target population, for a certain variable the same micro data should be isolated from the SSD.

Next, the table set should be estimated using the selected data. The more register variables are available, the more cross tabulations can be obtained by mere counting. One of the merits of the SSD is that several linked registers indeed áre available, thus generating many opportunities for high-quality information on cross links between different fields. Register counts from the SSD are always consistent, irrespective of the specific target population or the object type one is focused on.

Cross tabulations relating to survey variables must evidently be estimated.[3] An important advantage of the SSD is that known (or estimated) population totals of register (or survey) variables can be used as auxiliary information in the estimation process. By choosing suitable auxiliary variables, it is well known that this improves the accuracy of these estimate considerably, as does the basic principle that for each estimate as much micro data as possible is used.

Estimates from survey data are always numerically consistent with each other when they are based on the same micro data, and the same weights are used. However,

[3] We focus here on usual estimation strategies, that is, estimation by weighting. Alternatively, all missing fields can be imputed by some suitable imputation strategy. The table set can then simply be ‘counted’ from the resulting complete data set. Mass imputation is common practice in, for instance, household statistics. Although imputation models are better when more register information is available, these models are never sufficiently rich to account for all significant data patterns between sample and register data, and may easily lead to oddities in the estimates, see Kooiman (1998). Therefore, traditional estimation by weighting is favored over mass imputation at Statistics Netherlands.

they are generally not numerically consistent with register counts, except for the few register variables used as auxiliary information in the estimation process. So, when estimating a table set using combined data from different sources, some tables in this table set are numerically consistent, but others are not. Note that for a fixed target population, consistency between tables in different table sets is automatically achieved when these tables exactly relate to the same variables, and the same weights (and records, of course) are used for both estimates.

In practice, especially for provisional estimates, numerical consistency and to a lesser degree accuracy of estimates is of minor importance compared to timeliness. Therefore, to produce fast, preliminary estimates, data from one single survey are mostly used, instead of using all combined data in the SSD. There is simply no time to wait until all (register and survey) data is available in the SSD. For these fast estimates, auxiliary information is not used optimally, resulting obviously in estimates of lower accuracy. However, when it comes to final figures, Statistics Netherlands' policy is to prevent numerical inconsistencies as much as possible. It will certainly not be possible to obtain full numerical consistency between *all* possible estimates and counts from the SSD. However, with the method of repeated weighting, well-defined table sets can be estimated consistently. In addition, consistency between table sets is obtained in part with repeated weighting. The repeated weighting strategy always leads to consistent population totals when it comes to register variables, irrespective of the table set or the target population one is looking at. Moreover, estimated totals of survey variables are under certain conditions also consistent, even when they concern different table sets: they are consistent when the same micro data and weights are used for these estimates. In short, repeated weighting automatically leads to much more numerical consistency between estimates than ever can be obtained with simple weighting procedures. In the next section, we give a detailed overview of this new estimation strategy.

3. Theory of repeated weighting

In this section, we describe the method of repeated weighting. For convenience, we first introduce some terminology and notation, based on Renssen *et al*. (2001).

3.1 Terminology and notation

In this paper, we are dealing with table estimates. Tables are formed by aggregating micro data. A table T is characterized by one *count variable*, specifying what exactly is counted in the table (such as total number of elements in some population, or total income of the elements in the population), and one or more *dimension variables*, specifying the background variables (such as sex, and age class) against which the count variable is counted. The former type of variable is also called *quantification* variable, the latter *classification* variables.

A classification variable is a variable whose values represent categories by which the population is partitioned into an exclusive and exhaustive set of subpopulations. Let X_g be such a classification variable. In practice, a classification variable X_g may partition the population in several different ways. For instance, the classification variable 'age class' may alternatively consist of 10-year classes, 5-year classes, or 1-year classes. With repeated weighting, classification variables are allowed to consist of several levels of classification, *as long as these levels are hierarchical*. The degree of classification of a classification variable will be denoted by a superscript r, that is, $X_g^{(r)}$. The larger r, the finer the classification. The minimum value of r equals 0, corresponding to a 'partitioning' of the population into one single class. The finest classification of X_g will be denoted by $X_g^{(R)}$. The levels of a classification variable X_g are hierarchical if and only if one or more classes of level r together make up one class of level $r-1$ for all $r \in [1,...,R]$.

Suppose that G classification variables are distinguished in the population. Each classification variable X_g has $R_g \geq 1$ hierarchical levels. A table T can be specified by the expression

$$T = [X_1^{(r_1)} \times \ldots \times X_G^{(r_G)}] \times Y ,$$

where Y denotes the count variable of interest in the table. The dimensional part of this table (the part between the square brackets) can alternatively be expressed as a G-vector $\mathbf{r} = (r_1, ..., r_G)'$, with $0 \leq r_g \leq R_g$. If $r_g = 0$, the corresponding classification variable X_g is, in fact, not present, in table T. For an M-way table, only M out of G components of $\mathbf{r}$ are nonzero.

For repeated weighting, it will be necessary to quickly evaluate common margins of tables. With the above notation, the common margin of two tables can easily be derived. Suppose that two tables T_1 and T_2 relate to the same count variable Y. The dimensions of T_1 are specified by the vector $\mathbf{r}_1 = (r_{11}, ..., r_{1G})'$, and the dimensions of T_2 by $\mathbf{r}_2 = (r_{21}, ..., r_{2G})'$. It is easy to see that the common margin $T_{1,2}$ of these two tables is specified by the G-vector

$$\mathbf{r}_{1,2} = \min(\mathbf{r}_1, \mathbf{r}_2) \equiv (\min(r_{11}, r_{21}), \min(r_{12}, r_{22}), \ldots, \min(r_{1G}, r_{2G}))'.$$

Whenever one or more components of $\mathbf{r}_{1,2}$ are nonzero, the two tables have a nontrivial common margin. If, for example, the variable 'age class' consists of the three levels 10-year classes (level 1), 5-year classes (level 2), and 1-year classes (level 3), and T_1 contains, among others, the classification variable 'age class' at the level of 5-year classes, whereas T_2 contains this variable at the level of 10-year classes, then these two tables obviously have a common margin related to 'age class' at the level of 10-year classes, the minimum of (2,1). When all components of $\mathbf{r}_{1,2}$ are zero, tables T_1 and T_2 have no common margins, except for the population total of the count variable Y.

When deriving approximate variance formulas for repeated weighting estimators, it will be convenient to use a vector notation to indicate tables. Suppose that level r of classification variable X_g consists of $P_g \geq 1$ mutually exclusive classes. We represent the score on classification variable $X_g^{(r)}$ of an element k in the population as a P_g-vector of dummy variables, that is, we define

$$\mathbf{x}_{gk}^{(r)} = (x_{gk1}, \ldots, x_{gkP_g})' ,$$

where

$$x_{gkp} = \begin{cases} 1 \text{ if element } k \text{ belongs to class } p \text{ of variable } X_g^{(r)} \\ 0 \text{ otherwise.} \end{cases}$$

The score of an element k in the population on the crossing $X_1^{(r_1)} \times \ldots \times X_G^{(r_G)}$ can be expressed in a similar way. The crossing now consists of $P = P_1 \times \ldots \times P_G$ mutually excluding classes, where each P_g depends on the level r_g of variable X_g in the crossing. A P-vector can be defined as

$$\mathbf{x}_k^{(\mathrm{r})} = (x_{k1}, \ldots, x_{kP})' ,$$

where

$$x_{kp} = \begin{cases} 1 \text{ if element } k \text{ belongs to class } p \text{ of the crossing } X_1^{(r_1)} \times \ldots \times X_G^{(r_G)} \\ 0 \text{ otherwise.} \end{cases}$$

Given a population consisting of N elements, table counts of a table T = $[X_1^{(r_1)} \times \ldots \times X_G^{(r_G)}] \times Y$, concerning the count variable Y, can, in vector notation, be written as

$$\mathbf{Y}_T = \sum_{k=1}^{N} \mathbf{x}_k^{(\mathbf{r})} y_k \equiv \sum_{k=1}^{N} \mathbf{y}_{k,T} ,$$

where y_k denotes the score of element k on the count variable Y. For frequency counts, y_k equals one, and could equally well be omitted. For count variables such as 'income', y_k may assume any real number. The vector $\mathbf{Y}_T$ consists of P components, corresponding to the $P = P_1 \times \ldots \times P_G$ mutually excluding categories of the crossing $X_1^{(r_1)} \times \ldots \times X_G^{(r_G)}$, that is, they correspond to the P cells in table T.

Often, the population totals for (the components of) $\boldsymbol{Y}_T$ are not know: they must be estimated from survey data. The P components in $\boldsymbol{Y}_T$ can be viewed as a collection of population characteristics that are to be estimated simultaneously. For a sample S of size n drawn from the population, the Horvitz-Thompson estimator for the table counts for table T corresponds to

$$\hat{\mathbf{Y}}_T^{HT} = \sum_{i \in S} d_i \mathbf{x}_i^{(\mathbf{r})} y_i \equiv \sum_{i \in S} d_i \mathbf{y}_{T,i} ,$$

where d_i corresponds to the usual inverse inclusion probability of element i in the sample, see Särndal *et al.* (1992).

3.1.1 Regression estimator

The generalized regression estimator (for short: regression estimator) is an important ingredient in repeated weighting. The regression estimator is designed to increase the accuracy of the estimate of some variable *Y* by using auxiliary variables $X_1, \ldots, X_J$ that are correlated with *Y*, and whose population totals are known, see Cassel et al. (1976). Instead of one specific target variable *Y*, one can also estimate the *P* cell totals $\mathbf{Y}_T$ of some table *T* with the regression estimator, using the known population totals $\mathbf{Y}_{T_M}$ of one of its margins T_M as the auxiliary variables. The regression estimator is then given by

$$\hat{\mathbf{Y}}_T^R = \hat{\mathbf{Y}}_T^{HT} + \mathbf{B}'_{T,T_M}\left(\mathbf{Y}_{T_M} - \hat{\mathbf{Y}}_{T_M}^{HT}\right)$$

where

$$\mathbf{B}_{T,T_M} = \left(\sum_{i\in S} d_i \mathbf{y}_{T_M,i}\mathbf{y}'_{T_M,i}\right)^{-1}\left(\sum_{i\in S} d_i \mathbf{y}_{T_M,i}\mathbf{y}'_{T,i}\right)$$

denotes the matrix of regression coefficients, see Särndal *et al.* (1992). Substituting the **B-**matrix in the expression for the regression estimator, it can be rewritten as

$$\begin{aligned}\hat{\mathbf{Y}}_T^R &= \sum_{i\in S} d_i \mathbf{y}_{T,i}\left(1 + \mathbf{y}'_{T_M,i}\left(\sum_{i\in S} d_i \mathbf{y}_{T_M,i}\mathbf{y}'_{T_M,i}\right)^{-1}\left(\mathbf{Y}_{T_M} - \hat{\mathbf{Y}}_{T_M}^{HT}\right)\right)\\ &\equiv \sum_{i\in S} d_i g_i \mathbf{y}_{T,i}\\ &= \sum_{i\in S} w_i \mathbf{y}_{T,i}\,.\end{aligned}$$

The regression weights w_i are given by the original design weights d_i, times a correction factor g_i. The correction factor g_i is generally close to unity. The important point with the regression estimator is that when the population totals of the auxiliary variables themselves, that is, the cells in the margin T_M, are estimated with the weights w_i, the known population totals $\mathbf{Y}_{T_M}$ are recovered, that is,

$$\hat{\mathbf{Y}}_{T_M}^R = \sum_{i\in S} d_i \mathbf{y}_{T_m,i}\left(1 + \mathbf{y}'_{T_M,i}\left(\sum_{i\in S} d_i \mathbf{y}_{T_M,i}\mathbf{y}'_{T_M,i}\right)^{-1}\left(\mathbf{Y}_{T_M} - \hat{\mathbf{Y}}_{T_M}^{HT}\right)\right) = \mathbf{Y}_{T_M}\,.$$

In other words, the regression weights w_i are *calibrated* on the known population totals $\mathbf{Y}_{T_M}$, see Deville (1988) and Deville and Särndal (1992). This property of the regression estimator is used in repeated weighting. Roughly speaking, repeated weighting works as follows. Some table needs to be estimated, but one or more margins of this table may already be known from register counts or may be estimated from different data sources. By using the regression estimator, the table estimate is calibrated on these margins. Thus by reweighting the survey weights, the table will be estimated consistently with its margins.

3.2 Preliminaries

Having introduced some basic notational matters, we can start with a description of the method of repeated weighting. It is assumed that a target population, variables, and a table set that needs to be estimated, are specified in advance. The process of repeated weighting consists roughly speaking of three steps, see Renssen *et al*. (2001). In the first step, relevant micro data on the target population are extracted from the SSD. From the extracted data, rectangular, completely filled, data blocks are constructed. In the second step, weights are assigned to each rectangular data block. In the final step, numerically consistent table estimates are generated from the rectangular data blocks. In Section 3.3, we focus on this third step. In this section, we describe the first two steps in more detail. They are necessary in order to carry out the third step, which is the actual estimation process.

3.2.1 Data blocks

As explained earlier, the SSD contains data from administrative registers and surveys. Given the knowledge on the target population, and the table set that needs to be estimated, relevant micro data must be extracted from the SSD. In Figure 1, an example of extracted micro data is given. The target population in this example is ‘jobs in The Netherlands’. A register, containing the complete population of jobs, is extracted from the SSD. This register contains several, but not all variables that are needed for the table set of interest. The remaining variables are obtained from survey data, in particular, from the Employment and Wages Survey (EWS) and the Labor Force Survey (LFS). The data from both surveys are linked to the register, and to each other. The resulting micro database is shown in Figure 1. A large part of the micro database is empty: only the shaded areas contain observations. As can be seen from the figure, the two surveys have a small overlap.

Next, rectangular, completely filled, data blocks are constructed from the selected micro data. Each block consists of all records that have a certain set of variables in common, and for each record all these common variables are included. For the present example, four rectangular data blocks can be constructed as shown in Figure 1. The first data block corresponds to the administrative registration, and contains all register variables. The second data block contains all elements in the largest of the two surveys (the EWS), and all register and EWS variables. The third data block contains all elements in the smallest of the two surveys (the LFS), and all register and LFS variables. The fourth rectangular data block contains all elements that are present in both the EWS and the LFS, and contains all register, EWS, and LFS variables.

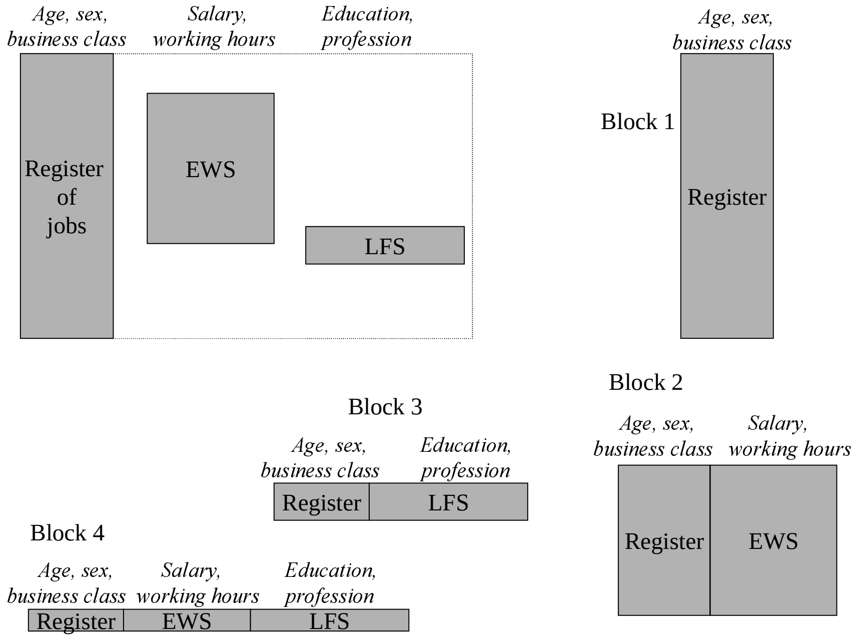

Figure 1. Example of micro data from the SSD, and the construction of rectangular data blocks.

In principle, one can also imagine a situation in which two or more surveys contain some common variables (not shown in Figure 1, but see Figure 3 in Section 5.1 for an example). In that case, a data block consisting of all elements in the union of these surveys can be constructed. In general, data blocks may therefore consist of register data, data from the union of two (or more) surveys, data from a single survey, or data from the cross section of two (or more) surveys. The table set of interest will be estimated from these rectangular data blocks. For each table, one first determines the most suitable rectangular data block from which this table can be estimated. In general, the most suitable data block will correspond to the largest data block (in number of records) that contains all relevant variables simultaneously. The more data an estimate will be based upon, the more accurate it will be. Of course, considerations regarding the quality of certain survey data may, in practice, be a reason to deviate from this natural choice. However, for simplicity we assume in the remainder of this paper that the most suitable data block always is the largest block that contains all variables in the table simultaneously. So, again referring to the example in Figure 1, the frequency table 'age class × working hours × education' is estimated from data block 4, but its margin (lower-dimensional aggregate) 'age class' is counted from the register, the margin 'age class × working hours' is estimated from data block 2, and the margin 'age class × education' is obtained from data block 3.

3.2.2 Weights

Before tables can be estimated at all, a (fixed) set of weights should be assigned to each of the rectangular data blocks to inflate from the samples to the population. This is done in the second step in the process of repeated weighting. These weights should adjust for sampling error and nonresponse, and preferably already meet some of the consistency requirements.

For a register data block, the weights are identical to unity, since each record in the (complete) register only counts for itself.[4] For data blocks consisting of survey data, the weights are derived from the inclusion probability π_i of elements in these surveys. In the remainder of this section, we will explain how these block weights are determined.

Suppose a survey S has initial survey weights (design weights) $d_i^{(S)} = 1/\pi_i$ that correct for unequal sampling probabilities. These initial survey weights may be reweighted to known population totals from the population register to correct for selective survey nonresponse and to reduce the variance of estimates. This requires a careful selection of the weighting scheme: in choosing auxiliary variables, the three basic principles that they should explain the response probabilities, explain the variation of the main study variables, and identify the most important domains, should be satisfied as much as possible, see, for instance, Lundström and Särndal (1999, 2002). Using the usual regression estimator, the adjusted survey weights $w_i^{(S)}$ are given by, see Särndal *et al.* (1992)

$$w_i^{(S)} = d_i^{(S)}(1 + \mathbf{x}_i'\mathbf{G}(\mathbf{X} - \hat{\mathbf{X}}^{HT(S)})) \equiv d_i^{(S)} g_i^{(S)},$$

where the g weight $g_i^{(S)}$ denotes the correction factor of the original design weights, $\mathbf{x}_\mathrm{i}$ denotes the vector of auxiliary variables, the vector $\mathbf{X}$ contains the known population totals for the auxiliary variables $\mathbf{x}_\mathrm{i}$, and

$$\hat{\mathbf{X}}^{HT(S)} = \sum_{i \in S} d_i^{(S)} \mathbf{x}_i$$

gives the Horvitz-Tompson estimator of the population totals for the variables $\mathbf{x}_\mathrm{i}$, using the initial survey weights. The matrix $\mathbf{G}$ is the (generalized) inverse of

$$\sum_{i \in S} d_i^{(S)} \mathbf{x}_i \mathbf{x}_i' \, .$$

Next, we derive (initial) block weights $d_i^{(B)}$ for data blocks B consisting of survey data. Suppose we have two surveys S_1 and S_2, with sample sizes of n_1 and n_2, and initial survey weights $d_i^{(S1)} = 1/\pi_{1i}$ and $d_i^{(S2)} = 1/\pi_{2i}$, respectively. The surveys may each be corrected for nonresponse, resulting in the adjusted weights $w_i^{(S1)}$ and $w_i^{(S2)}$. Given these two surveys, three situations can be distinguished:

[4] Note that, for some table sets, one might be interested in the average over some time period, instead of the situation at a certain reference date. In that case, the register block contains the records of all elements that were a member of the population during (a fraction of) this time period. The weight of a record is then given by the fraction of the time period that the record was an element of the population, instead of unity.

1. Data block B consists of a single survey S_k. In this case, the initial weights $d_i^{(B)}$ of the elements in the data block are simply given by the adjusted survey weights $w_i^{(Sk)}$.

2. Data block B consists of the union $S_1 \cup S_2$ of the two surveys. Assuming that both surveys have an empty cross section, the initial weights of the elements in the block can be derived in several ways. We choose[5]

$$d_i^{(B)} = \begin{cases} \lambda^{(S1)} w_i^{(S1)} \text{ for } i \in S_1 \\ \lambda^{(S2)} w_i^{(S2)} \text{ for } i \in S_2, \end{cases}$$

   where the relative accuracy of the surveys is taken into account by the factors $\lambda^{(S1)}$ and $\lambda^{(S2)} = 1 - \lambda^{(S1)}$. These λ-factors should reflect the quality of estimates from the surveys. A simple choice is to take $\lambda^{(S)}$ proportional to the size of S, that is, $\lambda^{(S1)} = n_1/(n_1 + n_2)$, or proportional to the effective size of S, that is, $\lambda^{(S1)} = n_{1\text{eff}}/(n_{1\text{eff}} + n_{2\text{eff}})$, where

$$n_{\text{eff}} = \frac{n}{1+L},$$
$$1+L = \frac{1}{n}\sum_{i=1}^{n} w_i^2 \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^2,$$

   is a measure for the variance-increasing effect of unequal sampling probabilities, see Kish (1992). For surveys with equal sampling probabilities, the effective size reduces to the real size. Note that the case of a data block consisting of one single survey can be viewed as a special case of the union of surveys, with $\lambda = 1$. Note also that the λ-factors given here are constant for all records in each survey S. In practice, there may be reasons to deviate from this simple choice. For instance, surveys with unequal sampling frames may require record-dependent λ-factors.

3. Data block B consists of elements in the cross section $S_1 \cap S_2$ of the two surveys. In this case, the initial block weights may be given by the product of the (adjusted) survey weights, that is, $d_i^{(B)} = w_i^{(S1)} w_i^{(S2)}$, which corresponds to the inverse probability that an element i in the population is sampled in both surveys simultaneously.

The initial block weights may not yet add up to known population totals from the register. Furthermore, these initial block weights can be reweighted to known population characteristics to correct for selective nonresponse and to reduce the variance of estimates. This again requires a careful selection of the *overall weighting scheme*. The resulting final block weights are, analogous to the adjusted survey weights, given by

[5] When the surveys S_1 and S_2 have a small overlap, the weights of the records that are present in both surveys are given by $d_i^{(B)} = \lambda^{(S1)} w_i^{(S1)} + \lambda^{(S2)} w_i^{(S2)}$ for $i \in S_1 \cap S_2$. The records that are only present in S_1 and S_2 get weights $d_i^{(B)} = \lambda^{(S1)} w_i^{(S1)}$ and $d_i^{(B)} = \lambda^{(S2)} w_i^{(S2)}$, respectively.

$$w_i^{(B)} = d_i^{(B)}(1 + \mathbf{x}_i'\mathbf{G}(\hat{\mathbf{X}} - \hat{\mathbf{X}}^{(B)})) \equiv d_i^{(B)} g_i^{(B)},$$

where the g weight $g_i^{(B)}$ gives the correction factor to the original block weights, $\mathbf{x}_\mathrm{i}$ again denotes the vector of auxiliary variables used in the weighting scheme, and

$$\hat{\mathbf{X}}^{(B)} = \sum_{i \in B} d_i^{(B)} \mathbf{x}_i$$

gives the estimated population totals for the variables $\mathbf{x}_\mathrm{i}$, using the initial block weights. The matrix **G** is the (generalized) inverse of

$$\sum_{i \in B} d_i^{(B)} \mathbf{x}_i \mathbf{x}_i' \ .$$

The vector $\hat{\mathbf{X}}$ contains (estimated or counted) population totals for the variables $\mathbf{x}_\mathrm{i}$ derived from larger, that is, more accurate, data blocks than the present data block *B*. The possibility to benchmark on survey variables in addition to register variables, allows not only for a better correction for nonresponse and reduction of variances, but can also be used to achieve some consistency between the estimates from different data blocks at this point. However, due to lack of degrees of freedom, it will in general not be possible to obtain full consistency between all possible estimates from different data blocks with one (fixed) set of weights $w_i^{(B)}$ for each data block *B*. To obtain numerically consistent estimates, it will therefore be necessary to apply an additional *reweighting scheme*, as will be explained in Section 3.3.

3.2.3 Prerequisites

Before turning our attention to the estimation strategy in repeated weighting, we focus on the requirements on the datasets that are included in the SSD. First, these datasets must be complete and edited on the micro level. Item nonresponse should, for instance, be imputed (if nothing else, then a category ‘Unknown’ can be used), or the record must be considered as complete nonresponse. Missing values and inconsistencies at the micro level in general cause unacceptable inconsistencies in the estimates. The records in the registrations and surveys should furthermore be equipped with a unique key, so that records can indeed be linked at the micro level. It is assumed that it is not only technically possible, but also legally allowed to link the register and survey data to each other. Protection of privacy is for some countries a reason to impose legal restrictions to matching data sets. However, in The Netherlands, Statistics Netherlands is under strict disclosure conditions allowed by law to link data sets, see e.g. Van der Laan (2000).

On the survey level, there are also a few restrictions. If several surveys are used to estimate a certain table set, the sampling frames of these surveys should preferably be equivalent, and identical to the frame of the population in the register. Moreover, if two or more surveys have variables in common, these variables can only be considered equivalent if they truly measure the same statistical concept. This requires that the surveys are harmonized in the sense that the routing, the question formulation, and the answering categories for this particular concept in both surveys are equivalent. It may lead to major biases in the estimates if variables are considered equivalent despite the fact that these requirements are not met. Finally, for the method of repeated weighting, a necessary requirement to classification variables is that they should be hierarchical if a variable consists of more than one classification level.

3.3 Consistent estimates

After isolating suitable micro data from the SSD, constructing rectangular data blocks from this micro data, and assigning weights $w_i^{(B)}$ to the records in these data blocks, one can finally start to estimate a given table set. As said before, we wish to estimate this table set in such a way that all tables will be numerically consistent with each other. Using the block weights $w_i^{(B)}$, this is not automatically guaranteed, since estimates are not necessarily made from the same data block. In particular, the combination of variables in a cross tabulation determines from which data block the table is estimated. Consequently, cross tabulations having one or more variables in common and differing in the other variables may be estimated from different data blocks, *i.e.*, with different records and different weights $w_i^{(B)}$. The margins of these cross tabulations with respect to the variables they have in common will therefore, in general, differ. This leads to inconsistent estimates. Referring again to the example in Figure 1, the aggregate 'age class × working hours' of the frequency table 'age class × working hours × education' estimated from data block 4, will in general not coincide with the (more accurate) estimate of 'age class × working hours' from data block 2. With the method of repeated weighting, such inconsistencies are prevented.

3.3.1 Frequency tables

Suppose that a table set consisting of K tables $\{T_1, T_2, T_3, \ldots, T_K\}$ must be estimated. Assume for now that all tables relate to frequency counts. The tables can, using the notation of Section 3.1, alternatively be represented by the vectors $\mathbf{r}_1$, $\mathbf{r}_2$, …, $\mathbf{r}_K$. Note that with this notation, any frequency table T_m (not necessarily a member of the above table set) is a margin of T_k if and only if $\mathbf{r}_{m,k} = \min(\mathbf{r}_m,\mathbf{r}_k) = \mathbf{r}_m$. To estimate the table set $\{T_1, T_2, T_3, \ldots, T_K\}$ fully consistent, the following procedure is adopted, see Kroese and Renssen (2000), and Renssen *et al.* (2001):

- Every cross tabulation $T_k \in \{T_1, T_2, T_3, \ldots, T_K\}$ will be based on the most suitable data block (the data block in which the statistician has most confidence, that is, the largest data block in general), in which all relevant variables occur simultaneously.

- If a cross tabulation T_k has a margin T_m that can be estimated from a larger data block, this margin T_m should be added to the table set (if not already present), and estimated before T_k is estimated. The margin T_m is estimated more accurately, and can serve as auxiliary information when estimating table T_k.

- All crosstabulations T_k that can be estimated consistently with the block weights $w_i^{(B)}$ of the most suitable data block, should be estimated before tables that cannot be estimated consistently with these block weights. Note that a table T_k cannot be estimated consistently using the block weights $w_i^{(B)}$ when T_k has a margin T_m that can be estimated from a larger data block whereas this margin is not included in the weighting model of the block from which T_k is to be estimated.

- If a crosstabulation T_k cannot be estimated consistently with the block weights of the most suitable data block, but a margin T_m of this table can, this margin should be added to the table set (if not already present), and estimated before T_k is estimated.

- Only if a crosstabulation T_k cannot be estimated consistently with the block weights of the most suitable data block, the table must be estimated by repeated weighting, that is, the block weights $w_i^{(B)}$ will be adjusted by some additional reweighting scheme, taking into account all tables T_1, …, T_{k-1} and extra margins T_m that are already estimated according to the rules under bullets 1, 2, 3 and 4.

Thus, each table in the table set is estimated using as many data as possible. If a table has a margin that can be estimated with more data, this margin is added to the table set, and estimated before the current table. The margin is thus estimated more accurately, and can serve as auxiliary information when estimating the current table. Equivalently, when a table has margins that can be estimated consistently, without reweighting, with the block weights $w_i^{(B)}$, these margins are added to the table set, and also estimated first. A table T_k can be estimated consistently with the block weights $w_i^{(B)}$ when all its (nontrivial) margins T_m that can be obtained from larger data blocks, are present in the (margins of the) terms of the overall weighting scheme of data block B.

Only when a table cannot be directly estimated consistently with the block weights $w_i^{(B)}$ – which are optimally designed for nonresponse correction and variance reduction – these weights are adjusted slightly, but only to estimate the present table. The weights are reweighted such that consistency is achieved with all other tables having variables in common with the table under consideration. Reweighting is in the first place performed for cosmetic purposes, and should therefore not lead to large changes in the estimates. For reweighting, the calibration properties of the regression estimator are used, as will be explained below.

To describe the reweighting strategy, we assume that all tables in table set $\{T_1, T_2, T_3, \ldots, T_K\}$ must be reweighted. One starts by estimating the first table T_1. This is usually a register count, or otherwise can be estimated directly with the block weights, since there is nothing yet with which this table should be consistent. Subsequently, the second table T_2 is estimated, but this table should be consistent with the estimate of T_1. This is achieved by determining the common margin $T_{1,2}$ of the two tables. This common margin is specified by $\mathbf{r}_{1,2} = \min(\mathbf{r}_1,\mathbf{r}_2)$. The margin $T_{1,2}$ can be obtained by aggregating T_1. T_2 is then estimated with repeated weighting, that is, with the regression estimator, using the (possibly estimated) population totals in

$T_{1,2}$ as auxiliary information, and the final block weights $w_i^{(B)}$ of the most suitable data block as initial weights. Then one continues to the next table in the list. Generally, table T_k is estimated by determining the common margins with tables T_1, …, T_{k-1}, that is, $T_{1,\mathrm{k}}$, …, $T_{k-1,\mathrm{k}}$, and subsequently calibrating on the (possibly estimated) population totals of the reweighting scheme $T_{1,\mathrm{k}}$ + … + $T_{k-1,\mathrm{k}}$. Table T_k will then obviously be numerically consistent with all earlier estimated tables T_1, …, T_{k-1} . One continues until all tables in the table set are estimated. The resulting repeated weighting estimates (RW-estimates) will be numerically consistent.

The reweighting scheme $T_{1,\mathrm{k}}$ + … + $T_{k-1,\mathrm{k}}$ is called the *minimal weighting scheme*, since it contains the minimum of population totals that must be used as auxiliary information to achieve consistency. The weighting scheme may contain ‘redundant’ terms. A term T_p is redundant if the weighting scheme contains another term T_q such that $T_{p,q} = T_p$. Redundant terms can be omitted from the weighting scheme; they do not contain any new auxiliary information that is not already present in the remaining terms. Moreover, when all terms in the minimal weighting scheme of a certain table are redundant with the terms in the overall weighting scheme of the data block, reweighting is not necessary.[6] The table can immediately be estimated using the block weights $w_i^{(B)}$. Note that the minimal weighting scheme can in principle be extended with additional terms. This allows for the use of additional auxiliary information to reduce variances of the estimates.

In the mathematical terms of Section 3.1, the repeated weighting estimator for a table T that must be estimated from data block B with reweighting scheme $T_1 + T_2$ + … + T_M, is given by the regression estimator

$$\begin{aligned}\hat{\mathbf{Y}}_T^{RW} &= \sum_{i\in B} w_i^{(B)}\mathbf{y}_{T,i} + \mathbf{B}'_{T,T_1+\ldots+T_M}\left(\hat{\mathbf{Y}}_{T_1+\ldots+T_M}^{RW} - \sum_{i\in B} w_i^{(B)}\mathbf{y}_{T_1+\ldots+T_M,i}\right)\\ &= \sum_{i\in B} w_i^{(B)}\mathbf{y}_{T,i} + \mathbf{B}'_{T,T_1+\ldots+T_M,T_1}\left(\hat{\mathbf{Y}}_{T_1}^{RW} - \sum_{i\in B} w_i^{(B)}\mathbf{y}_{T_1,i}\right) + \ldots +\\ &\qquad + \mathbf{B}'_{T,T_1+\ldots+T_M,T_M}\left(\hat{\mathbf{Y}}_{T_M}^{RW} - \sum_{i\in B} w_i^{(B)}\mathbf{y}_{T_M,i}\right),\end{aligned}$$

where the final block weights $w_i^{(B)}$ are used as initial weights, and

$$\begin{aligned}\hat{\mathbf{Y}}_{T_1+\ldots+T_M}^{RW} &= (\hat{\mathbf{Y}}'^{RW}_{T_1},\ldots,\hat{\mathbf{Y}}'^{RW}_{T_M})',\\ \mathbf{y}_{T_1+\ldots+T_M,i} &= (\mathbf{y}'_{T_1,i},\ldots,\mathbf{y}'_{T_M,i})',\end{aligned}$$

and the regression coefficient matrix $\mathbf{B}$ is given by

[6] Here we assume that all margins of the table that can be estimated from larger data sets, are estimated first.

$$\begin{aligned}\mathbf{B}_{T,T_1+\ldots+T_M} &= \left(\sum_{i\in B} w_i^{(B)} \mathbf{y}_{T_1+\ldots+T_M,i} \mathbf{y}'_{T_1+\ldots+T_M,i}\right)^{-1} \left(\sum_{i\in B} w_i^{(B)} \mathbf{y}_{T_1+\ldots+T_M,i} \mathbf{y}'_{T,i}\right) \\ &\equiv \begin{pmatrix} \mathbf{B}_{T,T_1+\ldots+T_M,T_1} \\ \vdots \\ \mathbf{B}_{T,T_1+\ldots+T_M,T_M} \end{pmatrix}.\end{aligned}$$

Note that the population totals of the auxiliary variables in the reweighting scheme may be counted from the register, or estimated by repeated weighting themselves. Since the regression estimator is only approximately unbiased, one might think that this repeated use of the regression estimator may lead to a bias in the RW-estimator. In the simulations of Section 5, we will come back to this.

3.3.2 Simple example

As a simple example of the estimation strategy, consider again Figure 1. Assuming that ‘salary’ is a quantification variable, we are dealing with six classification variables. Suppose that the number of classification levels of each of these six classification variables is as indicated in Table 1.

Table 1. Overview of classification variables and the number of classification levels for each variable.

Classification variable	***Number of levels***
Age class	1
Sex	1
Business class	4
Working hours per week	1
Educational level	2
Professional level	3

Assume subsequently that the block weights $w_i^{(B)}$ of each data block that contains survey data, are calibrated on the overall weighting schemes shown in Table 2. The superscripts in this table refer to the level of classification. The last column in this table expresses the terms in the overall weighting scheme of each data block in terms of a vector **r**, as introduced in Section 3.1.

Table 2. Overview of the overall weighting schemes for the data blocks containing survey data.

Block	***Overall weighting scheme***	***Vector* r**
2	Age$^{(1)}$ × sex$^{(1)}$	(1,1,0,0,0,0)
3	Age$^{(1)}$ × sex$^{(1)}$	(1,1,0,0,0,0)
4	Age$^{(1)}$ × sex$^{(1)}$ + working hours$^{(1)}$ + education$^{(1)}$	(1,1,0,0,0,0) + (0,0,0,1,0,0) + (0,0,0,0,1,0)

Now suppose that the table set $\{T_1, \ldots, T_6\}$ given in Table 3 must be estimated. For each target table, the most suitable data block is given, as well as the vector **r** that alternatively defines each table. Note that for each $T_k \in \{T_1, \ldots, T_6\}$, the margins that can be estimated from a larger data block are also present in the table set $\{T_1, \ldots, T_6\}$.

Table 3. Overview of target frequency distributions and their reweighting scheme.

Table	***Description***	***Block***	***Vector* r**	***Reweighting scheme***
T_1	age$^{(1)}$ × sex$^{(1)}$ × business class$^{(4)}$	1	(1,1,4,0,0,0)	Register count
T_2	age$^{(1)}$ × working hours$^{(1)}$	2	(1,0,0,1,0,0)	Consistent with block weights
T_3	age$^{(1)}$ × sex$^{(1)}$ × education$^{(2)}$	3	(1,1,0,0,2,0)	Consistent with block weights
T_4	business class$^{(2)}$ × education$^{(1)}$	3	(0,0,2,0,1,0)	business class$^{(2)}$ + education$^{(1)}$
T_5	working hours$^{(1)}$ × education$^{(1)}$	4	(0,0,0,1,1,0)	Consistent with block weights
T_6	age$^{(1)}$ × working hours$^{(1)}$ × education$^{(1)}$	4	(1,0,0,1,1,0)	age$^{(1)}$ × working hours$^{(1)}$ + age$^{(1)}$ × education$^{(1)}$ + working hours$^{(1)}$ × education$^{(1)}$

Table T_1 can simply be counted from the register. The reweighting scheme of table T_2 is, in principle, given by the common margin $T_{1,2}$ of this table with T_1. This margin is defined by $\mathbf{r}_{1,2} = \min(\mathbf{r}_1,\mathbf{r}_2) = (1,0,0,0,0,0)$, which is already contained in the overall weighting scheme of data block 2. Table T_2 can thus be estimated consistently using the block weights. A similar situation arises for T_3, and T_5. The reweighting schemes of both tables are contained in the overall weighting scheme of the data blocks from which these tables should be estimated, and reweighting is thus not necessary.

A different situation arises for table T_4. This table has a reweighting scheme given by $T_{1,4} + T_{2,4} + T_{3,4}$. After removing all redundancy, we are left with the margins $\mathbf{r}_{1,4} = \min(\mathbf{r}_1,\mathbf{r}_4) = (0,0,2,0,0,0)$, or, equivalently, business class$^{(2)}$, and $\mathbf{r}_{3,4} = \min(\mathbf{r}_3,\mathbf{r}_4) = (0,0,0,0,1,0)$, which corresponds to education$^{(1)}$. These margins are not all contained in the overall weighting scheme of data block 3, which makes reweighting necessary to achieve consistency. The reweighting scheme of T_6 can be derived analogously.

3.3.3 General quantification variables

We have shown how a set of frequency tables can be estimated consistently with repeated weighting. Often however, one is also interested in tables related to other quantification variables such as ‘total income’. In principle, the same estimation strategy can be used. For each table, the minimum reweighting scheme is determined (the terms in this weighting scheme now relate to the quantification variable of the table), and the table is then estimated using the population totals in the reweighting scheme as auxiliary information.

However, a problem may arise when one is also interested in tables related to *average incomes*. To calculate an average income, the total income in some class must be divided by the frequency count of elements in this class. The frequency table may, however, be obtained with a reweighting scheme that differs from the reweighting scheme of the total income-table. Also, the data blocks from which both tables are estimated might be different entirely. As a consequence, it may happen that the estimated average income in some class is lower than the estimated average income in another class, whereas in the micro data, the incomes of all records in the first class are higher than the incomes of records in the second class. Especially when one of the dimension variables in the table relates to income classes, this may lead to impossible estimates.

This problem is prevented by adding the underlying frequency table to the minimum weighting scheme of any table relating to a general quantification variable. Thus, when estimating ‘total income by age class’, the frequency counts for ‘age class’ are always used as auxiliary information in the estimation process. This frequency count is added to the reweighting scheme of the ‘total income by age class’-table. In this way, the average income in a certain class obviously will never be lower than the income of the element with the lowest income in that particular class, and never be higher than the highest income in that class. In addition, by adding the frequency table to the minimum weighting scheme, the efficiency of the estimator is also increased.

A special case arises when a classification variable is related to a certain quantification variable, for instance, when the classification variable ‘income in classes’ is related to the quantification variable ‘income’. To ensure plausible table estimates for tables related to income, the underlying frequency table should be extended with the extra dimension ‘income in classes’, and added to the reweighting scheme of the table. So, for instance, when estimating ‘total income by age class’, the frequency counts for ‘age class × income class’ must be used as auxiliary information in the estimation process.

A disadvantage is that the extended frequency counts in the reweighting scheme of a table may take a large claim on the degrees of freedom of the system, especially when the classification variable that relates to a certain quantification variable, has many categories. In the spirit of the estimation strategy proposed in Renssen (1998), by defining new quantification variables that are based on the classification variable related to a certain quantification variable, a more economical estimation strategy that still ensures consistency, may be obtained. This still needs to be worked out.

3.3.4 Splitting-up procedure

In Section 3.3.1, we have seen that the reweighting scheme of a certain table $T_k \in \{T_1 ,\ldots, T_K\}$ depends on the tables $T_1, \ldots, T_{k-1}$ that have already been estimated. Consequently, with repeated weighting, the actual table estimates depend on the order in which the tables are estimated. Although differences in general will be small, this might be considered as an undesired side effect of the method.

Fortunately, this ‘order problem’ can be prevented by fixing the order of all estimates. One way to do so, is by using the so-called ‘splitting-up procedure’. In the splitting-up procedure, all lower dimensional aggregates of a table are estimated first. If, for instance, the three-way frequency table ‘sex × working hours × education’ is to be estimated, see Figure 1, first the one-way tables ‘sex’, ‘working hours’, and ‘education’ are estimated. Subsequently, the two-way tables ‘sex × working hours’, ‘sex × education’, and ‘working hours × education’ are estimated, taking the one-way tables into account. Finally, the target table is estimated, taking the two-way tables into account. Since all tables are estimated from the most suitable data block, this will solve the order problem.

In general, the margins of a table T can be derived quickly using the notation of Section 3.1. Any table T can be specified by a vector $\mathbf{r} = (r_1, r_2, \ldots, r_G)'$. Suppose that T is an M-way table, that is, M out of G components of $\mathbf{r}$ satisfy $r_g > 0$. This table has M ‘most-detailed’ margins. A most detailed margin of T is obtained by replacing the level r_g of one of the M classification variables with $r_g > 0$ by its next lower classification level $r_g - 1$. Obviously, when $r_g > 1$, this most-detailed margin is again an M-way table. When $r_g = 1$, the resulting margin will be an (M–1)-way table. In the splitting-up procedure, any table is estimated by reweighting to the (estimated or counted) population totals of its most-detailed margins. The margins themselves are estimated in the same way, that is, by reweighting to the (estimated or counted) population totals of their most-detailed margins.

Note that the use of the splitting-up procedure is optional. The advantage is that all tables are estimated in a well-defined order, which makes the estimation process reproducible. An additional advantage is that for a fixed target population, the splitting-up procedure can also ensure consistency *between* table sets. A requirement is that in the estimation process of these table sets for coinciding variables the same micro data and final block weights are used. Despite these clear advantages, the splitting-up procedure sometimes leads to estimation problems. We mention these limitations in Section 3.5. Also, one might think that more calibrations on estimated tables, lead to an increase in the variance of estimates. We return to this issue in Section 5.

3.4 Variance estimation

In the previous section, we have explained the estimation strategy in repeated weighting. In this section, we derive approximate variance formulas for the repeated weighting estimator (RW-estimator). These variance formulas will only be valid when certain simplifying conditions, such as small sampling fractions, constant λ-factors, and non-overlapping surveys, are met. The formulas are therefore, in principle, only applicable in the relatively simple situation of one or more surveys linked to a register, where the surveys have no record overlap and have equal sampling frames. In Sections 3.4.1 to 3.4.4, we focus on the simple case of non-overlapping surveys. In Section 3.4.5, we show that with some extra approximations and operations, the approximate variance formulas can, after all, be applied in the case of overlapping surveys as well.

3.4.1 Assumptions and basic idea

For now, we assume that we are dealing with non-overlapping surveys only. All data blocks that are generated from the micro data contain data from one single complete survey, or from the union of two or more complete surveys, see Section 3.2.1. We assume that the weight of a record originating from a certain survey S in a data block is given by the adjusted survey weight $w_i^{(S)}$ times a survey-dependent, *but otherwise constant* factor $\lambda^{(S)}$. This requires that the surveys have equal sampling frames. For blocks consisting of only one survey, $\lambda = 1$.

The basic idea in the derivation of variance formulas, is that the RW-estimator is a regression estimator, in which the population totals of the auxiliary variables are estimated themselves, again via the regression estimator, and with counted or estimated population totals for the auxiliary variables, see Knottnerus (2003). Each RW-estimator consists of a whole ‘tree structure’ of regression estimators that all contribute to a single RW-estimator. The knots in the tree correspond to regression estimators, and the leaves (endpoints) correspond to simple HT-estimators and register counts. The usual Taylor method for deriving approximate variance formulas for the regression estimator (see Särndal *et al*. (1992)), can thus be used to derive (approximate) variance formulas for the RW-estimator. A similar approach is used in Estevao and Särndal (2002) to derive variances in two-phase sampling.

In deriving variance formulas, we assume that the stochastic properties of the regression coefficients **B** in the regression estimator can be neglected. Furthermore, we assume that the surveys are independent, have small sampling fractions, and, as said before, no overlap.

The estimation of the approximate variance of a table T that is estimated with repeated weighting, roughly consists of the following steps:

1. The estimated table T is considered as a vector $\hat{\mathbf{Y}}_T^{RW}$ of RW-estimators. The number of components in this vector is equal to the number of cells in the table T. The estimator $\hat{\mathbf{Y}}_T^{RW}$ is written as a linear combination of Horvitz-Thomson (HT-estimators) at the survey level, plus a constant term.
2. Terms in the linear combination of HT-estimators are grouped by survey. For each survey, we estimate the variance, using the Hansen-Hurwitz (HH) variance formula for surveys with replacement. Here we use the fact that for

small surveys without replacement, the variance of the HT-estimator is approximately equal to the HH-variance.

3. The HH-variances per survey are summed to obtain the variance of the RW-estimator. Here, we assume that the surveys are independent, and the covariances between the surveys disappear. Despite the fact that the surveys are assumed to be non-overlapping, they can be considered independent when their sampling sizes are small compared to the population size.

In the next subsections, the above steps are worked out in more detail.

3.4.2 Decomposition of RW-estimator in HT-estimators

First of all, we will prove that under the assumptions that we are dealing with small, independent, but non-overlapping surveys, and that the stochastic properties of the regression coefficients **B** can be neglected, the RW-estimator of a table *T* can approximately be written as a linear combination of HT-estimators at survey level plus a constant term, that is,

$$\hat{\mathbf{Y}}_T^{RW} = \sum_{k=1}^{K} \sum_{t \in T(S_k)} \mathbf{M}_{S_k,t} \hat{\mathbf{Y}}_t^{HT(S_k)} + \mathbf{c}\,. \tag{1}$$

The first summation runs over all surveys S_k (k = 1, …,K) used in the estimation process. The second summation runs over all tables t that are used as auxiliary information in the RW-estimator, that is, all tables t from which the HT-estimator $\hat{\mathbf{Y}}_t^{HT(S_k)}$, obtained from S_k, contributes to the original RW-estimator.

Each HT-estimator $\hat{\mathbf{Y}}_t^{HT(S_k)}$ in the summation is preceded by a constant matrix $\mathbf{M}_{S_k,t}$, that indicates how the corresponding HT-estimator of table t from survey S_k contributes to the RW-estimator of table T. This matrix has as many rows as the number of cells in T, and as many columns as the number of cells in t. $\mathbf{M}_{S_k,t}$ roughly consists of products of (parts of) regression coefficient matrices **B**, and a λ-factor. In Section 3.4.3, we give an small example in which these matrices $\mathbf{M}_{S_k,t}$ are determined.

The vector **c** has as many components as the table T, and consists of a linear combination of register counts, multiplied by regression coefficient matrices and possibly a λ-factor. Since the stochastic properties of these regression matrices are neglected, the vector **c** can be considered as constant. For the variance estimation, the vector **c** is not relevant.

The proof that each RW-estimator can be written as a linear combination of HT-estimators is given in a few steps. We assume that we have P data blocks B_p (p = 1, …, P), that are constructed from K surveys S_k (k = 1, …, K) linked to register data. The data blocks are ordered in such a way that the block weights $w_i^{(Bp)}$ of data block B_p are only calibrated on RW-estimators from data blocks B_1, …, B_{p-1}, and on register counts. This will be called the 'hierarchy of data blocks'. In practice, block

B_1 will often correspond to the largest data block (in number of records), and data block B_P to the smallest. To proceed, we first of all introduce two definitions.

Definition 1. A pseudo RW-estimator from data block B is an RW-estimator for which reweighting is not necessary, because the corresponding estimator using the block weights $w_i^{(B)}$ is already numerically consistent with all register counts, and all possible RW-estimators from data blocks which lie higher (earlier) in the hierarchy of blocks.

Definition 2. A true RW-estimator from data block B is an RW-estimator for which reweighting ís necessary, because the corresponding estimator using the block weights $w_i^{(B)}$ is not numerically consistent with all register counts and all possible RW-estimators from data blocks which lie higher (earlier) in the hierarchy of blocks.

Corollary. The reweighting scheme of the first *true* RW-estimator from data block B contains as terms at least one or more register counts and/or RW-estimators from data blocks that lie higher (earlier) in the hierarchy of blocks, and possibly pseudo RW-estimators from data block B.

The corollary is an intrinsic property of the procedure of repeated weighting: It follows automatically from the splitting-up procedure, and if this procedure is not applied, from the rules given in Section 3.3.1. In the simple example of Section 3.3.2, and in particular in Table 3, we also have seen this. Using the above definitions, we will prove that all RW-estimators can be written as a linear combination of HT-estimators.

Theorem 1. All block-estimators which use the initial block weights $d_i^{(B)}$ as weights, are approximately linear combinations of HT-estimators at survey level.

Proof. All survey-estimators with the adjusted survey weights $w_i^{(S)}$ as weights are true regression estimators,[7] and consequently are (approximately, that is, neglecting the stochasticity of the regression coefficients) linear combinations of HT-estimators at survey level. Since the surveys have no record overlap, every data block B consists of the union of one or more surveys. The initial block weights $d_i^{(B)}$ are given by the corresponding adjusted survey weights $w_i^{(S)}$, multiplied by a constant, survey dependent factor $\lambda^{(S)}$. Consequently, all block-estimators which use the initial block

[7] A true regression estimator uses only variables of which the population totals are known (register counts) as auxiliary information.

weights $d_i^{(B)}$ as weights, are linear combinations of true regression estimators at survey level, and therefore approximately linear combinations of HT-estimators. ■

Theorem 2a. Let data block B_1, apart from the register block, be the first data block in the hierarchy B_1, …, B_P, from which estimates can be obtained. Every block-estimator which uses the final block weights $w_i^{(B1)}$ as weights, can be written as a linear combination of HT-estimators.

Proof. If the overall weighting scheme of a data block B only contains register variables, and no estimates from other data blocks, all block-estimators with $w_i^{(B)}$ as weight, are approximately linear combinations of register counts and block-estimators which use the initial block weights $d_i^{(B)}$ as weights. Using Theorem 1, these estimators can therefore be written as a linear combination of HT-estimators. Data block B_1 satisfies the condition, since it is the first data block in the hierarchy. Every block-estimator from B_1 which uses the block weights $w_i^{(B1)}$ as weights, can therefore be written as a linear combination of HT-estimators. ■

Theorem 2b. Let B_1, apart from the register block, be the first data block in the hierarchy B_1, …, B_P, from which estimates can be obtained. Every RW-estimator from B_1 can be written as a linear combination of HT-estimators.

Proof. We distinguish between pseudo and true RW-estimators. Every pseudo RW-estimator from B_1 is, in fact, a block-estimator from B_1 which uses the block weights $w_i^{(B1)}$ as weights, and can, according to Theorem 2a, be written as a linear combination of HT-estimators.

To prove that also every true RW-estimator from B_1 can be written as a linear combination of HT-estimators, we first note that the *first* true RW-estimator from B_1 always has a reweighting scheme that only contains register counts and pseudo RW-estimators from B_1. The first true RW-estimator can thus approximately be written as a linear combination of register counts and block-estimators with $w_i^{(B1)}$ as weights, and, using Theorem 2a, approximately as a linear combination of HT-estimators. It then follows via induction that also the second, third, fourth etc. true RW-estimator from B_1 approximately can be written as a linear combination of HT-estimators, as the minimum reweighting scheme of the *j*-th true RW-estimator only can contain register counts, pseudo, and the first to the (j–1)-th true RW-estimators from B_1. ■

Theorem 3. Every RW-estimator can approximately be written as a linear combination of HT-estimators.

Proof. The proof follows from induction. Let B_p (apart from the register block) be the p-th data block in the hierarchy B_1, …, B_P, from which RW-estimates can be obtained. According to Theorem 2b, the theorem is valid for RW-estimators from B_1. Assume that the theorem is also valid for all RW-estimators from B_1, …., B_p. It suffices to prove the theorem for RW-estimators from B_{p+1}.

Due to the hierarchy in the data blocks, every block-estimator from B_{p+1} with $w_\mathrm{i}^{(Bp+1)}$ as weights, can approximately be written as a linear combination of register counts, block-estimator from B_{p+1} with the initial block weights $d_\mathrm{i}^{(Bp+1)}$ as weights, and RW-estimators from B_1, …, B_p. Using Theorem 1 and the induction assumption, these estimators can approximately be written as HT-estimators.

Next, we distinguish again between pseudo and true RW-estimators. Every pseudo RW-estimator from B_{p+1} is, in fact, a block-estimator from B_{p+1} which uses the block weights $w_i^{(Bp+1)}$ as weights, and can be written as a linear combination of HT-estimators.

Every *first* true RW-estimator from B_{p+1} always has a reweighting scheme that only can contain register counts, RW-estimators from B_1 to B_p, and pseudo RW-estimators from B_{p+1}. The first true RW-estimator can thus approximately be written as a linear combination of register counts, RW-estimators from B_1 to B_p, and block-estimators with $w_i^{(Bp+1)}$ as weights. Each of those can approximately be written as a linear combination of HT-estimators (the register counts of course give rise to a constant term).

It then follows immediately that also the second, third, fourth etc. true RW-estimator from B_{p+1} approximately can be written as a linear combination of HT-estimators. ■

We have shown that every RW-estimator can approximately be written as a linear combination of HT-estimators. The problem is, of course, to find the correct linear combination. This is just a matter of writing out all regression estimators, until the level of HT-estimators at survey level is reached. In the next section, we give a small example.

3.4.3 Example of decomposition

Consider two surveys S_1 and S_2, with sample sizes n_1 and n_2, respectively. Let B_1 indicate the largest data block that can be constructed, that is, B_1 consists of the union of these surveys. Let $\lambda^{(S1)} = 1-\lambda^{(S2)} = n_1/(n_1+n_2)$. Assume that the final block weights $w_i^{(B1)}$ of B_1 are calibrated on the overall weighting scheme given by the register counts of table T_1, and the adjusted survey weights $w_i^{(S2)}$ of S_2 are obtained by calibration on the register counts of table T_2+T_3 Assume that S_1 is not calibrated, that is, the adjusted weights $w_i^{(S1)}$ are equal to the original design weights $d_i^{(S1)}$.

We want to decompose the RW-estimator of a table T, which is estimated from B_1. Suppose for simplicity that we are dealing with a pseudo RW-estimator. We then have

$$\begin{aligned}
\hat{\mathbf{Y}}_T^{RW(B1)} &= \sum_{i \in B_1} w_i^{(B1)} \mathbf{y}_{T,i} \\
&= \sum_{i \in B_1} d_i^{(B1)} \mathbf{y}_{T,i} + \mathbf{B}'_{T,T_1} \left(\mathbf{Y}_{T_1} - \sum_{i \in B_1} d_i^{(B1)} \mathbf{y}_{T_1,i} \right) \\
&= \lambda^{(S1)} \sum_{i \in S_1} w_i^{(S1)} \mathbf{y}_{T,i} + \lambda^{(S2)} \sum_{i \in S_2} w_i^{(S2)} \mathbf{y}_{T,i} \\
&\qquad - \mathbf{B}'_{T,T_1} \left(\lambda^{(S1)} \sum_{i \in S_1} w_i^{(S1)} \mathbf{y}_{T_1,i} + \lambda^{(S2)} \sum_{i \in S_2} w_i^{(S2)} \mathbf{y}_{T_1,i} \right) + \mathbf{c},
\end{aligned}$$

where the vector **c** contains known population totals from register counts. The estimates based on the adjusted survey weights of S_2 can be written out further, giving

$$\begin{aligned}
\hat{\mathbf{Y}}_T^{RW(B1)} &= \lambda^{(S1)} \sum_{i \in S_1} d_i^{(S1)} \mathbf{y}_{T,i} - \lambda^{(S1)} \mathbf{B}'_{T,T_1} \sum_{i \in S_1} d_i^{(S1)} \mathbf{y}_{T_1,i} \\
&\quad + \lambda^{(S2)} \left(\sum_{i \in S_2} d_i^{(S2)} \mathbf{y}_{T,i} + \mathbf{B}'_{T,T_2+T_3} \left(\mathbf{Y}_{T_2+T_3} - \sum_{i \in S_2} d_i^{(S2)} \mathbf{y}_{T_2+T_3,i} \right) \right) \\
&\quad - \lambda^{(S2)} \mathbf{B}'_{T,T_1} \left(\sum_{i \in S_2} d_i^{(S2)} \mathbf{y}_{T_1,i} + \mathbf{B}'_{T_1,T_2+T_3} \left(\mathbf{Y}_{T_2+T_3} - \sum_{i \in S_2} d_i^{(S2)} \mathbf{y}_{T_2+T_3,i} \right) \right) + \mathbf{c} \\
&= \lambda^{(S1)} \hat{\mathbf{Y}}_T^{HT(S1)} - \lambda^{(S1)} \mathbf{B}'_{T,T_1} \hat{\mathbf{Y}}_{T_1}^{HT(S1)} + \lambda^{(S2)} \hat{\mathbf{Y}}_T^{HT(S2)} - \lambda^{(S2)} \mathbf{B}'_{T,T_1} \hat{\mathbf{Y}}_{T_1}^{HT(S2)} \\
&\quad - \lambda^{(S2)} \mathbf{B}'_{T,T_2+T_3,T_2} \hat{\mathbf{Y}}_{T_2}^{HT(S2)} - \lambda^{(S2)} \mathbf{B}'_{T,T_2+T_3,T_3} \hat{\mathbf{Y}}_{T_3}^{HT(S2)} \\
&\quad + \lambda^{(S2)} \mathbf{B}'_{T,T_1} \mathbf{B}'_{T_1,T_2+T_3,T_2} \hat{\mathbf{Y}}_{T_2}^{HT(S2)} + \lambda^{(S2)} \mathbf{B}'_{T,T_1} \mathbf{B}'_{T_1,T_2+T_3,T_3} \hat{\mathbf{Y}}_{T_3}^{HT(S2)} + \mathbf{c},
\end{aligned}$$

so the RW-estimator is written out as HT-estimators at survey level, analogous to the decomposition in Equation (1). The **M**-matrices can be read off immediately. Indeed, they consist of products of (parts of) regression coefficient matrices **B**, and λ-factors. Note that the regression coefficient matrices $\mathbf{B}_{T,T_1}$ and $\mathbf{B}_{T_1,T_2+T_3}$ are based on different micro data and weights, *i.e.*, on the union of the surveys, and S_2, respectively.

In general, the decomposition tree of an arbitrary RW-estimators can be much bigger than in the above example, leading to products of more (parts of) **B** matrices. The basic idea is always the same though: all regression estimators are written out until the level of HT-estimators and population totals from register counts is achieved.

3.4.4 Approximate variance of the RW-estimator

Once the RW-estimator is decomposed in HT-estimators at survey level, its variance can readily be derived. We have

$$\begin{aligned} \mathrm{V}(\hat{\mathbf{Y}}_T^{RW}) &= \mathrm{V}\left(\sum_{k=1}^{K} \sum_{t \in T(S_k)} \mathbf{M}_{S_k,t} \hat{\mathbf{Y}}_t^{HT(S_k)} + \mathbf{c} \right) \\ &= \sum_{k=1}^{K} \mathrm{V}\left(\sum_{t \in T(S_k)} \mathbf{M}_{S_k,t} \sum_{i \in S_k} d_i^{(Sk)} \mathbf{y}_{t,i} \right). \end{aligned} \qquad (2)$$

The constant vector **c** can be ignored, since it does not contribute to the variance. In the second line we also used the assumption of independent surveys. For independent surveys, the variance of the sum of estimates from different surveys equals the sum of the variance of these survey estimates.

For each survey S_k we subsequently define new variables $\mathbf{z}_{T,i}^{(Sk)}$ satisfying

$$\mathbf{z}_{T,i}^{(Sk)} = \sum_{t \in T(S_k)} \mathbf{M}_{S_k,t} \mathbf{y}_{t,i} \ .$$

The new variable $\mathbf{z}_{T,i}^{(Sk)}$ is called ‘superresidual’, since it is the total residual of all the regressions which ultimately lead to the RW-estimator, see Knottnerus (2003). For a single regression of one survey variable Y on one register variable X, the superresidual would equal the normal residual $e_i = y_i - \mathbf{B}_{Y,X}\, x_\mathrm{i}$.

With these superresiduals, the variance can be written as

$$\mathrm{V}(\hat{\mathbf{Y}}_T^{RW}) = \sum_{k=1}^{K} \mathrm{V}\left(\sum_{i \in S_k} d_i^{(Sk)} \mathbf{z}_{T,i}^{(Sk)} \right).$$

The expression in the sum on the right-hand side of this equation is the variance of the HT-estimator of the population total of the superresiduals from survey S_k . For small sample sizes, the HT-estimator can be approximated with the Hansen-Hurwitz estimator for samples drawn with replacement, that is,

$$\hat{\mathbf{Z}}_T^{HT(Sk)} = \sum_{i \in S_k} d_i^{(Sk)} \mathbf{z}_{T,i}^{(Sk)} \cong \frac{1}{n_k} \sum_{i \in S_k} \frac{\mathbf{z}_{T,i}^{(Sk)}}{p_i^{(Sk)}} = \hat{\mathbf{Z}}_T^{HH(Sk)} \ ,$$

where n_k denotes the sample size of survey S_k, and the selection probabilities $p_i^{(Sk)}$ for small sampling fractions in a with-replacement sample are given by

$$p_i^{(Sk)} \cong \frac{\pi_i^{(Sk)}}{n_k} = \frac{1}{n_k d_i^{(Sk)}} \ .$$

The variance of each component of $\hat{\mathbf{Z}}_T^{HH(Sk)} = \left(\hat{Z}_{T,1}^{HH(Sk)}, \ldots, \hat{Z}_{T,P}^{HH(Sk)} \right)'$ can be estimated with the Hansen-Hurwitz variance estimator. For the p-th cell of table T we thus have, see Särndal *et al*. (1992),

$$\begin{aligned}\hat{\mathrm{V}}\left(\hat{Z}^{HH(Sk)}_{T,p}\right)&=\frac{1}{n_k}\frac{1}{n_k-1}\sum_{i\in S_k}\left(\frac{z^{(Sk)}_{T,pi}}{p_i}-\hat{Z}^{HH(Sk)}_{T,p}\right)^2\\&=\frac{n_k}{n_k-1}\left(\sum_{i\in S_k}\left(d_i^{(Sk)}z^{(Sk)}_{T,pi}\right)^2-\frac{1}{n_k}\left(\sum_{i\in S_k}d_i^{(Sk)}z^{(Sk)}_{T,pi}\right)^2\right).\end{aligned}\tag{3}$$

Note that the HH-variance formula does not require any knowledge on second order inclusion probabilities.

Altogether, we find for the estimated variance of the RW-estimator of table *T*

$$\hat{\mathrm{V}}(\hat{\mathbf{Y}}^{RW}_{T})\cong\sum_{k=1}^{K}\hat{\mathrm{V}}(\hat{\mathbf{Z}}^{HH(Sk)}_{T})\,.\tag{4}$$

In the derivation of the variance estimator, we have made some assumptions with respect to the surveys:

1. The surveys are non-overlapping.
2. The λ-factors that indicate the relative accuracy of surveys contributing to a certain rectangular data block, are constant (they are only survey-, and not record-dependent).
3. The sampling fractions of the surveys are small ($1 << n_k << N$).
4. The surveys are approximately independent.

With the first two assumptions, the RW-estimator can be written as a linear combination of HT-estimators. With the third assumption, the variance can be approximated with the HH-variance estimator. With the fourth assumption, the variance of the sum of estimators from different surveys can be approximated with the sum of the variance per survey.

In practice, not all surveys will satisfy these assumptions. An example will be given in Section 4.1. Especially when the λ-factors that indicate with which relative factor a survey contributes to a rectangular data block, are not constant, the derivation of variances is more complicated. In the next subsection we will come back to this.

In the derivation of the variance estimator, we also made some approximations and simplifications. These include:

1. The stochastic properties of the regressions coefficient matrices **B** are neglected.
2. The simplified variance estimator of the regression estimator is used, that is, the *g* weights are set equal to 1 in deriving variances.
3. The variance estimators are not corrected for nonresponse.

With the first approximation, RW-estimators can be approximated by linear combinations of HT-estimators. However, in the decomposition of RW-estimators, not only the size, but also the number of **B**-matrices can become very large. It is not a priori clear that the stochastic properties of the product of many **B**-matrices may be neglected. In the simulation study of Section 5, we will shortly come back to this. In DACSEIS, a European research project for variance estimation in complex survey designs, this assumption will be tested out in more detail, see Boonstra et al. (2003).

The neglect of the g-weights leads to the calculation of unconditional variances, and may lead to an underestimation of the true variance, for the variation of the g-weights is left out of the calculation. Although table estimates are corrected for nonresponse with a proper choice of the survey weighting scheme and the overall weighting scheme of the data blocks, the variance estimators are not corrected for this. Note also that nonresponse can cause large deviations from unity in the g-weights. To really take into account the effect of nonresponse on the variances of the table estimates, one should treat the surveys as two-phase samples, and calculate variance estimators from there, see Lundström and Särndal (2002). This leads, however, to (much) more complicated variance estimators.

3.4.5 Overlapping surveys

One of the assumptions for variance estimation of RW-estimators in the previous sections was that surveys should not have an overlap. In practice, however, surveys sometimes *do* have an overlap. In this section, we focus on the estimation of variances of RW-estimators from the overlap of two independent surveys S_1 and S_2.[8] To estimate variances in the case of overlapping surveys, we have to make the extra assumption that the surveys themselves are not corrected for nonresponse. The adjusted survey weights $w_i^{(S)}$ are thus equal to the initial design weights $d_i^{(S)}$ for each survey. Nonresponse correction has to be performed at the data block level.

For overlapping surveys, RW-estimators based on records in the overlap of S_1 and S_2 can not be written out as a linear combination of HT-estimators at survey level, simply because not all records of S_1 and S_2 are a member of the cross section. A solution could be to consider three completely disjoint surveys S_1', S_2', and S_3' instead of the two overlapping surveys S_1 and S_2. Survey S_1' corresponds to the n'_1 out of n_1 records that are only available in S_1 and not in S_2, survey S_2' corresponds to the n'_2 out of n_2 records that are only present in S_2 and not in S_1, and S_3' corresponds to the $n'_3 = n_1 - n'_1 = n_2 - n'_2$ records present in both S_1 and S_2. From these disjoint new surveys, rectangular data blocks must be constructed that correspond to the original S_1, S_2, and $S_1 \cap S_2$. The initial block weights should coincide with the desired weights $d_i^{(S1)}$, $d_i^{(S2)}$, and $d_i^{(S1)} d_i^{(S2)}$ for these blocks. It is easy to see that this can only be achieved by choosing record-dependent (as opposed to constant) λ-factors that indicate the relative accuracy of the surveys of which the data blocks are constructed. This complicates the practical estimation of the variances, since these record-dependent factors λ_i must be absorbed in the HT-estimates of Y-variables instead of in the constant matrix **M** in Equation (1).

This complication can be circumvented by assuming that the data blocks corresponding to S_1, S_2, and $S_1 \cap S_2$ are obtained from three completely independent surveys S_1, S_2, and $S_3 = S_1 \cap S_2$, and with survey weights $d_i^{(S1)}$, $d_i^{(S2)}$, and $d_i^{(S3)} = d_i^{(S1)} d_i^{(S2)}$, respectively. Each data block consists of one of the surveys, with $\lambda = 1$. Since we assume that the surveys are not corrected for nonresponse, any RW-estimator from the cross section $S_1 \cap S_2$ can subsequently be written as a linear combination of HT-estimators on S_1, S_2, and S_3. In this way, we obtain superresiduals $\mathbf{z}_{T,i}^{(Sk)}$ on S_1, S_2, and S_3. The RW-estimator of some table T is thus given by

$$\hat{\mathbf{Y}}_T^{RW} = \sum_{k=1}^{3} \sum_{i \in S_k} d_i^{(Sk)} \mathbf{z}_{T,i}^{(Sk)} .$$

Of course, we cannot estimate the variance of the RW-estimator by adding the variance of the superresiduals per survey, since the surveys are, in reality, not independent. The records in S_3 are, after all, also present in S_1 and S_2. Therefore, we continue by re-organizing the superresiduals according to the three disjoint groups given by surveys S_1', S_2', and S_3' from the previous paragraph, instead of the surveys S_1, S_2, and S_3. This gives

$$\begin{aligned}\hat{\mathbf{Y}}_T^{RW} &= \sum_{i \in S_1'} d_i^{(S1)} \mathbf{z}_{T,i}^{(S1)} + \sum_{i \in S_2'} d_i^{(S2)} \mathbf{z}_{T,i}^{(S2)} \\ &\quad + \sum_{i \in S_3'} \left(d_i^{(S1)} \mathbf{z}_{T,i}^{(S1)} + d_i^{(S2)} \mathbf{z}_{T,i}^{(S2)} + d_i^{(S3)} \mathbf{z}_{T,i}^{(S3)} \right)\end{aligned}$$

Since S_1', S_2', and S_3' are disjoint, the HH-variance per survey can be calculated and added to obtain an approximate estimation for the variance of the RW-estimator. The estimated variance of the p-th cell of table T is given by

$$\begin{aligned}\hat{\mathrm{V}}\left(\hat{Y}_{T,p}^{RW}\right) &= \frac{n_1'}{n_1'-1} \sum_{i \in S_1'} \left(d_i^{(S1)} z_{T,pi}^{(S1)} \right)^2 - \frac{1}{n_1'-1} \left(\sum_{i \in S_1'} d_i^{(S1)} z_{T,pi}^{(S1)} \right)^2 \\ &\quad + \frac{n_2'}{n_2'-1} \sum_{i \in S_2'} \left(d_i^{(S2)} z_{T,pi}^{(S2)} \right)^2 - \frac{1}{n_2'-1} \left(\sum_{i \in S_2'} d_i^{(S2)} z_{T,pi}^{(S2)} \right)^2 \\ &\quad + \frac{n_3'}{n_3'-1} \sum_{i \in S_3'} \left(d_i^{(S1)} z_{T,pi}^{(S1)} + d_i^{(S2)} z_{T,pi}^{(S2)} + d_i^{(S3)} z_{T,pi}^{(S3)} \right)^2 \\ &\quad - \frac{1}{n_3'-1} \left(\sum_{i \in S_3'} \left(d_i^{(S1)} z_{T,pi}^{(S1)} + d_i^{(S2)} z_{T,pi}^{(S2)} + d_i^{(S3)} z_{T,pi}^{(S3)} \right) \right)^2 ,\end{aligned}$$

where n'_1, n'_2, and n'_3 denote the sample sizes of S_1', S_2', and S_3'. In fact, it can be shown that the alternative method mentioned in the beginning of this section, with a suitable choice of the record-dependent λ-factors and weights $d_i^{(Sk)}$ of the surveys

[8] Note that, in order to have a non-negligible overlap, at least one of the surveys must be rather large.

S_1', S_2', and S_3', leads to the same variance estimator. In both cases, survey nonresponse can only be corrected for at the data block level. The resulting variance estimation is therefore only a crude approximation of the true variance.

3.5 Limitations

With repeated weighting, a fully consistent set of tables can be estimated. The weights that are used for each table estimate are either the final block weights (which were optimally chosen for reduction of variance and bias due to nonresponse), or weights that are slightly adjusted but still close to these block weights.[9] Although the estimation strategy for repeated weighting seems quite simple, the method is not without practical complications. In this section, we shortly mention some of these complications, which at present limit the applicability of the estimation strategy to 'relatively simple' table sets.

3.5.1 Harmonization

In the SSD, sample surveys are linked to registrations. If these surveys have variables in common, a separate rectangular data block consisting of the records from the union of these surveys can be created. Cross tabulations concerning these common variables may be estimated more accurately from the union of these surveys. After all, the variance of an estimate will be smaller when more data are available. A requirement is, however, that the definitions of the common variables in both surveys be the same. The routing, question formulation, answering categories etc., should be equivalent. Preferably, the sampling frames of both surveys should also be the same. It may lead to major biases in the estimates when the definitions of the common variables in the surveys differ, and records of both surveys are swept together. As a consequence, harmonization of the surveys is an important requirement for the successful implementation of the SSD. In practice, harmonization of the surveys may not be that simple to achieve, since the purposes of surveys may be quite different, naturally resulting in different definitions of variables. Statistics Netherlands puts at present considerable effort into this.

The requirement that categorical variables should have a hierarchical structure imposes some limitations on the flexibility of the method. This can be viewed as a disadvantage of repeated weighting, since one is no longer completely free to choose different categories for similar variables, depending on what one is interested in a particular table. However, one has to realize that also for an effective disclosure control of linked tables, a hierarchical structure of the variables is required.

[9] Note that with repeated weighting, there is no unique set of weights with which all tables from a certain source are estimated. The estimation process is therefore less transparent and the results are more difficult to reproduce by external researchers working on the same data.

3.5.2 Empty cells

A second complication is related to the occurrence of empty cells as a consequence of survey zeros. A problem arises when the interior of a cross tabulation has to be calibrated on some counted or estimated population total, but in the rectangular data block from which the table must be estimated, there are no records satisfying the conditions. It will then be impossible to find a solution for the repeated weighting estimator that satisfies the restrictions from the reweighting scheme. These empty-cell estimation problems arise in particular when the surveys have different sampling frames, or when certain groups in the population are heavily underrepresented in one or more of the surveys, and detailed estimates of this subpopulation or its complement are desired. Also, the splitting-up procedure might cause extra empty-cell problems.

One way to tackle this problem, is to combine several categories in the classification variables where the problem occurs. Due to the required consistency between all tables in a table set, these categories must be combined in all estimates, or, alternatively, an extra hierarchical level, in which these categories are combined, has to be added to the classification variable. The first option leads to information loss, and the second option will be difficult to implement in the process of consistent weighting, because it may be difficult to find cell combinations that solve all empty-cell problems and satisfy the required hierarchy simultaneously. The use of synthetic-like estimators such as the pseudo-Bayes estimators in Bishop et al. (1975), may be another way to treat empty-cell problems. The idea is to fill problematic empty cells with a small contribution, and solve the regression equations, see Houbiers (2003). Further research should point out whether or not such synthetic-like estimators can indeed be used to solve empty-cell problems in practical situations. Of course, one should also be cautious for improper use of these estimators. One might too easily be inclined to estimate tables with too little data. For this, synthetic estimators are not meant.

3.5.3 Edits

A third point is related to edit rules between variables. Edit rules define extra constraints between estimates. If consistency between all tables in a table set is required, then edit rules have to be taken into account as well. This is especially true if cross tabulations are estimated from different rectangular data blocks. It could, for example, easily happen that the number of people having a driver’s license in some small area, exceeds the number of people who are 18 years and older, see Kroese and Renssen (2000). In The Netherlands, no person younger than 18 years can have a driver’s license. As a consequence, when estimating a cross tabulation on possession of a driver’s license, one has to take the variable ‘age’ into account. Note that calibration on the underlying frequency table in case of tables related to general quantification variables, in fact constitutes a special case of edits. Especially when there are many edit rules between variables, the extra restrictions due to the edit rules absorb a large fraction of the degrees of freedom of the system. Therefore, the method of repeated weighting is less suitable when many edit rules between variables are present.

A problem related to edit rules arises when tables in one table set relate to different object types such as persons and households, and consistency between household characteristics and persons' characteristics is required. For instance, the total number of persons in these households should equal the total number of persons in the population. General integrated estimation procedures are available to ensure such consistency, see Lemaître and Dufour (1987). However these techniques are not yet incorporated in the method of repeated weighting. For simple cases, consistency between tables for different object types might also be obtained by introducing suitable quantification variables for households, and calibrating these variables on population totals for persons. Persons belonging to one household, will then have different weights. In more complicated situations, this simple strategy may not suffice. Generally, when using repeated weighting, one has to limit oneself to table sets relating to one single object type, or a subpopulation thereof.

3.5.4 Inconsistent margins

A fourth complication is related to a combination of empty-cell problems and edit rules. Suppose that a two-way table should be estimated from survey data, and both margins are already obtained from other data sources: the first margin corresponds to a register count, the second margin is obtained from survey data. Now suppose that the interior of the two-way table is (partially) diagonal. It is easy to see that it might not be possible to find a solution for the interior if both margins are unequal: the margins are inconsistent with the available data in the interior of the two-way table. The risk for such inconsistent margins increases when the interior of a table which one wishes to estimate, contains more survey zeros, or when (too) little data is available. Inconsistent margins can also be caused by the presence of (hidden) edit rules between variables. Use of the splitting-up procedure can also result in inconsistent margins, since all lower-dimensional margins of a table are estimated independently from the best data block.

3.5.5 Break-down of asymptotic behavior regression estimator

A fifth point relates to the limits of repeated weighting itself. In repeated weighting, the number of constraints on which needs to be calibrated can become quite large, especially when one is dealing with detailed crosstabulations and/or quantitative variables. With the increase in the number of constraints, the stability of the weights becomes less: the adjusted weights start to deviate more from the final block weights according to the distance function used, and they can even become negative. A large variability in the (adjusted) weights leads to higher variances. So, although the mean square error of the regression estimator initially decreases with the number of constraints, eventually it increases when more and more auxiliary variables are used in the estimation process, see Silva and Skinner (1997).

It is intuitively clear that repeated weighting can lead to lower variances as long as cell sizes are sufficiently large such that the regression estimator is asymptotically unbiased, and the number of restrictions is not too large such that the weights remain stable. After all, one works from ‘outside to inside’: the margins of a target table are pinned down by accurate estimates from large data sets, which leaves less variability for the interior of this table, even when the interior must be estimated from a smaller data set. In addition, the auxiliary variables used for reweighting are very well correlated with the target variables since the weighting model consists of margins of the tables that one wants to estimate. Of course, variance reduction will be less once the most important variables are inserted in the weighting model to determine the final block weights $w_i^{(B)}$. But when cell sizes are small and the number of constraints is large, the repeated weighting estimator can become less efficient than the estimator based on the block weights, and repeated weighting breaks down, as will be shown in the simulations in Section 5. This break-down point of the repeated weighting estimator is a topic for further research.

3.5.6 Variance estimation

A sixth complication is related to the variance of RW-estimators. The variance formulas derived in Section 3.4.4 are only valid for independent surveys with small sampling fractions. In practice, sampling fractions of the surveys might be rather large, and the assumption that the surveys can be approximated with a with-replacement sample might be incorrect. When the sampling fractions are very large, estimates from these surveys may be considered as register counts, and thus not contributing to the variance. However, for intermediate sampling fractions, it is not clear how variance formulas should be derived. This is a topic of research in the DACSEIS-project, see Boonstra et al. (2003). Moreover, the λ-factors indicating the relative accuracy of surveys, are assumed to be constant. However, unequal sampling frames, or general quality considerations with respect to the survey data may require a record-depend choice of λ. The ‘best’ choice of these λ-factors may even differ for each table to be estimated! This is not yet incorporated in the variance formulas.

Another problem related to variances is that, in practice, the target population in a survey might differ completely from the target population in the table set that one wants to estimate. For instance, in the survey, persons were interviewed, but the table set relates to jobs. In the variance estimation, it is not possible to switch from one target population to another target population. Adjustment of the survey weights to correct for nonresponse and reduce the variance of estimates, can therefore not be taken in to account. Instead, one has to use the adjusted survey weights as if they are the design weights for the target population of the table set.

Summarizing, the method of repeated weighting appears not suitable when many edit rules between variables are present, or when lack of data causes many empty-cell problems. The risk for the latter is high when one wants to make detailed estimates with too little data, as might happen when one is interested in estimates on subpopulations of the target population. Since the estimates on the particular subpopulation should be consistent with estimates on the target population, the complement of the subpopulation must be included in the estimate. Empty-cell problems can easily occur when this complement is small.

Repeated weighting is suitable to estimate relatively simple table sets, which do not involve too many edit rules, and where sufficient micro data are available. In the next section we give two examples of table sets which were estimated at Statistics Netherlands, using repeated weighting.

4. Applications

Statistics Netherlands has developed a prototype software package to estimate table sets with repeated weighting. The software package is called VRD and can be used once rectangular data blocks are extracted from the micro data. After entering all relevant meta data about the categories of classification variables, the hierarchical relations between different levels of each classification variable, and the composition of the rectangular data blocks, the user can indicate which tables should be estimated. These tables are subsequently estimated consistently, each table from the most suitable data block. The software package is still under construction. The basic functionalities are built in, but more advanced functionalities which deal with empty-cell problems and edit rules, are not yet included. Variances of the estimates can, however, be estimated under the assumptions mentioned in Section 3.4.4. In the case of overlapping surveys, superresiduals can in principle be calculated inside VRD, but the re-organizing over disjoint surveys, and the actual variance estimation must be performed outside VRD.

With this software package, two table sets were estimated at Statistics Netherlands, see Statistics Netherlands (2002). The first table set involves the Structure of Earnings Survey, and the second table set relates to search behavior on the labor market of persons on social welfare. In Sections 4.1 and 4.2, we discuss these applications of the method in more detail.

4.1 Structure of Earnings Survey

The Structure of Earnings Survey (SES) is a publication on jobs in The Netherlands and the (average) hourly, monthly, and yearly wages for these jobs, against some relevant background variables such as business classification, and age, gender, and educational and professional level of the persons having these jobs. The target population is the ‘jobs of persons living in The Netherlands, aged 15 to 64, excluding the institutional population’. In line with the policy to minimize the response burden, Statistics Netherlands does not conduct a separate survey among businesses to collect the data for the SES, since these data can also be obtained from a combination of other sources, in particular, the registers of jobs and persons, and survey data from the Employment and Wages Survey (EWS) and the Labor Force survey (LFS). The SES describes the situation at December 31, so in that case, the register of persons is linked to the register of jobs at this reference date. The population consists of approximately 6.6 million jobs. Survey data from the EWS and the LFS are also needed for the SES estimates.

The first survey, the EWS, is a large two-stage survey among businesses on wages and hours of employment of their employees on December 31, each year. In the first stage of this survey, companies are sampled, in the second stage employees of these companies are sampled. The data are mainly collected electronically from companies and salary administration offices. Often, information on all employees of a company is received, instead of only the data corresponding to the employees in the survey. Also, administration offices often send information on all companies, instead of only those that are sampled. This leads to some selectivity in the data. For instance, government employees are overrepresented, and small companies are underrepresented. Due to this selectivity, the variation in the (adjusted) survey weights is quite big. From the data in the EWS, the yearly, monthly, and hourly salary per employee can be calculated. The EWS consists of approximately 2.8 million records.

The second survey, the LFS, is a household survey. In this survey, data related to education, labor characteristics, and labor market position of persons are collected.[10] The LFS is a continuous survey, meaning that the sampling and surveying of persons is spread throughout the year, instead of on one particular day. However, for variables which are relatively fixed in time (such as level of education and profession which are used in the SES), the LFS can without much error be considered as being held at one particular date. Each year, approximately 100,000 persons are sampled in the LFS. Roughly half of them have a job. To increase the mass, data from two LFS’s (the years symmetric around December 31) are used. This amounts to approximately 100,000 records of persons with a job.

The survey data are linked to the register at the reference date. Figure 1 in Section 3.2.1 shows the linked data sets used for the SES. In this figure, the EWS and LFS are linked to the register of jobs to which person’s characteristics from the register of persons are added. As can be seen from the figure, the surveys have a partial record overlap. The overlap consists of approximately 50,000 records. Note that it is assumed that the LFS data are still/already valid at the reference date of the SES (December 31), although the actual surveying is performed sometime (but within twelve months) around the reference date. Note also that one single person in the LFS may have several jobs simultaneously. In such a case, the LFS record is copied (including its adjusted survey weight) to all jobs belonging this particular person.

[10] Lack of harmonization is the reason that, for instance, the ISSS data on education can not yet be used together with the LFS data to estimate tables related to education in the SES table set.

As mentioned in Section 3.2.1, four rectangular data blocks can be constructed from the micro data. The initial block weights for data blocks 2 and 3 are given by the (adjusted) survey weights for the EWS and the LFS. The initial survey weights of data block 4 are given by the product of the EWS and LFS weights of each record. This product is, however, bounded to a maximum, to prevent exceptionally large weights for records that accidentally have a large EWS ánd a large LFS weight. An overall weighting scheme given by the register counts of ‘age class × sex + business class’ is subsequently applied to data blocks 2, 3, and 4. This weighting scheme should correct for nonresponse, reduce the variance of estimates, and already provide some numerical consistency between estimates and register counts. In principle, RW-estimates from data blocks 2 and 3 could also be used in the overall weighting scheme of data block 4, but this is not done here.

A given table set is subsequently estimated with repeated weighting. The dimensional part of the tables in the table set basically consists of [$\Gamma \in$ {1 (for population totals of count variables), sex, age class, sex × age class, business class, size class, ethnicity, working hours}], [$\Gamma \times$ educational level], [$\Gamma \times$ professional level], and [educational level × professional level]. The count variables include 1 (for frequency tables), and the general quantification variables ‘monthly salary’, ‘yearly salary’, and ‘total hours worked’. For each estimate, the largest rectangular data block – in terms of number of records – that contains all relevant variables simultaneously, is used. Average monthly and yearly salaries are subsequently estimated from the ratio of total salaries and the corresponding frequency counts. Average salaries per hour are estimated from the ratio of total salaries per month and total hours worked. Note that all tables on general quantification variables are calibrated on the underlying frequency table.

Due to empty cells, a few tables in the SES-table set can not be estimated consistently. Empty-cell problems occur mainly with detailed frequency tables related to the classification variable ‘business class[(4)]’. This register variable has a very fine classifications, and when crossing with fine classifications of, for instance, ‘educational level[(2)]’, this might lead to rare combinations which are present in the LFS, but not in the cross section of the EWS and the LFS. Moreover, tables related to general quantification variables require, on average, larger adjustments in the reweighting process. The calibration on the underlying frequency table leads in some cases to empty-cell problems. Most tables can nevertheless be estimated without problems, and these estimates are published in Statistics Netherlands (2002).

Variances of these estimates can not be estimated with the formulas of Sections 3.4.4, because

- There is an overlap between the EWS and the LFS.
- The sampling fraction of the EWS is ½, that is, not particularly small.
- The target population of the LFS concerns persons, and not jobs, so there is a switch in target population of the survey and the target population of the table set.

Assuming that estimates from the EWS can be considered as register counts, we show in Section 5.2.4 that the method proposed in Section 3.4.5 gives reasonable estimates of the variance of the RW-estimates from the overlap, given that the sample size is sufficient.

4.2 Search behavior on the labor market.

The second application of repeated weighting is much simpler: only one survey (the LFS) is linked to a register. The table set of interest relates to search behavior on the labor market of 'persons on social welfare, living in The Netherlands, aged 15 to 64, excluding the institutional population'. All tables are frequency counts. The register is now obtained by linking data from the register of jobs and the register of social welfare payments to the MBA, and selecting all persons belonging to the target population. Persons' characteristics, and information regarding (type of) social welfare payments and job characteristics can be counted from the linked registers. Other variables of interest, such as 'educational level' and 'search behavior' are only available from survey data, in particular, from the LFS. Since the latter variable changes rapidly in time, survey data must be linked to the register at the survey date. Consequently, an (approximate) average register must be used, instead of a register at a certain reference date. Figure 2 schematically shows the micro data used to estimate the table set with repeated weighting.

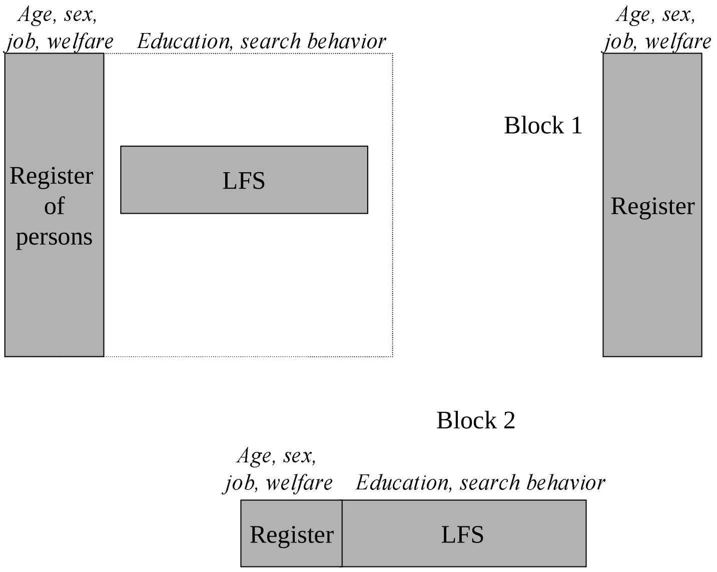

Figure 2. Micro data and data blocks for the table set on search behavior on the labor market.

From the micro data, two rectangular data blocks can be constructed, as indicated in Figure 2. Note that in this case, the persons in the register block have a weight according to the fraction of the year that they were part of the target population, instead of unity. The persons in the LFS data block have the usual (adjusted) survey weights as initial block weights. These weights are benchmarked on certain register counts using an overall weighting scheme. The tables in the table set can subsequently be estimated consistently using repeated weighting. Since in this application only one survey is linked to the register, variances of the table estimates can be obtained in VRD.

Many tables in this table set relate to subpopulations of persons having a special type of social welfare. Since these estimates should be consistent with the complete target population of persons having any kind of social welfare, the complement of the subpopulation must be included in the table estimate. The dimensions of such table should therefore be extended with the classification ‘having this special type of social welfare or not’. When the complement, or the subpopulation itself, is small, this might lead to empty-cell problems. Most tables in the table set can however be estimated without problems. Table estimates and variances of these estimates are published in Statistics Netherlands (2002).

5. Simulations

In Section 3, we have described the method of repeated weighting, and derived approximate variance formulas for RW-estimators. In this section, which is based on Van Duin and Snijders (2003), we discuss the results of two simulations studies in which the quality of RW-estimators is investigated. Also, the variance formulas for these RW-estimators are tested. In particular, we focus on the following issues:

- The bias of the RW-estimates, where tables are estimated with and without splitting-up.
- Comparison of variances of table estimates, where the tables are estimated (not consistent) with final block weights, and (consistent) with repeated weighting, with and without splitting-up.
- Test of the approximate variance formulas.

Since the regression estimator is only asymptotically unbiased, one might in first instance fear that the bias of the RW-estimator may grow excessively due to repeated application of the regression estimator, especially when the splitting-up procedure is applied. Similarly, the variance formulas derived for the RW-estimator are based on an approximation by means of a Taylor linearization of the regression estimator. Therefore, repeated application of the Taylor linearization in the RW-estimators may result in a poor approximation of the variance estimator. In the simulations we hope to get insight into the approximation of both the variance formulas and the bias. Furthermore, we hope to get insight into the accuracy, that is, the mean square error, of the RW-estimator compared to the accuracy of the regression estimator.

Intuitively, we expect that the repeated weighting estimator gives better estimates (in the sense of lower variances) than the estimates based on the final block weights, when the number of records contributing to these (table cell) estimates is sufficiently large (N >> n >> 1), and aggregates of these estimates can be obtained from larger data sets. On the other hand, we expect that repeated weighting does not improve the accuracy of the estimates when the number of contributing records is very small, and consequently, variances are very high to begin with. We expect that repeated weighting breaks down when the assumptions that underlie the asymptotic behavior of the regression estimator are violated. To investigate the above issues, two simulations studies are performed. In the next section, we describe the set-up of these simulation studies. In Sections 5.2 and 5.3, we show the results.

5.1 Simulation set-up

In the simulations, we focus on some (artificial) population of jobs. The first step in the simulation study consists of the generation of a 'complete population register'. This complete population register should contain all variables used in the simulations, and all elements of the target population. The population register is based on the EWS, see Section 4.1. For each record in the EWS, the variable 'educational level' is imputed. The imputations are based on the actual probability distribution of education in the LFS, per class of 'age class × sex × ethnicity'. Each record in the EWS is subsequently copied as many times as the corresponding (truncated) weight. If the weight is non-integer, the decimal part of it gives the probability for one more copy. The result is a complete population register, containing approximately 6.6 million elements. Variables in the population register are, among others, the classification variables 'sex', 'age class', 'business classification', 'working hours per week', and 'educational level', and the quantification variable 'monthly salary'. All classification variables consist of only one level of classification.

Although all variables are available in the complete population register, from now we assume that only the variables 'sex', 'age class', and 'business classification' are true register variables. The variables 'working hours per week', 'monthly salary', and 'educational level' are collected in surveys.

The basic idea in both simulations is that two surveys are drawn according to some design from the population. Subsequently, rectangular data blocks are constructed, weights are attached to the records in these data blocks, and a given table set is estimated. The table set is estimated with repeated weighting, both with and without splitting-up. In the latter case, all margins that can be obtained from larger data blocks, are added to the table set. Similarly, all margins that can be estimated consistently with the final block weights, are added. These margins are estimated first. In the splitting-up case, all most-detailed margins are added, irrespective from which data block they must be estimated. In addition to the RW-estimates, the target tables in the table set are estimated with the final block weights, thus without reweighting for consistency. In the non-splitting-up case, the variance of the target tables is also estimated. This procedure is repeated a number of times. After each run, table estimates and variance estimates are stored. From these table estimates, biases and (simulation) variances can be calculated, and the simulation variances can be compared with the average of the estimated variances.

The first simulation (Simulation 1) is similar to the situation in the SES, see Figure 1. Survey S_1 contains variables ‘working hours per week’ and ‘monthly salary’. Survey S_2 contains the variable ‘educational level’. The sampling size of S_1 is approximately 2.8 million records, about half of the population. These records are drawn with the EWS-inclusion probabilities. The sampling size of S_2 is about 100,000 records. These records are drawn with inclusion probabilities which resemble the inclusion probabilities of jobs (not persons) in the LFS. About half of the records in S_2 are also present in S_1.

The second simulation consists of two parts: Simulation 2A and 2B. In both parts, survey S_1 is again assumed to contain the variables ‘working hours per week’ and ‘monthly salary’. Survey S_2 contains in addition the variable ‘educational level’. The surveys have no overlap, but they do have variables in common. In Simulation 2A, both surveys are drawn using simple random sampling, and the sampling sizes of both surveys are 100,000 records. To avoid overlap, first a sample of 200,000 records is drawn, and this sample is divided at random in two parts of 100,000 records each. In Simulation 2B, both surveys are drawn with unequal inclusion probabilities. The records in S_1 are drawn with the LFS-inclusion probabilities, as in simulation 1. This results again in a sample of approximately 100,000 records. The records in S_2 are drawn with an adapted version of the EWS inclusion probabilities from simulation 1. The EWS inclusion probabilities are adapted such that the size of the resulting sample is about 100,000 instead of 2.8 million records. In addition, the internal variation of the weights is reduced by a factor of about 2, by rescaling the weights towards their average value. Figure 3 shows the micro data, and the corresponding data blocks in this simulation.

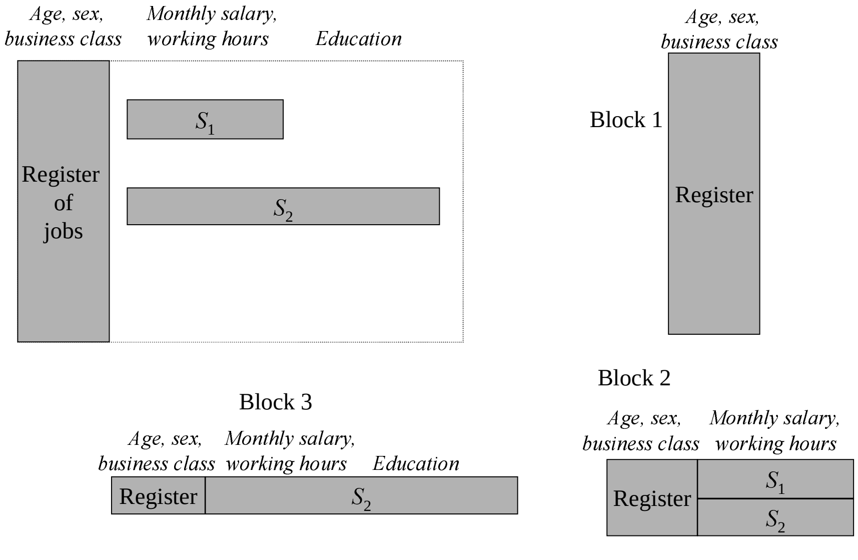

Figure 3. Overview of micro data and data blocks in second simulation.

As can be seen from this figure, data block 2 consists of the union of the two surveys. For Simulation 2A, we take $\lambda^{(S1)} = n_1/(n_1 + n_2) = 1/2$, and $\lambda^{(S2)} = n_2/(n_1 + n_2) = 1 - \lambda^{(S1)}$. For Simulation 2B, we use the effective sample sizes instead of the real sample sizes to determine the λ-factors, see Section 3.2.2. We thus get $\lambda_{\text{eff}}^{(S1)}$ =

$n_{1\text{eff}}/(n_{1\text{eff}} + n_{2\text{eff}}) =$ 0.81, and $\lambda_{\text{eff}}^{(S2)} = n_{2\text{eff}}/(n_{1\text{eff}} + n_{2\text{eff}}) = 1 - \lambda_{\text{eff}}^{(S1)} = 0.19$, given the inclusion probabilities of S_1 and S_2. Note that compared to sample S_1, the effective size of sample S_2 is reduced by a factor of 4 due to the large variation in the inclusion probabilities. Since both surveys have small sampling fractions, and have no overlap, the variances of RW-estimators can be estimated with the formulas from Section 3.4.4.

In all simulations, the final block weights of the data blocks are benchmarked on the register frequency counts 'sex × age class + business class'. The target table set in the simulations consists of the tables

T_1 = [sex × working hours per week × educational level] × 1

T_2 = [sex × working hours per week × educational level] × monthly salary.

By taking the ratio of these two tables, average monthly salaries can be calculated.

Given the two target tables, and the structure and overall weighting scheme of the data blocks in both simulations, one can easily derive the margins that must be added to the table set to guarantee optimal use of the data. Table 4 gives an overview of all tables that must be estimated when the splitting-up procedure is not applied. Table 5 gives an overview of all tables that must be estimated when the splitting-up procedure ís applied. In both tables, the last two lines relate to the actual target tables T_1 and T_2. Note that all tables that must be estimated from data block 4 in Simulation 1, are estimated from data block 3 in Simulations 2A and 2B.

Table 4. Overview of tables for repeated weighting without splitting-up.

Table set (without splitting-up)	***Block***	***Reweighting scheme***
[age × sex × business class] × 1	1	Register count
[sex × working hours] × 1	2	Consistent with block weights
[sex × working hours] × monthly salary	2	Consistent with block weights
[sex × education] × 1	3	Consistent with block weights
[sex × working hours × education] × 1	4 / 3	[sex × education] × 1 + [sex × working hours] × 1
[sex × working hours × education] × monthly salary	4 / 3	[sex × working hours × education] × 1 + [sex × working hours] × monthly salary

Table 5. Overview of tables for repeated weighting with splitting-up.

Table set (with splitting-up)	***Block***	***Reweighting scheme***
[age × sex × business class] × 1	1	Register count
[sex × working hours] × 1	2	Consistent with block weights
[sex × working hours] × monthly salary	2	Consistent with block weights
[sex × education] × 1	3	Consistent with block weights
[working hours × education] × 1	4 / 3	[working hours] × 1 + [education] × 1
[education] x monthly salary	4 / 3	[education] × 1 + monthly salary
[sex × education] × monthly salary	4 / 3	[sex × education] × 1 + [sex] × monthly salary + [education] × monthly salary
[working hours × education] × monthly salary	4 / 3	[working hours × education] × 1 + [working] × monthly salary + [education] × monthly salary
[sex × working hours × education] × 1	4 / 3	[sex × education] × 1 + [sex × working hours] × 1 + [working hours × education] × 1
[sex × working hours × education] × monthly salary	4 / 3	[sex × working hours × education] × 1 + [sex × working hours] × monthly salary + [sex × education] × monthly salary + [working hours × education] × monthly salary

5.2 Results Simulation 1

In the first simulation, $M = 600$ runs are performed according to the set-up described before. In each run $m \in \{1, 2, \ldots, M\}$, table estimates for T_1 and T_2 are obtained, both with and without the splitting-up procedure. The non-splitting-up estimates are indicated with $\hat{\mathbf{Y}}_{T_1,m}^{RW}$ and $\hat{\mathbf{Y}}_{T_2,m}^{RW}$, and the splitting-up estimates are denoted with $\hat{\mathbf{Y}}_{T_1,m}^{RW+}$ and $\hat{\mathbf{Y}}_{T_2,m}^{RW+}$. Moreover, these tables are also estimated without repeated weighting, that is, with the final block weights $w_i^{(B4)}$ as weights. These estimates are denoted by $\hat{\mathbf{Y}}_{T_1,m}^{SW}$ and $\hat{\mathbf{Y}}_{T_2,m}^{SW}$.

From these estimates, average monthly salaries can be obtained by dividing the total monthly salaries in T_2 by the corresponding frequency counts in T_1. Tables related to average monthly salaries are denoted as $\hat{\mathbf{Y}}_{T_2/T_1,m}^{RW}$, $\hat{\mathbf{Y}}_{T_2/T_1,m}^{RW+}$, and $\hat{\mathbf{Y}}_{T_2/T_1,m}^{SW}$, respectively, where it is understood that each components of $\hat{\mathbf{Y}}_{T_2,m}^{RW}$ is divided by the corresponding component of $\hat{\mathbf{Y}}_{T_1,m}^{RW}$, and similarly for the other two estimated ratios.

Variances for $\hat{\mathbf{Y}}_{T_1,m}^{RW}$ and $\hat{\mathbf{Y}}_{T_2,m}^{RW}$ are estimated, using the method for overlapping surveys proposed in Section 3.4.5. These variance estimates are denoted by $\hat{\mathrm{V}}\left(\hat{\mathbf{Y}}_{T_1,m}^{RW}\right)$ and $\hat{\mathrm{V}}\left(\hat{\mathbf{Y}}_{T_2,m}^{RW}\right)$, respectively. In the variance estimation, we assume that estimates from the large survey S_1 can be considered as register counts. The overall weighting scheme of data blocks 3 and 4 is taken into account in the variance calculation. The variance of the p-th cell in the ratio $\hat{\mathbf{Y}}^{RW}_{T_2/T_1,m}$ can be obtained from usual Taylor linearization, see Särndal et al. (1992),

$$\hat{\mathrm{V}}\left(\hat{Y}^{RW}_{T_2/T_1,pm}\right) \cong \frac{1}{\left(\hat{Y}^{RW}_{T_1,pm}\right)^2}\Bigg[\hat{\mathrm{V}}\left(\hat{Y}^{RW}_{T_2,pm}\right) + \left(\hat{Y}^{RW}_{T_2/T_1,pm}\right)^2 \hat{\mathrm{V}}\left(\hat{Y}^{RW}_{T_1,pm}\right) - 2\hat{Y}^{RW}_{T_2/T_1,pm}\mathrm{C}\left(\hat{Y}^{RW}_{T_1,pm}, \hat{Y}^{RW}_{T_2,pm}\right)\Bigg], \tag{7}$$

where $\mathrm{C}\left(\hat{Y}^{RW}_{T_2,pm}, \hat{Y}^{RW}_{T_2,pm}\right)$ denotes the covariance between the p-th cell totals of tables T_1 and T_2. The covariance is estimated with

$$\hat{\mathrm{C}}\left(\hat{Y}^{RW}_{T_2,pm}, \hat{Y}^{RW}_{T_2/T_1,pm}\right) = \frac{1}{2}\left[\hat{\mathrm{V}}\left(\hat{Y}^{RW}_{T_1,pm} + \hat{Y}^{RW}_{T_2,pm}\right) - \hat{\mathrm{V}}\left(\hat{Y}^{RW}_{T_1,pm}\right) - \hat{\mathrm{V}}\left(\hat{Y}^{RW}_{T_2,pm}\right)\right].$$

5.2.1 Cell size

Both tables T_1 and T_2 are estimated from data block 4. This data block consists on average of 44,967 records. Table 6 in appendix A shows an overview of the average cell sizes of the target table [sex × working hours per week × education level]. As can be seen from the overview in this table, the sizes of some cells in the target tables are very small, and other cells have reasonably large size. We expect that the former cells will have relatively large biases and high variances and that repeated weighting does not improve the accuracy of these estimates much, if at all. For the cells which have larger size, we expect that repeated weighting lowers the variances, compared to the estimates based on the final block weights.

As said in Section 4.1, the variation in inclusion weights of the records in the EWS, that is, in S_1, is rather large. As a consequence, the weights in data block 4 also vary a lot, resulting in larger variances of the estimates than one would expect when dealing with equal inclusion probabilities. When the weights are uncorrelated with the target variable, the variation in these weights contributes a factor $1 + L$ to the variance, where

$$1 + L = \frac{1}{n}\sum_{i=1}^{n} w_i^2 \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^2 .$$

The increase in variance can also be viewed as if the sample has an effective size $n_{\mathrm{eff}} = n/(1+L)$. For data block 4, $1 + L = 10.5$. Hence, the effective block size is only 4,278, instead of 44,967 records. The value of L varies somewhat from cell to cell (the maximum is 12.1, the minimum equals 8.8). The effective cell size for the target table is also shown in Table 6 in Appendix A. Some cells appear to have a very low effective sizes, which will result in very large variances and biases, and poor variance estimates.

5.2.2 Bias

Given the results of M = 600 runs, the bias of the RW-estimate $\hat{\mathbf{Y}}^{RW}_{T_2/T_1}$ can be estimated with

$$\hat{\mathrm{B}}\left(\hat{\mathbf{Y}}^{RW}_{T_2/T_1}\right) = \frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{Y}}^{RW}_{T_2/T_1,m} - \mathbf{Y}_{T_2/T_1} = \mathrm{E}\left(\hat{\mathbf{Y}}^{RW}_{T_2/T_1}\right) - \mathbf{Y}_{T_2/T_1}\,, \tag{8}$$

where $\mathbf{Y}_{T_2/T_1}$ are the average monthly salaries in the population. This population average is obtained from the complete population register. Similar expression are calculated for the biases of $\hat{\mathbf{Y}}^{RW+}_{T_2/T_1}$ and $\hat{\mathbf{Y}}^{SW}_{T_2/T_1}$. The resulting biases, relative to the average salaries in the population, $\mathbf{Y}_{T_2/T_1}$, are shown in Table 9 in Appendix B.

As can be seen from this table, both the SW-estimator (which is based on the final block weights, and is in fact a usual regression estimator), as well as the RW-estimator have a bias. Especially for cells with a small (effective) cell size, the bias can be large. In section 5.2.3, we calculate whether these biases are significant; the significant biases are marked grayish. The bias is considerably smaller when more records are contributing to a cell. For cells with a sufficiently large cell size, the bias does not systematically seem to increase or decrease due to repeated weighting, compared to the SW-estimator; for some cells there is an increase, for other cells a decrease. However, when the assumptions underlying the asymptotic behavior of the regression estimator are violated, and the bias of the SW-estimator is already large, the RW-estimator performs even worse and results in a larger bias compared to the SW-estimator. The results in Table 9 do not show that the splitting-up procedure has an increasing effect on the bias of the RW-estimates. We thus conclude that repeated weighting does not cause an excessive increase in the bias of the estimates, as long as the cell sizes are sufficiently large such that the regression estimator is asymptotically unbiased.

5.2.3 Simulation variance

Next, we investigate the variance of the M average monthly salary tables estimated in the simulation. The simulation variance for the p-th cell in the ratio table $\hat{\mathbf{Y}}^{RW}_{T_2/T_1}$ is given by

$$\hat{\mathrm{S}}^2\left(\hat{Y}^{RW}_{T_2/T_1,p}\right) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Y}^{RW}_{T_2/T_1,pm} - \mathrm{E}(\hat{Y}^{RW}_{T_2/T_1,p})\right)^2\,. \tag{9}$$

Similar expressions are calculated for the components of $\hat{\mathbf{Y}}^{RW+}_{T_2/T_1}$ and $\hat{\mathbf{Y}}^{SW}_{T_2/T_1}$. In Table 12 in appendix C, an overview is given of the standard errors, relative to the average salary in the population, $\mathbf{Y}_{T_2/T_1}$, in each cell. As can be seen from this table, standard errors might be rather large, especially for cells with low (effective) size.

The table shows also that the variances of RW-estimates are, in general, lower than the variances of the SW-estimates based on the final block weights. This is more clearly visible from Figure 4a in Appendix C. This figure shows the relative difference between the SW- and RW-standard error, relative to the SW-standard error in each cell p of the target table, that is

$$\frac{\hat{S}\left(\hat{Y}^{SW}_{T_2/T_1,p}\right)-\hat{S}\left(\hat{Y}^{RW}_{T_2/T_1,p}\right)}{\hat{S}\left(\hat{Y}^{SW}_{T_2/T_1,p}\right)},$$

against a measure of the quality of the cell for which we take the estimated coefficient of variation

$$cve_p=\frac{\hat{S}\left(\hat{Y}^{SW}_{T_2/T_1,p}\right)}{Y_{T_2/T_1,p}}.$$

Note that denominator in this expression contains the true average salaries in the population. Apparently, in this case repeated weighting leads to a reduction of about 10 percent in the standard errors of the estimates. This is due to better use of auxiliary information: some margins of this table are obtained from the very large S_1-survey. In Figure 4b, the relative difference in the standard deviations of the RW- and RW+-estimates is shown. Differences are rather small, although the standard deviation in the splitting-up case is slightly bigger.

In order to see whether the differences between the variances of the SW- and RW-estimator are significant, we have to estimate the variance of the quantity

$$\begin{aligned}\hat{Q}_p &= \hat{S}^2\left(\hat{Y}^{SW}_{T_2/T_1,p}\right)-\hat{S}^2\left(\hat{Y}^{RW}_{T_2/T_1,p}\right)\\ &=\frac{1}{M}\sum_{m=1}^{M}\frac{M}{M-1}\left[\left(\hat{Y}^{SW}_{T_2/T_1,pm}-\mathrm{E}(\hat{Y}^{SW}_{T_2/T_1,p})\right)^2-\left(\hat{Y}^{RW}_{T_2/T_1,pm}-\mathrm{E}(\hat{Y}^{RW}_{T_2/T_1,p})\right)^2\right]\\ &\equiv\frac{1}{M}\sum_{m=1}^{M}\hat{Q}_{pm}\end{aligned}$$

in each cell p of the table. We thus have

$$\hat{S}^2\left(\hat{Q}_p\right)=\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_{pm}-\hat{Q}_p\right)^2 .$$

The difference between the two variances is significant when

$$\left|\hat{S}^2\left(\hat{Y}^{SW}_{T_2/T_1,p}\right)-\hat{S}^2\left(\hat{Y}^{RW}_{T_2/T_1,p}\right)\right|>\frac{1.96\,\hat{S}\left(\hat{Q}_p\right)}{\sqrt{M}}.$$

The significance of the difference between the variances of the SW- and RW+-estimator, and between the RW- and RW+-estimator can be derived similarly. Significant differences are marked grayish in Table 12 in Appendix C. The relative standard error of the SW-estimator is marked grayish when its variance differs significantly from the variances of both the RW- as well as the RW+-estimator. The relative standard errors of the RW- and RW+-estimator are marked grayish when they differ significantly from each other. In Figure 4, significant differences are indicated with boxes around the points. The results suggest that repeated weighting leads in this case to more accurate estimates. However, the gain in accuracy is slightly less when the splitting-up procedure is used.

From the standard deviations in Table 12, one can also derive whether the biases given in Table 9 in Appendix B are significant. These biases are significant when the bias is larger than the error on the (average) table estimate, that is, when

$$\left|\hat{\mathrm{B}}\left(\hat{\mathbf{Y}}_{T_2/T_1}^{RW}\right)\right| > \frac{1.96 \times \hat{\mathrm{S}}\left(\hat{\mathbf{Y}}_{T_2/T_1}^{RW}\right)}{\sqrt{M}},$$

and similar expressions for $\hat{\mathbf{Y}}_{T_2/T_1}^{RW+}$ and $\hat{\mathbf{Y}}_{T_2/T_1}^{SW}$. Significant biases are marked grayish in Table 9.

5.2.4 Variance estimate

Finally, we test the variance formulas from Section 3.4.5, using the software package VRD. Since S_1 has a large sampling fraction, we cannot approximate this survey with a with-replacement survey. Instead, all estimates obtained from S_1 are considered as register counts. They do not contribute to the variance. As said before, variances are only estimated in the non-splitting-up case. In each simulation $m \in \{1,2,\ldots,M\}$, the variance $\mathrm{V}\left(\hat{Y}_{T_2/T_1,pm}^{RW}\right)$ of cell p of the RW-estimate $\hat{\mathbf{Y}}_{T_2/T_1,m}^{RW}$ is estimated using the ratio formula (7). For each simulation m we subsequently check whether the average salaries in the population lie inside the (estimated) 95%-confidence interval, that is, whether

$$\hat{Y}_{T_2/T_1,pm}^{RW} - 1.96\sqrt{\hat{\mathrm{V}}\left(\hat{Y}_{T_2/T_1,pm}^{RW}\right)} \leq Y_{T_2/T,p} \leq \hat{Y}_{T_2/T_1,pm}^{RW} + 1.96\sqrt{\hat{\mathrm{V}}\left(\hat{Y}_{T_2/T_1,pm}^{RW}\right)}. \qquad (10)$$

The coverage probability P_p of cell p is given by the fraction of simulations for which the true population value of the average salary lies within the estimated confidence interval. Figure 7 in Appendix D shows the coverage probabilities for each cell in the target table, versus the quality of the cell (measured with the estimated coefficient of variation (cve)). Ideally, the coverage probability should be 95%. However, as one can see from the figure, the coverage probability lies below 95% for almost all cells, and drops even well below 95%, when the cve becomes large. Since the coverage probability has a binomial distribution, its standard deviation is given by $\sqrt{P_p(1-P_p)/(M-1)}$. From this it follows that the coverage probability P_p deviates significantly from 95%, when it is larger than about 96.5%, or smaller than about 93%. Significant deviations are indicated with boxes around the points in Figure 7. It is evident that the variance formula underestimates the true variance, especially for cells with high cve, or, equivalently, low effective cell size. This is also expected, because the assumptions underlying the asymptotic behavior of the regression estimator are violated for cells with a small effective size.

Furthermore, the estimates from S_1 are considered as register counts, the correction factors g_i of the regressions are not incorporated in the variance formulas, and the stochastic properties of the regression coefficients **B** are neglected. These approximations are of course not made in the simulation variance. Only for a few table cells with a sufficiently low cve, the estimated standard deviation indeed gives a correct 95% confidence interval.

5.2.5 Variance of aggregate

The main conclusions in this first simulation study can be summarized as that repeated weighting does not lead to *excessive* biases of table estimates, that the variances of the target table estimates are rather high from the start due to small effective sample sizes, that repeated weighting decreases these variances, but only slightly in this particular case, and that the variance is underestimated with the variance formulas, especially for cells with a very small effective size. It is instructive to calculate variances for a less-detailed table, where variances are expected to be lower, and the variance formulas are expected to perform better. Therefore, in this subsection, we focus on the tables

T_3 = [sex × educational level] × 1

T_4 = [sex × educational level] × monthly salary,

which are aggregates of the tables in the previous sections. Note that the frequency table T_3 is estimated from data block 3, and T_4 is calibrated on this frequency table. Average monthly salaries are again calculated by taking the ratio of T_4 and T_3. The bias and simulation variance of the average monthly salaries are shown in Table 15 in Appendix E, for both the RW- and the SW-estimate. Repeated weighting obviously has a positive effect on the accuracy of these estimates: the standard deviation of the RW-estimates is significantly lower than the standard deviation of the SW-estimates.

The standard deviations of the RW-estimates can again be estimated, and coverage probabilities can be calculated. The results are not shown here. Only for one cell (the one with the smallest cell size, that is, the cell Male, Upper secondary A/B), the coverage probability turns out to deviate significantly from 95%, showing that the variance formulas do perform rather well when cell sizes are sufficiently large.

5.3 Results Simulation 2

In the second simulation, M = 600 runs are performed according to the two set-ups 2A and 2B described in Section 5.1. In Simulation 2A, samples are drawn with equal inclusion probabilities. In Simulation 2B, the samples are drawn with unequal inclusion probabilities. In particular, for sample S_2 the internal variation of weights is rather large. Therefore we use the effective sample sizes instead of the real sample size to calculate the λ-factors of both surveys.

In each run $m \in \{1, 2, \ldots, M\}$, table estimates for T_1 and T_2 are again obtained, both with repeated weighting (with and without the splitting-up procedure), and without repeated weighting. From these estimates, average monthly salaries are obtained. Variances for $\hat{\mathbf{Y}}_{T_1,m}^{RW}$ and $\hat{\mathbf{Y}}_{T_2,m}^{RW}$ are estimated, using the variance formulas derived in Section 3.4.4. The variance of the p-th cell in the ratio $\hat{\mathbf{Y}}_{T_2/T_1,m}^{RW}$ is estimated using Eq. (7).

5.3.1 Cell size

Both tables T_1 and T_2 are estimated from data block 3, or equivalently, from the data of S_2. In Simulation 2A, this data block consists of exactly 100,000 records. The effective size is equal to the real size in this case. In Simulation 2B the real sample size is approximately 99,000 records, and the effective sample size is about 18,000. Table 7 and Table 8 in Appendix A show an overview of the real and effective average cell sizes of the target table [sex × working hours per week × education level] in Simulation 2A and Simulation 2B, respectively. The cell sizes are much larger than in the first simulation; especially the worst cells have profited more than proportional from the increase in micro data used for the estimates in this simulation. We thus expect that biases and variances of the estimates are much smaller. Furthermore, we expect that the estimated variance resembles the simulation variance much better.

5.3.2 Bias

Given the results of the $M = 600$ simulations, the biases of the estimated average salaries $\hat{\mathbf{Y}}_{T_2/T_1}^{RW}$, $\hat{\mathbf{Y}}_{T_2/T_1}^{RW+}$ and $\hat{\mathbf{Y}}_{T_2/T_1}^{SW}$ are calculated with Equation (8). The results are shown in Table 10 and Table 11 in Appendix B for Simulation 2A and Simulation 2B, respectively. Indeed, the biases are much smaller than those in Table 9; the grayish-colored cells again show which biases are significant. For Simulation 2A, the biases do not deviate significantly from zero for most cells in the table, as is also expected for such a ‘clean’ simulation. In Simulation 2B, the biases seem to increase somewhat, but not much, due to the reweighting strategy compared to the biases of the SW-estimator. Again, the difference in biases of the with and without splitting-up RW-estimates is small.

5.3.3 Simulation variance

Next, the simulation variances for the three estimators are computed according to Equation (9). The relative standard deviations of the average monthly salaries are shown in Table 13 and Table 14 in Appendix C, for Simulation 2A and 2B, respectively. Indeed these standard deviations are much smaller than the corresponding ones from Simulation 1 in Table 12. Figure 5 and Figure 6 in Appendix C show again the relative differences between the standard deviations of th SW- and RW-estimates, and the relative difference between the standard deviations of the RW- and RW+-estimator.

From Figure 5a, we again conclude that for Simulation 2A repeated weighting leads to a (significant) gain in accuracy for most cells in the table. However, the gain is small, much smaller than in Simulation 1. This is due to the fact that data block 2 in this case is not so much larger than data block 3, such that calibration on margins from data block 2 does not lead to much gain in accuracy. Figure 5b shows that there is no significant difference between repeated weighting with and without splitting-up.

The results of Simulation 2B shown in Table 14 and Figure 6 are similar to those obtained in Simulation 1. There is a (significant) gain in accuracy due to repeated weighting, but the gain is less when the splitting-up procedure is used. Note that Simulation 2B in fact is similar to Simulation 1: the effective size of S_1 is about four times larger than the effective size of S_2, and therefore almost a 'register' compared to sample S_2. The difference between the results of Simulation 1 and Simulation 2B is therefore mainly caused by the fact that the effective size of the data block from which the target tables are estimated is about a factor of four larger in the latter case.

5.3.4 Variance estimate

Finally, the estimated variances for the RW-estimator are used to estimate 95% confidence intervals, and again we check whether the true population average salaries lie within this interval, see Equation (10). The coverage probabilities P_p of the estimated 95% confidence interval for the p-th cell in the target table are plotted versus the quality of the cell (measured in terms of the estimated coefficient of variation (cve)) in Figure 8 and Figure 9 in Appendix D for Simulation 2A and 2B, respectively.

For both simulations we find that the coverage probabilities drop below 95% when the quality of the cell (or, equivalently, the effective size) becomes less. In Simulation 2A, we see that variances are estimated well for most cells. Only a few cells have a coverage probability that is significantly less than 95%, showing that in the case of sufficiently large sample sizes the variance formulas work well. In Simulation 2B, the effective cell sizes are much smaller, and the variance formulas start to underestimate the true variance.

In Figure 10 in Appendix D, we have plotted the coverage probability of Simulation 1, 2A, and 2B versus the effective cell size for each cell in the target table. The figure is truncated at an effective cell size of 400 records. From the figure it is clear that the different simulations behave more or less equivalent. For cells with a sufficiently large effective size, say more than 100 records, the variances are estimated well with the variance formulas. But for cells with a smaller effective size, the estimated variance underestimates the true variance. Apparently, the assumptions on which the variance estimator are based, are not satisfied anymore. Further research is required to see how the variance formulas can be improved in the case of small cell sizes.

6. Discussion

Statistics Netherlands has constructed a social statistical database (SSD) in which data from administrative sources are linked to each other and to survey data. From this database, a huge amount of information related to social statistics can be obtained, either by mere counting, or by using any general estimation technique. However, an important issue at Statistics Netherlands is that estimates should be – as much as possible, at least – numerically consistent. At Statistics Netherlands, a new estimation strategy is developed with which well-defined table sets can be estimated consistently. This estimation strategy is called 'repeated weighting', since it is based on a repeated use of the calibration properties of the regression estimator. In this paper we gave an overview of the method of repeated weighting.

Repeated weighting is applicable in case of relatively simple and well-defined table sets, for which consistency ís, and timeliness is not important. Apart from numerical consistency, an important additional advantage of the estimation method is that the accuracy of the estimates is improved due to a better use of auxiliary information. The method is, however, not suitable when

- The tables in the table set relate to different object types (for instance, persons and households) and characteristics of one object type are related to characteristics of another object type.
- There are many edit rules between the variables in the table set.
- There are many quantification variables, especially when these variables are related to classification variables with very fine classifications.
- The classification variables in the table set cannot be squeezed into a hierarchical structure.
- The tables in the table set relate to several different subpopulations of the target population, and these subpopulations, or their complements, are small.
- Certain subpopulations of the target population are heavily underrepresented (or worse, not present at all) in one or more of the surveys used in the estimation process, or more general, when too little survey data is available.

In general, with repeated weighting it will certainly *not* be possible to obtain numerical consistency between *all* estimates from the SSD. It will however ensure numerical consistency *within* a well-defined table set. In addition, there will be consistency *between* table sets when it comes to population totals of register variables, even if these population totals of register variables are margins of (estimated) tables which also contain survey variables. Moreover, consistency between table sets is guaranteed if these table sets relate to the same target population, and are estimated with the splitting-up procedure, using the same micro data and final block weights for each variable that occurs in both table sets. When combining data from different sources, repeated weighting yields obviously much more numerical consistency between estimates than other, more simple, estimation procedures.

Depending on the particular purposes of some publication, and the estimates one wants to make for this publication, one can decide to use repeated weighting, or any other suitable estimation strategy. At Statistics Netherlands, two table sets have thus far successfully been estimated using repeated weighting. The first table set relates to the structure of earnings survey. The second table set involves search behavior on the labor market of persons on social welfare. At present, repeated weighting is also used to meet Eurostat's demands with respect to the 2001 Census, see Van der Laan (2000).

Appendix A: Cell size

Table 6. Real and effective average cell sizes in the target table of Simulation 1.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	n_{cell}	29	101	84	239	1818	43	180	296	566	478
	$n_{\text{eff., cell}}$	3	9	9	27	178	4	17	32	55	46
Lower secondary	n_{cell}	46	157	126	377	3025	71	270	481	905	768
	$n_{\text{eff., cell}}$	4	14	12	41	292	7	25	50	91	73
Upper secondary C	n_{cell}	35	121	83	172	1115	75	276	382	714	650
	$n_{\text{eff., cell}}$	3	10	8	18	106	7	24	39	69	61
Upper secondary A/B	n_{cell}	27	91	60	132	896	56	189	250	479	517
	$n_{\text{eff., cell}}$	2	8	6	14	86	5	16	25	46	47
Tertiary C	n_{cell}	80	273	241	859	7549	200	759	1442	2840	2645
	$n_{\text{eff., cell}}$	7	24	24	92	736	18	67	146	290	248
Tertiary B	n_{cell}	27	96	98	394	3509	93	369	749	1495	1333
	$n_{\text{eff., cell}}$	3	8	10	43	343	8	34	77	146	128
Tertiary A and higher	n_{cell}	12	40	47	210	1904	26	106	242	493	455
	$n_{\text{eff., cell}}$	1	3	5	24	190	3	9	24	49	43

Table 7. Real (and effective) cell sizes in the target table of Simulation 2A.

Gender x Working hours	Male					Female				
Education Level	<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	101	309	199	577	4332	135	486	566	1105	936
Lower secondary	153	470	322	850	7400	205	687	881	1649	1505
Upper secondary C	135	381	221	456	2985	240	758	720	1376	1361
Upper secondary A/B	98	268	150	354	2419	173	517	488	941	1109
Tertiary C	276	859	578	1899	17806	619	2006	2647	5124	5226
Tertiary B	92	309	245	826	8006	295	982	1377	2711	2513
Tertiary A and higher	51	142	119	437	4299	66	258	454	888	868

Table 8. Real and effective cell sizes in the target table of Simulation 2B.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	n_{cell}	71	244	201	577	4302	100	414	655	1268	1083
	$n_{\text{eff., cell}}$	12	36	44	95	857	20	75	137	237	190
Lower secondary	n_{cell}	109	368	295	875	7033	159	601	1042	1977	1682
	$n_{\text{eff., cell}}$	19	57	50	155	1335	29	107	213	385	289
Upper secondary C	n_{cell}	83	290	195	408	2630	169	616	835	1568	1445
	$n_{\text{eff., cell}}$	10	39	31	70	484	27	91	171	297	224
Upper secondary A/B	n_{cell}	63	215	141	316	2124	124	421	546	1049	1142
	$n_{\text{eff., cell}}$	8	31	23	56	381	19	63	107	189	183
Tertiary C	n_{cell}	186	633	552	1975	17404	447	1691	3116	6167	5788
	$n_{\text{eff., cell}}$	28	88	91	345	3349	71	279	594	1198	1021
Tertiary B	n_{cell}	66	224	230	901	8090	208	818	1625	3255	2919
	$n_{\text{eff., cell}}$	11	29	47	172	1553	29	139	315	603	554
Tertiary A and higher	n_{cell}	28	94	108	483	4385	57	232	524	1072	995
	$n_{\text{eff., cell}}$	3	14	16	81	872	12	42	88	207	197

Appendix B: Bias

Table 9. Relative bias of the average monthly salary in Simulation 1.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	SW	-2,8%	-0,6%	-0,2%	1,1%	0,2%	-1,8%	1,5%	0,7%	0,7%	0,2%
	RW	2,1%	2,3%	0,6%	0,9%	0,0%	1,5%	1,8%	0,7%	0,5%	0,2%
	RW+	0,9%	2,0%	0,6%	1,0%	0,0%	0,5%	1,4%	0,5%	0,4%	0,0%
Lower secondary	SW	-2,1%	-0,9%	0,0%	0,8%	0,4%	-3,3%	0,0%	0,5%	0,4%	0,6%
	RW	-1,0%	1,1%	0,7%	0,5%	0,3%	0,7%	0,0%	0,1%	0,1%	0,3%
	RW+	-1,5%	1,0%	0,6%	0,6%	0,3%	0,2%	-0,3%	0,2%	0,2%	0,4%
Upper secondary C	SW	-2,1%	-2,8%	-2,0%	0,4%	0,4%	-3,1%	1,6%	-0,2%	-0,1%	-0,1%
	RW	-1,4%	-0,7%	-1,3%	0,2%	0,1%	-4,8%	1,1%	-0,4%	-0,4%	-0,1%
	RW+	-1,0%	-0,7%	-1,3%	0,4%	0,1%	-4,7%	1,3%	-0,3%	-0,3%	-0,3%
Upper secondary A/B	SW	-1,8%	-0,2%	-1,8%	1,0%	-0,3%	2,5%	-0,6%	-0,3%	0,8%	1,1%
	RW	-0,1%	2,2%	-0,8%	0,7%	-0,4%	4,1%	-0,7%	-0,3%	0,5%	0,8%
	RW+	-0,5%	2,0%	-1,0%	0,8%	-0,4%	3,9%	-0,6%	-0,3%	0,5%	0,9%
Tertiary C	SW	-1,8%	-3,0%	-0,4%	0,5%	0,2%	-1,5%	0,8%	0,2%	0,3%	0,4%
	RW	-0,8%	-2,1%	0,8%	0,3%	0,0%	-0,6%	0,8%	0,1%	0,1%	0,2%
	RW+	-0,6%	-1,8%	0,8%	0,3%	0,1%	-0,1%	0,7%	0,1%	0,1%	0,2%
Tertiary B	SW	1,7%	1,2%	0,6%	0,6%	0,1%	0,0%	-0,5%	0,6%	0,3%	-0,3%
	RW	5,8%	4,7%	1,5%	0,4%	-0,1%	0,2%	-1,0%	0,6%	0,0%	-0,3%
	RW+	4,9%	4,8%	1,4%	0,4%	-0,1%	-0,3%	-0,8%	0,6%	0,0%	-0,3%
Tertiary A and higher	SW	4,9%	0,4%	8,7%	1,7%	-0,2%	3,7%	-0,4%	0,6%	0,6%	0,0%
	RW	10,7%	3,8%	9,5%	1,4%	-0,4%	9,1%	-2,0%	0,6%	0,4%	-0,2%
	RW+	9,3%	3,2%	9,8%	1,5%	-0,4%	8,1%	-2,0%	0,5%	0,3%	-0,1%

Table 10. Relative bias of the average monthly salary in Simulation 2A.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	SW	-0,31%	-0,16%	0,29%	-0,02%	-0,03%	-0,21%	0,06%	0,00%	0,04%	0,05%
	RW	-0,19%	-0,16%	0,24%	0,01%	-0,02%	-0,17%	0,03%	-0,02%	0,02%	0,06%
	RW+	-0,21%	-0,16%	0,24%	0,01%	-0,02%	-0,21%	0,03%	-0,02%	0,02%	0,06%
Lower secondary	SW	-0,11%	0,04%	-0,40%	-0,04%	0,00%	0,60%	-0,06%	0,05%	0,00%	-0,06%
	RW	0,10%	0,07%	-0,41%	-0,01%	0,01%	0,61%	-0,08%	0,02%	-0,02%	-0,05%
	RW+	0,09%	0,06%	-0,41%	-0,02%	0,01%	0,59%	-0,09%	0,02%	-0,02%	-0,05%
Upper secondary C	SW	-0,07%	-0,27%	0,18%	0,00%	0,01%	-0,29%	-0,01%	0,06%	0,04%	-0,04%
	RW	0,07%	-0,24%	0,18%	0,03%	0,02%	-0,27%	-0,04%	0,04%	0,02%	-0,03%
	RW+	0,06%	-0,24%	0,18%	0,03%	0,02%	-0,26%	-0,03%	0,04%	0,02%	-0,03%
Upper secondary A/B	SW	-0,36%	-0,23%	-0,29%	0,09%	-0,10%	0,17%	0,28%	0,06%	0,00%	-0,10%
	RW	-0,22%	-0,20%	-0,30%	0,11%	-0,08%	0,22%	0,26%	0,04%	-0,03%	-0,08%
	RW+	-0,23%	-0,20%	-0,30%	0,12%	-0,08%	0,21%	0,26%	0,04%	-0,03%	-0,09%
Tertiary C	SW	-0,07%	-0,20%	0,10%	0,02%	-0,03%	-0,05%	0,02%	0,05%	0,03%	-0,03%
	RW	0,09%	-0,18%	0,10%	0,04%	-0,01%	-0,03%	0,00%	0,03%	0,01%	-0,02%
	RW+	0,10%	-0,18%	0,09%	0,04%	-0,01%	-0,01%	-0,01%	0,03%	0,01%	-0,02%
Tertiary B	SW	-0,62%	-0,09%	-0,01%	0,02%	0,02%	-0,08%	0,04%	-0,06%	0,01%	0,00%
	RW	-0,36%	-0,05%	-0,03%	0,04%	0,03%	-0,10%	0,01%	-0,08%	-0,01%	0,01%
	RW+	-0,36%	-0,05%	-0,03%	0,05%	0,03%	-0,09%	0,02%	-0,08%	-0,01%	0,01%
Tertiary A and higher	SW	-0,07%	0,03%	0,07%	0,13%	0,00%	-0,02%	0,19%	-0,05%	0,13%	0,05%
	RW	0,04%	0,07%	0,02%	0,16%	0,01%	0,03%	0,10%	-0,07%	0,11%	0,06%
	RW+	0,03%	0,09%	0,03%	0,15%	0,01%	0,01%	0,12%	-0,06%	0,11%	0,06%

Table 11. Relative bias of average monthly salary in Simulation 2B.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	SW	0,50%	-0,59%	0,64%	0,31%	0,04%	-0,30%	-0,13%	0,12%	-0,07%	0,15%
	RW	1,85%	-0,45%	0,02%	0,14%	-0,14%	0,89%	-0,07%	-0,02%	-0,23%	0,06%
	RW+	1,49%	-0,42%	0,09%	0,15%	-0,14%	0,53%	-0,10%	-0,03%	-0,25%	0,03%
Lower secondary	SW	-1,31%	-0,55%	-0,13%	-0,18%	0,05%	0,39%	-0,16%	0,12%	0,24%	-0,04%
	RW	-1,61%	-0,34%	-0,51%	-0,35%	-0,12%	0,98%	-0,09%	-0,12%	0,05%	-0,18%
	RW+	-1,40%	-0,43%	-0,48%	-0,36%	-0,11%	0,92%	-0,17%	-0,08%	0,04%	-0,18%
Upper secondary C	SW	-1,51%	0,49%	0,46%	0,36%	-0,07%	-0,39%	0,05%	0,11%	0,17%	0,04%
	RW	-1,62%	0,91%	0,18%	0,19%	-0,27%	-0,70%	-0,07%	-0,11%	0,01%	-0,09%
	RW+	-1,54%	0,82%	0,14%	0,20%	-0,26%	-0,44%	-0,07%	-0,10%	0,04%	-0,05%
Upper secondary A/B	SW	-0,47%	0,29%	0,17%	0,52%	-0,09%	0,05%	0,73%	-0,08%	0,10%	0,43%
	RW	0,39%	0,71%	-0,21%	0,33%	-0,24%	0,79%	0,64%	-0,23%	-0,07%	0,27%
	RW+	0,21%	0,63%	-0,24%	0,35%	-0,23%	0,61%	0,63%	-0,24%	-0,05%	0,31%
Tertiary C	SW	-0,47%	-0,54%	0,47%	0,06%	0,04%	-0,77%	-0,07%	0,08%	-0,03%	0,03%
	RW	-0,51%	-0,62%	0,17%	-0,07%	-0,10%	0,17%	-0,03%	-0,08%	-0,19%	-0,10%
	RW+	-0,47%	-0,52%	0,14%	-0,08%	-0,10%	0,25%	-0,01%	-0,09%	-0,19%	-0,11%
Tertiary B	SW	1,00%	-0,56%	1,37%	0,02%	0,03%	0,09%	0,08%	0,11%	-0,04%	-0,06%
	RW	2,71%	0,03%	0,93%	-0,11%	-0,14%	-0,52%	0,02%	0,00%	-0,21%	-0,17%
	RW+	2,43%	0,03%	0,92%	-0,09%	-0,14%	-0,55%	0,03%	-0,01%	-0,21%	-0,18%
Tertiary A and higher	SW	2,03%	1,02%	1,23%	0,48%	0,03%	0,53%	-0,46%	0,10%	0,19%	0,09%
	RW	3,23%	1,46%	0,63%	0,35%	-0,14%	1,72%	-0,90%	-0,10%	0,03%	-0,03%
	RW+	2,92%	1,32%	0,67%	0,34%	-0,13%	1,35%	-0,85%	-0,09%	0,03%	-0,03%

Appendix C: Standard deviation

Table 12. Relative standard deviations of average monthly salaries in Simulation 1.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	SW	29,2%	27,5%	24,0%	11,1%	3,9%	31,7%	12,9%	7,9%	5,4%	5,8%
	RW	33,1%	27,0%	21,2%	10,3%	3,6%	36,3%	12,9%	7,5%	5,1%	5,7%
	RW+	34,9%	27,0%	22,2%	10,5%	3,6%	38,2%	13,3%	7,6%	5,2%	5,8%
Lower secondary	SW	37,9%	27,5%	17,5%	8,7%	2,8%	24,0%	11,2%	7,6%	4,8%	5,6%
	RW	28,1%	21,7%	15,5%	8,0%	2,7%	25,5%	10,3%	6,1%	4,3%	4,6%
	RW+	30,3%	22,7%	16,0%	8,1%	2,6%	27,2%	10,8%	6,4%	4,4%	5,0%
Upper secondary C	SW	44,6%	24,7%	16,9%	12,2%	5,7%	42,5%	17,1%	8,6%	5,1%	5,2%
	RW	33,3%	23,6%	16,7%	11,9%	4,9%	27,5%	14,9%	7,7%	4,9%	4,8%
	RW+	38,4%	23,9%	17,3%	12,2%	4,9%	29,3%	15,7%	7,8%	4,9%	5,1%
Upper secondary A/B	SW	41,2%	25,7%	20,7%	13,5%	5,1%	41,6%	20,1%	9,2%	6,1%	7,4%
	RW	40,3%	25,5%	20,9%	12,7%	4,9%	37,7%	18,8%	9,0%	5,9%	6,4%
	RW+	40,8%	25,9%	21,4%	12,9%	4,8%	40,2%	19,7%	9,0%	6,0%	6,8%
Tertiary C	SW	30,8%	21,6%	11,8%	5,6%	1,6%	21,3%	8,1%	4,0%	2,4%	2,9%
	RW	21,2%	14,2%	10,2%	4,4%	1,3%	15,2%	6,6%	3,2%	2,0%	2,0%
	RW+	23,7%	14,7%	10,4%	4,5%	1,3%	16,5%	7,1%	3,2%	2,0%	2,2%
Tertiary B	SW	39,3%	23,2%	17,8%	7,1%	2,4%	39,2%	13,0%	5,2%	3,5%	3,4%
	RW	39,1%	22,7%	16,5%	6,7%	2,2%	22,5%	10,8%	4,7%	3,1%	3,1%
	RW+	42,9%	23,4%	16,7%	6,9%	2,2%	23,5%	11,4%	4,8%	3,2%	3,2%
Tertiary A and higher	SW	40,7%	27,8%	25,1%	10,1%	3,3%	27,1%	24,6%	9,3%	5,4%	6,6%
	RW	67,5%	28,0%	24,4%	9,4%	3,2%	35,5%	14,7%	9,0%	5,3%	5,6%
	RW+	67,6%	28,7%	25,5%	9,8%	3,1%	36,9%	16,6%	9,1%	5,4%	6,2%

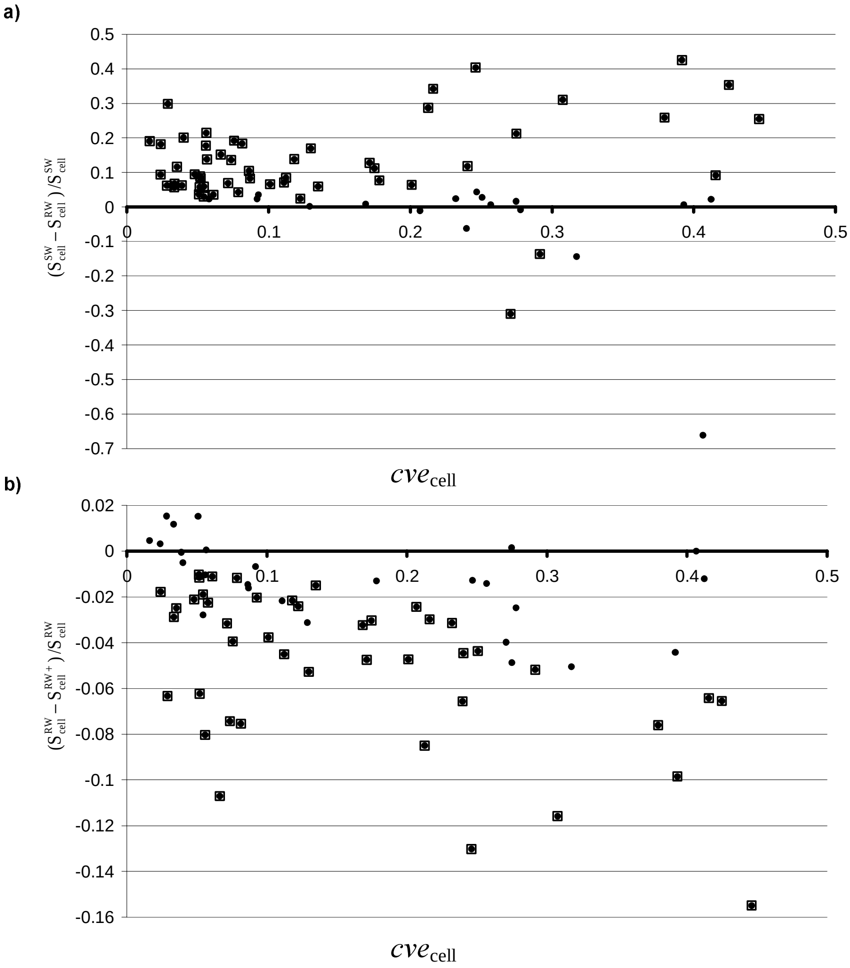

Figure 4. Relative difference between standard deviations of the SW- and RW-estimator (a), and the RW- and RW+-estimator (b) versus the quality of each cell in Simulation 1.

Table 13. Relative standard deviations of average monthly salaries in Simulation 2A.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	SW	7,85%	5,36%	5,78%	2,74%	0,80%	7,00%	3,27%	1,98%	1,33%	1,44%
	RW	7,73%	5,17%	5,48%	2,68%	0,79%	6,89%	3,20%	1,96%	1,30%	1,42%
	RW+	7,73%	5,17%	5,49%	2,68%	0,79%	6,88%	3,21%	1,96%	1,30%	1,42%
Lower secondary	SW	6,91%	4,30%	4,24%	2,27%	0,58%	6,24%	2,47%	1,59%	1,07%	1,12%
	RW	6,85%	4,14%	4,11%	2,21%	0,56%	5,99%	2,41%	1,54%	1,03%	1,08%
	RW+	6,85%	4,13%	4,10%	2,21%	0,56%	6,00%	2,42%	1,54%	1,03%	1,09%
Upper secondary C	SW	8,26%	4,58%	5,27%	3,11%	1,09%	6,88%	2,92%	1,95%	1,29%	1,22%
	RW	7,96%	4,52%	5,23%	3,05%	1,08%	6,85%	2,82%	1,92%	1,28%	1,20%
	RW+	7,93%	4,51%	5,24%	3,05%	1,08%	6,86%	2,83%	1,91%	1,27%	1,20%
Upper secondary A/B	SW	8,34%	4,99%	6,68%	3,47%	1,05%	7,43%	3,33%	2,34%	1,55%	1,33%
	RW	8,22%	4,96%	6,60%	3,41%	1,05%	7,34%	3,28%	2,31%	1,52%	1,31%
	RW+	8,21%	4,96%	6,59%	3,42%	1,05%	7,35%	3,28%	2,32%	1,52%	1,31%
Tertiary C	SW	5,48%	3,26%	2,75%	1,36%	0,34%	3,55%	1,42%	0,89%	0,59%	0,62%
	RW	4,94%	2,94%	2,63%	1,22%	0,33%	3,23%	1,30%	0,80%	0,54%	0,55%
	RW+	4,93%	2,94%	2,63%	1,22%	0,33%	3,23%	1,30%	0,80%	0,54%	0,55%
Tertiary B	SW	9,23%	4,25%	4,60%	1,89%	0,54%	5,42%	2,10%	1,08%	0,83%	0,84%
	RW	9,21%	4,13%	4,47%	1,84%	0,52%	5,11%	1,99%	1,03%	0,78%	0,81%
	RW+	9,22%	4,13%	4,46%	1,84%	0,52%	5,09%	2,00%	1,03%	0,78%	0,81%
Tertiary A and higher	SW	10,23%	6,73%	6,27%	2,49%	0,77%	9,18%	4,54%	1,92%	1,38%	1,31%
	RW	10,06%	6,64%	6,09%	2,43%	0,75%	9,22%	4,46%	1,90%	1,36%	1,31%
	RW+	10,07%	6,63%	6,10%	2,43%	0,76%	9,24%	4,47%	1,91%	1,36%	1,31%

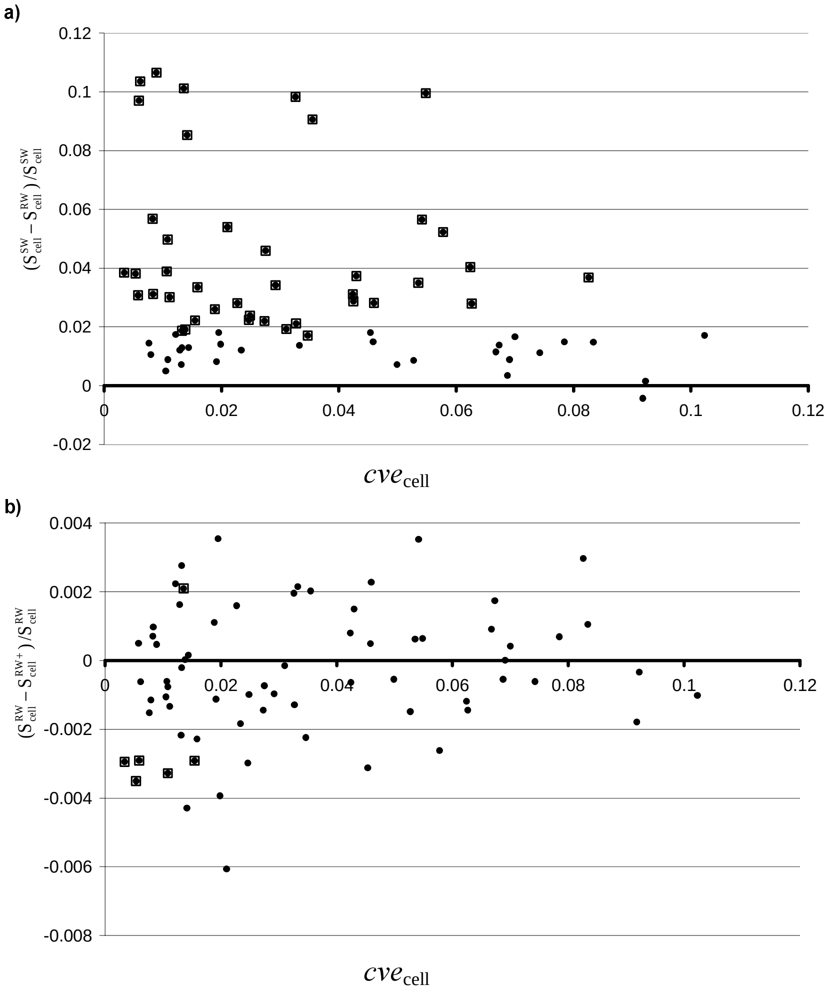

Figure 5. Relative difference between standard deviations of the SW- and RW-estimator (a), and the RW- and RW+-estimator (b) versus the quality of each cell in Simulation 2A.

Table 14. Relative standard deviations of average monthly salaries in Simulation 2B.

Gender x Working hours		Male					Female				
Education Level		<4	4-12	12-20	20-35	35+	<4	4-12	12-20	20-35	35+
Primary or less	SW	18,81%	14,03%	13,12%	5,77%	1,75%	15,81%	6,32%	3,89%	2,47%	2,84%
	RW	19,01%	13,18%	11,87%	5,52%	1,72%	16,39%	6,07%	3,79%	2,42%	2,81%
	RW+	18,93%	13,18%	12,00%	5,53%	1,71%	16,45%	6,13%	3,84%	2,43%	2,83%
Lower secondary	SW	23,92%	12,99%	8,73%	4,08%	1,29%	18,65%	4,83%	3,90%	2,18%	2,67%
	RW	19,38%	12,10%	8,10%	3,87%	1,21%	17,72%	4,73%	3,48%	2,05%	2,53%
	RW+	20,40%	12,12%	8,10%	3,88%	1,21%	17,76%	4,80%	3,54%	2,06%	2,54%
Upper secondary C	SW	24,13%	12,68%	10,31%	6,06%	2,43%	25,48%	7,86%	4,49%	2,45%	3,11%
	RW	20,64%	12,39%	9,92%	5,99%	2,33%	21,09%	7,32%	4,26%	2,31%	2,89%
	RW+	21,54%	12,36%	9,98%	6,01%	2,32%	21,79%	7,37%	4,28%	2,32%	2,92%
Upper secondary A/B	SW	22,85%	12,19%	15,03%	6,92%	2,42%	21,08%	11,20%	4,26%	2,91%	3,41%
	RW	22,32%	12,02%	14,46%	6,69%	2,38%	20,51%	10,44%	4,17%	2,82%	3,14%
	RW+	22,59%	12,04%	14,60%	6,72%	2,37%	20,58%	10,58%	4,18%	2,83%	3,20%
Tertiary C	SW	16,86%	10,26%	6,76%	2,68%	0,78%	10,59%	3,50%	1,97%	1,20%	1,44%
	RW	12,61%	7,84%	5,75%	2,31%	0,67%	9,00%	3,10%	1,64%	1,01%	1,14%
	RW+	13,12%	7,87%	5,80%	2,31%	0,67%	9,13%	3,15%	1,65%	1,02%	1,15%
Tertiary B	SW	21,28%	11,46%	9,31%	3,82%	1,22%	20,37%	6,45%	2,29%	1,64%	1,74%
	RW	21,36%	11,06%	8,74%	3,63%	1,09%	15,51%	5,59%	2,18%	1,54%	1,60%
	RW+	21,76%	11,23%	8,76%	3,66%	1,09%	15,81%	5,68%	2,19%	1,55%	1,63%
Tertiary A and higher	SW	25,95%	18,38%	14,40%	5,09%	1,86%	16,21%	12,20%	5,45%	2,69%	3,17%
	RW	26,83%	18,07%	13,81%	4,98%	1,74%	17,00%	10,41%	5,01%	2,67%	3,02%
	RW+	26,34%	18,07%	13,77%	5,01%	1,74%	17,01%	10,82%	5,07%	2,68%	3,04%

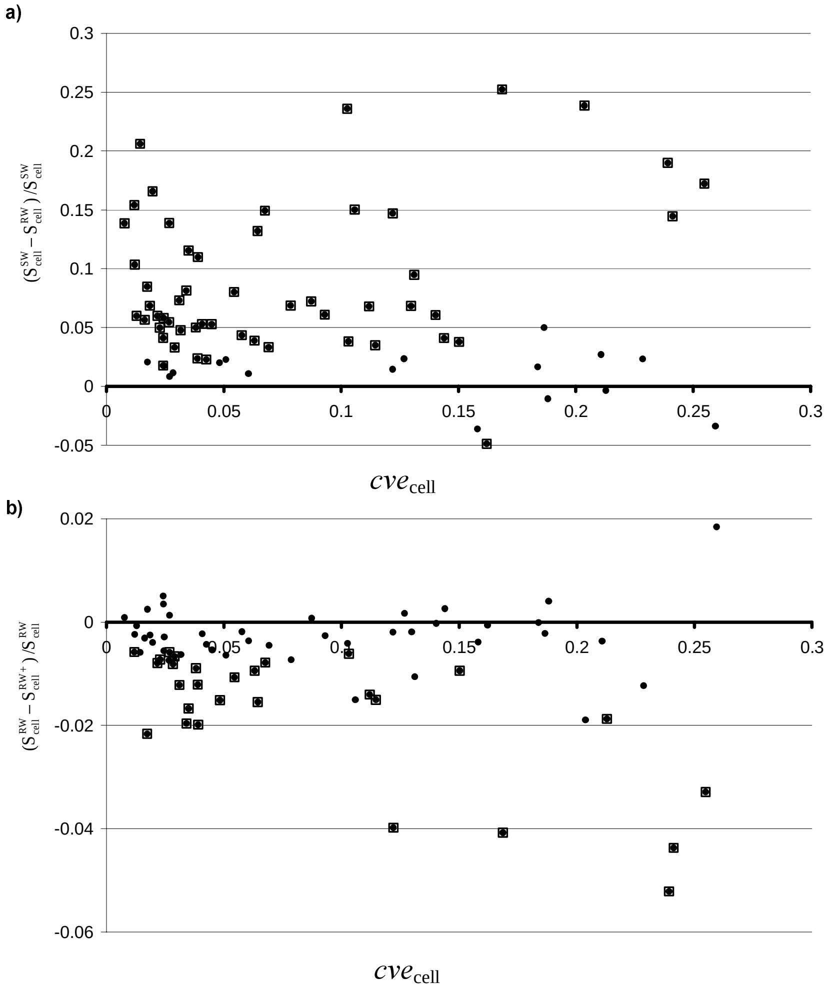

Figure 6. Relative difference between standard deviations of the SW- and RW-estimator (a), and the RW- and RW+-estimator (b) versus the quality of each cell in Simulation 2B.

Appendix D: Confidence interval

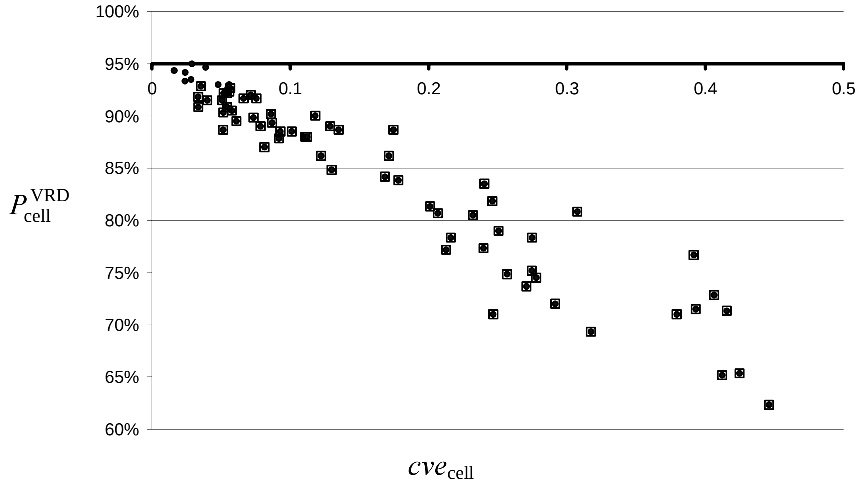

Figure 7. Coverage probability versus the quality of the cell for Simulation 1.

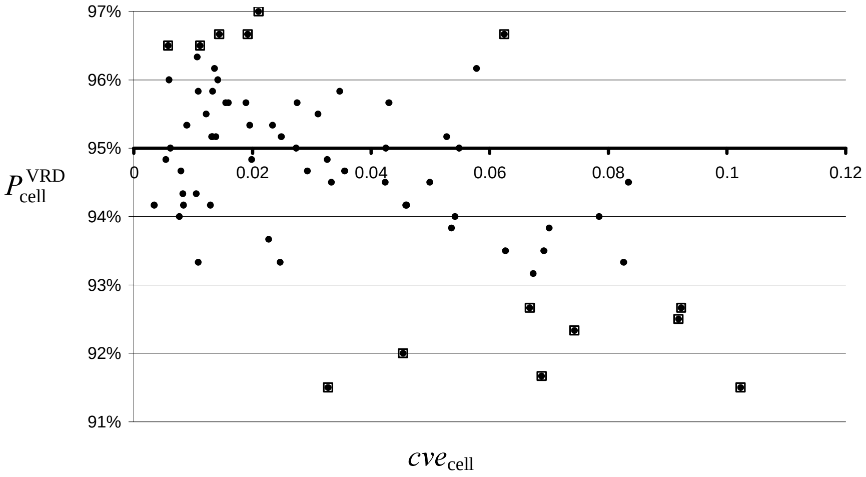

Figure 8. Coverage probability versus the quality of the cell for Simulation 2A.

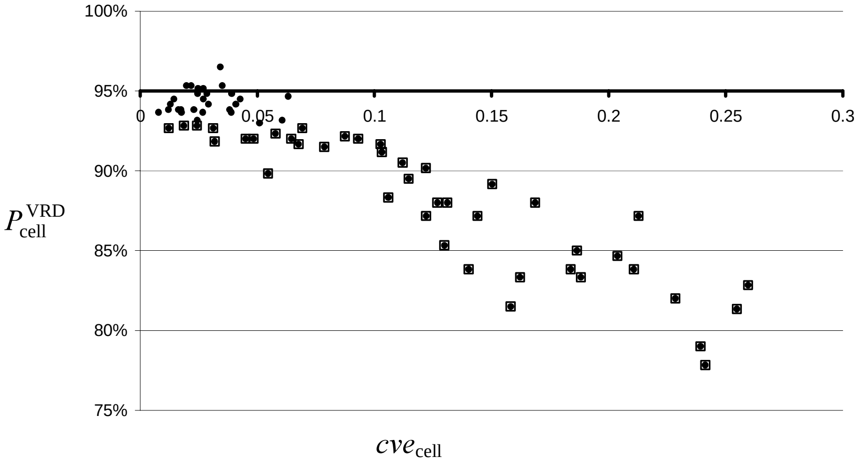

Figure 9. Coverage probability versus the quality of the cell for Simulation 2B.

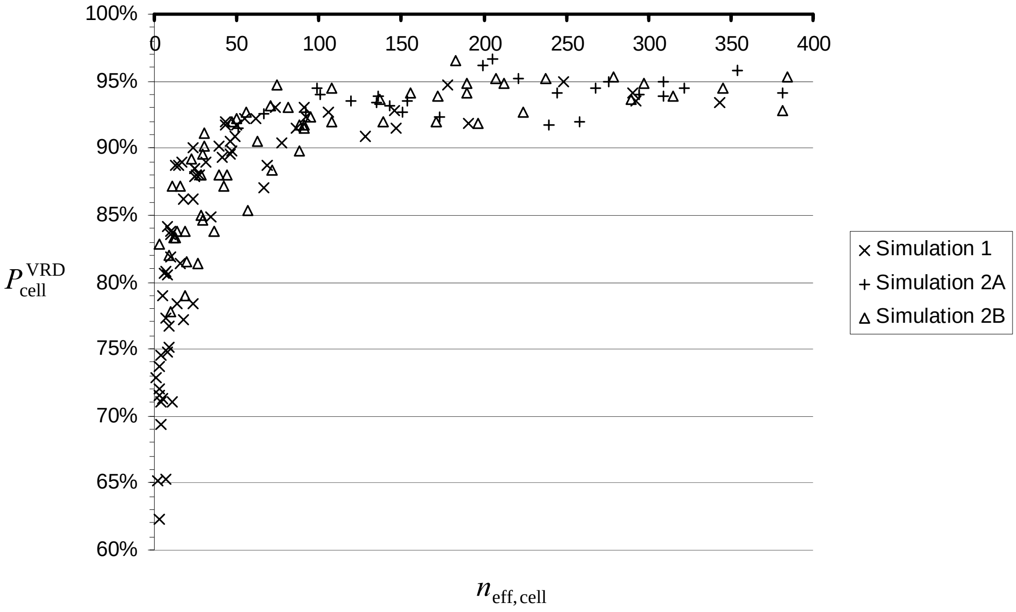

Figure 10. Coverage probability versus effective cell size for all simulations.

Appendix E: Aggregated table

Table 15. Relative bias (left table) and simulation standard deviations (right table) of average monthly salaries.

Sex		Male	Female
Education level			
Primary or less	SW	0.35%	0.04%
	RW	-0.04%	-0.07%
Lower secondary	SW	0.55%	0.61%
	RW	0.12%	0.04%
Upper secondary C	SW	-0.32%	-0.73%
	RW	0.31%	-0.28%
Upper secondary A/B	SW	-0.47%	0.56%
	RW	-0.63%	0.11%
Tertiary C	SW	0.25%	0.38%
	RW	0.04%	0.19%
Tertiary B	SW	0.29%	0.38%
	RW	-0.15%	-0.22%
Tertiary A and higher	SW	-0.09%	0.11%
	RW	0.09%	-0.05%

Sex		Male	Female
Education level			
Primary or less	SW	6.7%	7.9%
	RW	3.9%	5.2%
Lower secondary	SW	5.2%	6.9%
	RW	2.8%	4.5%
Upper secondary C	SW	8.9%	6.8%
	RW	5.4%	5.0%
Upper secondary A/B	SW	9.1%	8.8%
	RW	5.5%	6.0%
Tertiary C	SW	2.7%	3.3%
	RW	1.5%	2.0%
Tertiary B	SW	4.5%	4.7%
	RW	2.4%	3.1%
Tertiary A and higher	SW	6.5%	8.9%
	RW	3.4%	5.2%

References

Bakker, B.F.M. (2002). Statistics Netherlands' approach to social statistics: The social statistical database. The Statistics Newsletter, issued by the OECD, October 2002.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Massachusetts.

Boonstra, H.J.H., Brakel, J.A. van den, Knottnerus, P., Nieuwenbroek, N.J., and Renssen, R.H. (2003). DACSEIS deliverable 7.2: A strategy to obtain consistency among tables of survey estimates. Research paper, BPA-no 1133-03-TMO, Statistics Netherlands, Heerlen.

Cassel, C.M., Särndal, C.-E., and Wretman, J.H. (1976). Some results on generalized difference estimation and general regression estimation for finite populations. Biometrika **63**, pp. 615-620.

Deville, J.-C. (1988). Estimation linéaire et redressement sur informations auxiliaires d'enquêtes par sondage. In : A. Monfort and J.J. Laffond (eds), *Essais en l'Honneur d'Edmond Malinvaud*. Paris: Economica, pp. 915-927.

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

Duin, C. van and Snijders, V. (2003). Simulation studies of repeated weighting. Discussion paper, Statistics Netherlands, Voorburg, in preparation.

Estevao, V.M. and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. Journal of Official Statistics **18**, pp. 233-255.

Houbiers, M. (2003). Towards a social statistical database and unified estimates at Statistics Nehterlands. Submitted to Journal of Official Statistics.

Kish, L. (1992). Weighting for unequal P_i. Journal of Offcial Statistics **8**, pp. 183-200.

Knottnerus, P. and Wiegert, R. (2002). DACSEIS deliverable 7.1: Questionaire on the use of register data for the labor force surveys. Research paper, BPA-no 1853-02-TMO, Statistics Netherlands, Voorburg.

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. Springer-Verlag, New York.

Kooiman, P. (1998). Mass imputation: Why not!?. Research paper, BPA-no 8792-98-RSM, Statistics Netherlands, Voorburg (in Dutch).

Kroese, A. H. and Renssen, R.H. (1999). Weighting and imputation at Statistics Netherlands. In: *Proceedings of the IASS Satellite Conference on Small Area Estimation*. Riga, Latvia, pp. 109-120.

Kroese, A.H. and Renssen, R.H. (2000). New applications of old weighting techniques; constructing a consistent set of estimates based on data from different surveys. In: *Proceedings of ICES II*. Buffalo NY: American Statistical Association, pp 831-840.

Laan, P. van der (2000). The 2001 Census in The Netherlands: Integration of registers and surveys. In: *Proceedings Insee-Eurostat seminar on censuses after 2001*. Paris, pp. 39-52.

Lemaître, G. and Dufour, J. (1987). An integrated method for weighting persons and families. Survey Methodology **13**, pp. 199-207.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. Journal of Official Statistics **15**, pp. 305-327.

Lundström, S. and Särndal, C.-E. (2002). *Estimation in the Presence of Nonresponse and Frame Imperfections*. Statistics Sweden.

Renssen, R.H. (1998). Use of statistical linking techniques in calibration estimation. Survey Methodology **24**, pp. 171-183.

Renssen, R.H., Kroese, A.H, and Willeboordse, A.J. (2001). Aligning estimates by repeated weighting. Research paper, BPA-no 491-01-TMO, Statistics Netherlands, Heerlen.

Silva, P.L.D.N. and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. Survey Methodology **23**, pp. 23-32.

Statistics Netherlands (2000). Special issue on ‘Integrating Administrative Registers and Household Surveys’. Netherlands Official Statistics **15**, Statistics Netherlands, Voorburg/Heerlen.

Statistics Netherlands (2002). SSD Special. Sociaal-Economische Maandstatistiek, **12**. Statistics Netherlands, Voorburg/Heerlen, pp. 72-92 (in Dutch).

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

Thomson, I. and Kleive Holmøy, A.M. (1998). Combining data from surveys and administrative record systems: The Norwegian experience. International Statistical Review **66**, pp. 201-221.