Journal of Official Statistics, Vol. 20, No. 1, 2004, pp. 55–75

# Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands

*Marianne Houbiers*[1]

Statistics Netherlands aims at improving the accuracy and reliability of estimates by using data from registers and surveys in an optimal way. To this end, Statistics Netherlands is constructing a Social Statistical Database, in which several registers are via a unique key linked to each other, as well as to data from sample surveys. All estimates related to social statistics will be obtained from this database. Many "estimates" can simply be counted from the (combined) registers. Moreover, the presence of ample register data offers far better opportunities for nonresponse correction of estimators from the surveys. Furthermore, by combining data from surveys having variables in common, the accuracy of estimators involving these variables can be improved. In addition, Statistics Netherlands prefers to publish a single figure for each statistical concept. Numerical consistency between estimates may be achieved by using the calibration properties of the regression estimator. In this article, we explain how the social statistical database is constructed, and how reliable, accurate, and numerically consistent tables can be estimated from it. We also mention some theoretical and practical problems, and discuss possible solutions.

*Key words:* Combining registers and surveys; consistent estimates; record linkage; regression estimator; repeated weighting.

## 1. Introduction

In recent years, detailed administrative registers on jobs and social welfare payments have become available at Statistics Netherlands. The availability of these registers allows for an improvement of the quality of estimates made at Statistics Netherlands in three important ways. First, by linking these registers to each other and to the Municipal Base Administration (MBA), detailed and accurate cross-tabulations on many topics concerning mainly social statistics can be obtained by mere counting. Second, these registers can be linked to survey data. With so much information on jobs and social welfare in the registers, the surveys can be corrected for selectivity due to nonresponse – the rates of which are generally quite high in The Netherlands – better than before, when only data from the population administrations from municipalities could be used for these purposes. Third,

[1] Statistics Denmark, Sejrøgade 11, DK-2100 Copenhagen, Denmark. Email: mhs@dst.dk (Formerly at the Department of Methods and Informatics, Statistics Netherlands, PO Box 4000, 2270 JM Voorburg, The Netherlands. The views expressed in the article are those of the author and do not necessarily reflect the policy of Statistics Netherlands.)

**Acknowledgment:** The author thanks Bert Kroese and Robbert Renssen, who are the "founding fathers" of the method of repeated weighting, for various interesting discussions on this topic. Paul Knottnerus is thanked for a continuous and constructive flow of criticism on the content of this article. The Associate Editor and the four referees are thanked for their careful reading of the manuscript, and the many suggestions and comments for improvement.

using the known population totals from register data as auxiliary information, the variances of the estimates from these surveys can be reduced. Clearly, one can greatly benefit from the use of these registers.

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 for legal and privacy reasons, matching 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, to increase the accuracy of estimates, and to secure (some) consistency between estimates from various sources (see for instance, Thomson and Kleive Holmøy 1998).

At Statistics Netherlands, the use of register data for social statistics is envisioned in the following way. By linking the registers for persons, jobs, and social security payments via a unique key to each other, as well as to survey data from sample surveys, a so-called Social Statistical Database (SSD) is constructed (see Statistics Netherlands 2000). All cross-tabulations concerning a certain target population can be subsequently extracted from the relevant part of the SSD, either by counting from the combined registers or by estimating from the survey data. Ideally, for the purpose of variance reduction, for each cross-tabulation all records in the SSD in which the relevant variables are present, are used. That is, an "estimate" is counted from the combined registers if all variables are present in these registers. If that is not the case, the estimate is obtained from a combination of two or more surveys, from one of the surveys, or from the cross-section of two or more surveys, depending on the variables required. In this way, Statistics Netherlands hopes to obtain accurate and reliable estimates from the SSD.

However, since not all estimates will be based on the same set of records, two estimates concerning the same variable may yield different results. For users of the statistical data, this may lead to some confusion about what is the "correct" number. Although the differences are, in principle, merely due to statistical noise, Statistics Netherlands has adopted the so-called one-figure policy, and tries to track down and remove such inconsistencies whenever possible. An important issue at Statistics Netherlands is of course to prevent inconsistencies in estimates in the first place. Therefore, a major goal has been to develop an estimation method that guarantees – as far 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 reaching this goal. Although this new estimation method is not yet applicable in all practical situations, it can be applied in the case of relatively simple and well-defined table sets, yielding consistent estimates.

In principle, mass imputation offers a simple alternative to estimation by weighting to achieve numerical consistency between estimates from the SSD. By using some suitable imputation strategy, all missing fields in the SSD can be imputed. Tables can then simply be "counted" from the resulting complete data set. 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.

In this article we recapitulate how Statistics Netherlands intends to construct the Social Statistical Database, and how accurate, reliable, and consistent estimates can be obtained from it. The article is organized as follows. In Section 2 we focus on the present state of the SSD, and give a specific example of a set of tables that can be estimated from it. In Section 3 we explain how consistent estimates can be obtained from the SSD using the method of repeated weighting. The construction of the SSD and estimating consistent tables from it may seem quite trivial in theory. However, in practice there are numerous problems to tackle. In Section 4 we mention some issues, which may cause complications in the process of constructing the SSD and estimating (consistent) tables from it. In this context, a comparison between the method of repeated weighting and mass imputation would be interesting, but this is beyond the scope of this article. In Section 5 we conclude and summarize.

## 2. Linking Registers and Surveys

For the construction of the SSD, several registers are linked to each other as well as to survey data sets. The registers that are available at present at Statistics Netherlands comprise the Municipal Base Administration (MBA), the jobs register, and the social welfare payments register. The first register contains information on age, gender, ethnicity, place of birth, place of residence, marital status etc., for persons in The Netherlands, except for illegals. The second register contains information (such as size class and business classification) on all jobs in The Netherlands. Via a unique key based 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 register contains information on social welfare payments (such as type of social welfare, amount, and duration of payment). This register can also via the social security number be linked to the persons and the jobs registers. All three registers are so-called volume registers, which means that they contain longitudinal information about all elements in the population during a certain time period. Therefore, they can be linked on any day of the year, thus creating a linked register on a certain reference date.[3] By linking the registers on two or more days of the year, and subsequently averaging, an (approximate) average register is

[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. This 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.

[3] 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 registers, and social welfare agencies. Clearly, the jobs and social welfare registers are not administrative registers 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 that, we refer to them as "registers" in this article.

obtained. Depending on the estimates one is interested in, an average register or a register on a certain reference date is used as the backbone of the estimation process. Many cross-tabulations can be counted from the combined registers.

In addition to linking the registers to each other, survey data are linked to the register data. In principle, all surveys of individuals and households 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), but in the near future, other survey data may be used as well. The EWS is a large two-phase survey among businesses and contains information on, for example, wages and hours of employment. The LFS is a household survey and contains variables such as occupation, education, and search behavior on the labor market. The ISSS is a survey of individuals and contains information related to, for instance, education and health. 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 registers on 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 occupation, can be linked "around" the survey date, that is, they can be linked to the register data on a certain desired reference date not too far from the survey date, as if they were collected on 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 carried out on the reference date, and a cross-section of the linked registers on this particular reference date can be used.

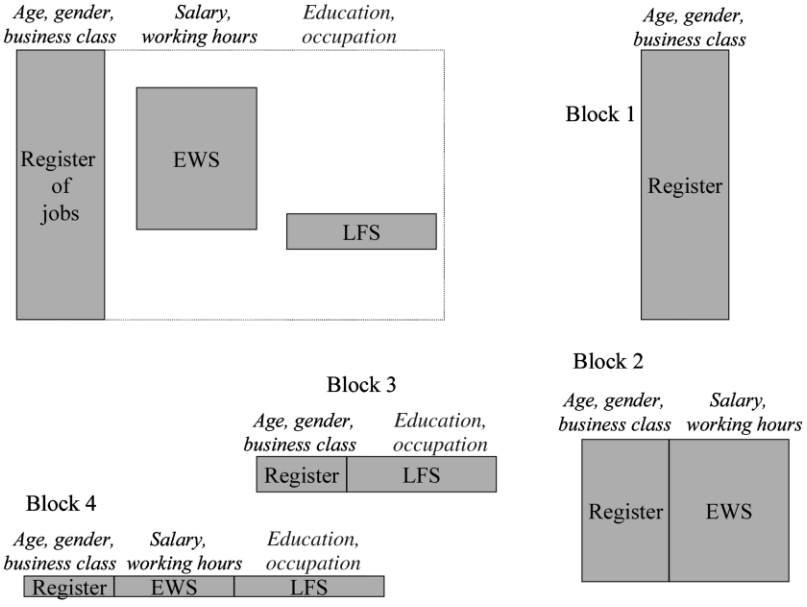


*Fig. 1. Example of linked registers and surveys used for the Structure of Earnings Survey*

The registers and surveys in the SSD were recently used to estimate the Structure of Earnings Survey (SES). The SES is a publication on jobs in The Netherlands and the (average) hourly, monthly, and yearly wages for these jobs, set 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 of minimizing the respondent 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 EWS and the LFS. The SES describes the situation as of December 31, so in that case, the register of persons is linked to the register of jobs on this reference date. Figure 1 shows the linked data sets used for the SES. In this figure, two surveys (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; most SES tables must be estimated from this overlap.

As mentioned before, in order to reduce the variance of estimates, each estimate from the SSD will be based on as many records as possible. For this reason, rectangular, complete data blocks are extracted from the linked data sets. Each data block contains all records that have a certain maximal set of variables in common. Figure 1 shows the extraction of rectangular data blocks from the linked data sets used for the SES. Owing to the partial record overlap of the two surveys, four rectangular data blocks can be created: (1) a data block containing all elements in the population and all variables in the register, (2) a data block containing all records in the largest of the two surveys (the EWS) and for each record all relevant variables from that survey and the register, (3) a data block containing all records from the smallest survey (the LFS) and all relevant variables from that survey and the register, and (4) a data block containing all records in the overlap of the two surveys, and all variables from both surveys, as well as the register variables. For each estimate, the largest rectangular data block – in terms of number of records – that contains all relevant variables simultaneously is, in principle, used. So, considering the data blocks in Figure 1, the frequency table "gender × working hours × education" must be estimated from data block 4, but the margin (lower-dimensional aggregate) "gender" can be counted from the register, the margin "gender × working hours" can be estimated from data block 2, and the margin "gender × education" can be obtained from data block 3.

Before estimates can be made from these rectangular data blocks, weights $w_i$ must be attached to the data to inflate from the samples to the population. For a data block consisting of register data only, the weights of the records are of course equal to unity.[4] For data blocks that consist of survey data (e.g., blocks 2-4 in Figure 1), the weights depend on the design of the surveys, the actual nonresponse, and the use of auxiliary information.

[4] Note that, for some table sets, one might be interested in the average over some time period, instead of the situation on 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.

More precisely, for data blocks 2 and 3, the block weights $w_i$ are given by the standard survey weights, which are, in addition, calibrated on (some of the) known population totals from data block 1 to correct for nonresponse and to reduce the variance of estimates. This requires a careful selection of the weighting model. In choosing auxiliary variables, the three basic requirements, that they should explain the response probabilities, explain the variation of the main study variables, and identify the most important domains, should be satisfied to the extent possible (see Lundström and Särndal 1999; 2002).

Since the two surveys are independent, the weights of the records in data block 4 are given by the product of the standard survey weights from each of the surveys. To correct for nonresponse and reduce the variance of the estimates from data block 4, these product weights can subsequently be calibrated not only on (some of the) known population totals from data block 1, but also on estimated population totals from data blocks 2 and 3 (see Renssen 1998). This requires again a careful selection of the weighting model. With these block weights $w_i$, cross-tabulations can be estimated from the data blocks. These estimates will automatically be consistent with the population totals used in the weighting model for nonresponse correction and variance reduction. By extending the weighting model for each data block with additional variables, more estimates based on these block weights will be immediately consistent. However, owing to lack of degrees of freedom, it is in general impossible to include all known crossings from the register and estimated crossings from larger data blocks in the weighting model of a certain data block. Therefore, some estimates from this data block may be numerically inconsistent with corresponding register counts and estimates from larger data blocks.

Cross-tabulations that cannot be estimated consistently with the block weights should be calculated with the method of repeated weighting. In the next section, we explain this method in more detail. In the remainder of this section, we focus on some important requirements regarding the data sets that are included in the SSD. First, these data sets 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 unit nonresponse. In general, missing values and inconsistencies at the micro level cause unacceptable inconsistencies in the estimates. Furthermore, the records in the registers and the surveys should 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 on the matching of 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). Finally, for the method of repeated weighting, when it comes to the variables a requirement is that they should be hierarchical if a variable consists of more than one classification level. For example, the variable "age" may be divided into age classes at several levels, such as 10-year classes, 5-year classes, and 1-year classes, as long as they are hierarchical. An additional level of 7-year classes would not be hierarchical and is therefore not allowed.

## 3. Consistent Estimates from the Social Statistical Database

Having constructed the rectangular data blocks and having assigned weights $w_i$ to the records in each data block, one can finally start to estimate tables from the Social Statistical Database (SSD). Because of the one-figure policy, all table estimates concerning a certain statistical topic should preferably be numerically consistent with each other. 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 being different in the other variables may be estimated from different data blocks, i.e., with different records and different weights. 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. Again referring to the Structure of Earnings Survey example (SES) in Figure 1, the margin "gender × working hours" of the frequency table "gender × working hours × education" estimated from data block 4 will in general not coincide with the more accurate estimate of "gender × working hours" from data block 2. In the Appendix, an example of these numerical inconsistencies is given. With the method of repeated weighting such inconsistencies are prevented. To estimate a fully consistent set of tables $\{T_1, T_2, \ldots, T_K\}$ from the SSD, the following procedure is adopted (see Kroese and Renssen 2000 and Renssen et al. 2001):

1. Every cross-tabulation $T_k$ $(k = 1, \ldots, K)$ 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. Tables from larger data blocks are estimated before tables from smaller data blocks, and each table is estimated using as many data as possible.
2. If a cross-tabulation $T_k$ has a margin $T_m$ that can be estimated from a larger data block, this margin 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 $T_k$.
3. All cross-tabulations $T_k$ that can be estimated consistently with the block weights $w_i$ 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 $w_i$ 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.
4. Suppose that a cross-tabulation $T_k$ cannot be estimated consistently with the block weights of the most suitable data block, but suppose that this table has a margin $T_m$ for which the most suitable data block is the same as the one for $T_k$, and $T_m$ can be estimated consistently with the block weights. In that case, the margin $T_m$ should be added to the table set (if not already present) and estimated with the block weights before $T_k$ is estimated.
5. If a cross-tabulation $T_k$ 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 $w_i$ will be adjusted by some additional reweighting scheme, taking into account all tables $T_1, \ldots, T_{k-1}$ that are already estimated according to the rules under points 1, 2, 3, and 4.

Thus, only when a table cannot be directly estimated consistently with the block weights $w_i$ – which are optimally designed for nonresponse correction and variance reduction – are these weights adjusted slightly, but only to estimate the table in question. The weights are adjusted such that the distance between the block weights and the adjusted weights (according to some distance function) is minimized, under the restriction that consistency is achieved with all other tables already estimated in the table set having variables in common with the table under consideration. For reweighting, the calibration properties of the generalized regression estimator (see Deville 1988 and Deville and Särndal 1992), are used, as will be explained below.

In the absence of nonresponse in a survey of $n$ elements from a population of $N$ elements, and using the known population totals $\vec{t}_x = (t_{x_1}, t_{x_2}, \ldots, t_{x_j})'$ of the $J$ auxiliary variables $X_1, X_2, \ldots, X_J$, the generalized regression estimator (GREG-estimator) $\hat{\vec{t}}_y^R = (\hat{t}_{y_1}^R, \ldots, \hat{t}_{y_P}^R)'$ for the population totals of the $P$ target variables $Y_1, Y_2, \ldots, Y_P$, is given by (see Cassel et al. 1976)

$$\hat{\vec{t}}_y^R = \hat{\vec{t}}_y^{HT} + \mathbf{B}_\pi'(\vec{t}_x - \hat{\vec{t}}_x^{HT}) \tag{1}$$

where the $(J \times P)$-matrix of estimated regression coefficients $\mathbf{B}_\pi$ is given by

$$\mathbf{B}_\pi = (\mathbf{X}'\Pi^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Pi^{-1}\mathbf{Y}) \tag{2}$$

and the direct, or Horvitz-Thompson, estimators for the population totals of the target and auxiliary variables are, respectively, given by the $P$- and $J$-vectors

$$\hat{\vec{t}}_y^{HT} = \mathbf{Y}'\Pi^{-1}\vec{i}_n$$

$$\hat{\vec{t}}_x^{HT} = \mathbf{X}'\Pi^{-1}\vec{i}_n$$

In the expressions above, the $(n \times J)$-matrix $\mathbf{X}$ denotes the matrix with scores $x_{ij}$ of record $i$, for $i = 1, 2, \ldots, n$, on auxiliary variable $X_j$, where $j = 1, 2, \ldots, J$, and, similarly, the $(n \times P)$-matrix $\mathbf{Y}$ has elements $y_{ip}$ with the scores of record $i$ on target variable $Y_p$, for $p = 1, 2, \ldots, P$. The $n \times n$ diagonal matrix $\Pi^{-1}$ has elements $1/\pi_i$, the inverse inclusion probability of record $i$ in the survey. The vector $\vec{i}_n$ is an $n$-vector with all elements equal to one. Note that the GREG-estimator (for simplicity called regression estimator in the following) for the population totals of the variables $X_j$ instead of the variables $Y_p$, would return exactly the known population totals for each $X_j$, which shows the calibration properties of the regression estimator.

The regression estimator in Equation (1) can be simplified when the matrix $\mathbf{X}$ contains a column of ones, or when a linear combination of two or more columns of $\mathbf{X}$ equals the vector $\vec{i}_n$. In the first case, the population total $N$ is explicitly used as an auxiliary variable. In the second case, two or more of the auxiliary variables $X_j$ correspond to the mutually exclusive categories of some categorical variable. It can easily be shown that in these cases we have

$$\hat{\vec{t}}_y^{HT} = \mathbf{B}_\pi'\hat{\vec{t}}_x^{HT}$$

and the regression estimator can be written in the simple projection form (see Särndal and Wright 1984)

$$\hat{\vec{t}}_y^R = \mathbf{B}'_\pi \vec{t}_x \tag{3}$$

This form of the regression estimator will prove useful for the process of repeated weighting. As explained earlier, reweighting is only necessary if a cross-tabulation cannot be estimated consistently with the block weights of the rectangular data block from which the table is to be estimated. If this happens to be the case, the block weights have to be adjusted somewhat so that consistency with all other cross-tabulations having margins in common with the table under consideration, is enforced. To obtain a consistent estimate for the target table, we first have to determine which margins the present table has in common with already estimated, consistent tables. These margins form the weighting model for repeated weighting; each margin corresponds to a term in the weighting model.[5] Here, the connection with the regression estimator becomes clear: the cell totals in the target table can be seen as the population totals of $P$ target variables $Y_1, \ldots, Y_P$, and the cell totals corresponding to the cross-tabulations in the weighting model can be considered as the population totals of $J$ auxiliary variables $X_1, \ldots, X_J$. In analogy with the known population totals $\vec{t}_x$ in the regression estimator, a $J$-vector $\vec{r}$ containing the counted or estimated population totals of the cells of the weighting model can be defined. Note that some of the terms in the weighting model may be redundant in the sense that they are dominated by other terms, that is, they are margins of these other dominant terms. Redundant terms can immediately be omitted from the weighting model; they do not add any additional information. The dominant terms should always be kept.

Since the tables (terms) in the weighting model are, by construction, margins of the target table, all tables in the weighting model are related to the same quantitative variable $Y$ as the target table.[6] For instance, they are all frequency tables, or they are all tables on income of people. In addition, each cell in the target table is, by construction, related to one or more cells in the weighting model.[7] Suppose that the target table has $P$ cells, and that the nonredundant margins in the weighting model correspond to $J$ cells. The estimated or counted population totals of these $J$ cells are recorded in the $J$-vector $\vec{r}$. The relationship between the cells of the target table and the cells of the weighting model can be expressed in a $(J \times P)$-matrix $\mathbf{L}$. The matrix $\mathbf{L}$ is defined such that an element $l_{jp}$ of this matrix equals 1 if cell $p$ of the target table contributes to cell $j$ of the weighting model, and zero otherwise. Moreover, there exists a clear relationship between the scores $y_{ip}$ of the $P$ target variables for record $i$, and the values $x_{ij}$ of the $J$ auxiliary variables corresponding to record $i$. After all, each record only contributes to one cell, say cell $p$, of the target table.

[5] This weighting model corresponds to the "minimal" weighting model required to obtain consistency between estimates. In principle, the weighting model can be extended with additional auxiliary variables that correlate with the variables in the target table to reduce the variance of the estimates further.

[6] The weighting model in repeated weighting may, as will be explained later, also contain terms with a different quantitative variable $Z$ in addition to terms related to $Y$.

[7] If for some reason (for instance, for the purpose of variance reduction) an extra term (table) is added to the weighting model, and this term contains a dimension variable not present in the target table, this extra dimension variable can without loss of generality be added to the target table. After calibrating, the target table can be aggregated with respect to this variable, resulting in the same table as there would have been if the extra dimension variable had not been added.

Therefore, $y_{ip} = y_i$ if record $i$ falls in cell $p$, and zero otherwise. The value of $y_i$ equals one for frequency tables, and takes an arbitrary real value for other quantitative variables such as income. By multiplying the scores $\vec{y}_i' = (y_{i1}, y_{i2}, \ldots, y_{ip}, \ldots, y_{iP}) = (0, \ldots, 0, y_i, 0, \ldots, 0)$ of record $i$ on the $P$ target variables on the left with the matrix $\mathbf{L}$, we obtain the scores $\vec{x}_i' = (x_{i1}, x_{i2,}, \ldots, x_{iJ})$ of record $i$ on the $J$ auxiliary variables. The scores are obviously equal to $y_i$ times the $p$-th column of the $\mathbf{L}$-matrix, or equivalently,

$$\mathbf{X}' = \mathbf{L}\mathbf{Y}'$$

As will be explained in the next section, irrespective of the quantitative variable $Y$ in the target table, the weighting model consists of at least a constant (the overall population total $N$) or one or more frequency tables, each having mutually excluding cells. Therefore, with repeated weighting, the simple regression estimator formula from Equation (3) can always be used. Thus we arrive at the following expression of the repeated weighting estimator for the $P$ cells of the target table

$$\begin{aligned}\hat{\vec{t}}_y^{RW} &= \mathbf{B}_w'\vec{r} \\ &= (\mathbf{Y}'\mathbf{W}\mathbf{X})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\vec{r} \\ &= (\mathbf{Y}'\mathbf{W}\mathbf{Y}\mathbf{L}')(\mathbf{L}\mathbf{Y}'\mathbf{W}\mathbf{Y}\mathbf{L}')^{-1}\vec{r} \\ &\equiv \hat{\mathbf{T}}\mathbf{L}'(\mathbf{L}\hat{\mathbf{T}}\mathbf{L}')^{-1}\vec{r}\end{aligned} \tag{4}$$

where the superscript $RW$ stands for repeated weighting, and $\mathbf{B}_w$ indicates that the matrix $\Pi^{-1}$ from Equation (2) must be replaced by the $(n \times n)$-diagonal matrix $\mathbf{W}$ that contains the block weights $w_i$ of the data block from which the target table is estimated (see Boonstra 2004). The $(P \times P)$-matrix $\hat{\boldsymbol{T}} = \mathbf{Y}'\mathbf{W}\mathbf{Y}$ is also diagonal. The $p$-th diagonal element $\hat{T}_{pp}$ of this matrix is given by the "regression" estimator (which uses the block weights $w_i$, see Section 2) of the population total of $Y^2$ in the $p$-th cell of the target table:

$$\hat{T}_{pp} = \sum_{i=1}^{n} w_i y_{ip} y_{ip} = \sum_{i \in \text{cell}\, p} w_i y_i^2 \tag{5}$$

For frequency tables, this corresponds to the estimated cell counts since $y_{ip} = 1$ if record $i$ belongs to cell $p$, and $y_{ip} = 0$ otherwise. Note that by multiplying the repeated weighting estimator of Equation (4) on the left with the matrix $\mathbf{L}$, the restrictions $\vec{r}$ are recovered. Indeed, by multiplying the $j$-th row of $\mathbf{L}$ with $\hat{\vec{t}}_y^{RW}$, exactly those cells of the target table are aggregated which contribute to the $j$-th cell total of the weighting model. This shows that the estimated table will be consistent with the restrictions in the weighting model, as desired.

In practice, it might happen that the tables in the weighting model do not all contain the same count variable $Y$ as the target table, but that one or more of the tables in the weighting model contain a different count variable $Z$ instead of $Y$. For instance, one might be interested in estimating the target table "total income by gender × education," and the weighting model may contain both the table "total income by gender" as well as the frequency table "gender × education." The count variable $Y$ of both the target table and the first term in the weighting model is given by "income," whereas the second term in the weighting model has a different count variable $Z$ corresponding to "frequency count." In that case Equation (4) for the repeated weighting estimator is still valid, but the definitions

of $\hat{\vec{t}}_y^{RW}$, $\mathbf{Y}$, $\mathbf{L}$ and $\hat{\mathbf{T}}$ must be modified somewhat. Suppose that the first $J_1$ components in the $J$-vector $\vec{r}$ are related to variable $Y$, and the remaining $J_2 = J - J_1$ components to the variable $Z$. The vector of target variables is now given by the $2P$ vector $(y_{i1}, \ldots, y_{iP}, z_{i1}, \ldots, z_{iP})'$ with the scores of record $i$ $(i = 1, \ldots, n)$ on both quantitative variables $Y$ and $Z$ in each cell of the target table, resulting in an $(n \times 2P)$ matrix $(\mathbf{Y}, \mathbf{Z})$ instead of $\mathbf{Y}$. Similarly, the matrix $\mathbf{L}$ is given by the $(J \times 2P)$-block-diagonal matrix

$$\boldsymbol{L} = \begin{pmatrix} \mathbf{L}_Y & 0 \\ 0 & \mathbf{L}_Z \end{pmatrix}$$

where the $(J_1 \times P)$-matrix $\mathbf{L}_Y$, and the $(J_2 \times P)$-matrix $\mathbf{L}_Z$ are defined as in the case where only one count variable is present. The scores $\mathbf{X}$ on the auxiliary variables obviously equal $\mathbf{X}' = \mathbf{L}(\mathbf{Y}, \mathbf{Z})'$. The matrix $\hat{\mathbf{T}}$ is now given by the $(2P \times 2P)$-matrix

$$\hat{\mathbf{T}} = \begin{pmatrix} \hat{\mathbf{T}}^{YY} & \hat{\mathbf{T}}^{YZ} \\ \hat{\mathbf{T}}^{ZY} & \hat{\mathbf{T}}^{ZZ} \end{pmatrix}$$

where each submatrix is a $(P \times P)$ diagonal matrix with elements analogous to Equation (5):

$$\hat{\vec{T}}_{pp}^{YZ} = \sum_{i=1}^{n} w_i y_{ip} z_{ip} = \sum_{i \in \text{cell}\, p} w_i y_i z_i$$

and similar expressions for the diagonal elements of the other submatrices. Obviously, the repeated weighting estimator $\hat{\vec{t}}_{yz}^{RW}$ now has $2P$ components, of which the first $P$ are related to the count variable $Y$, and the other $P$ to the count variable $Z$. Depending on the count variable of the target table, either part of this estimate is the final table in which one is interested. This table will be consistent not only with all other tables having margins in common with the target table, but also with the extra restrictions in the weighting model concerning the other quantitative variable.

As a practical example of the reweighting procedure, consider again the target frequency table "gender × working hours × education" which must be estimated from data block 4 in Figure 1. This table must be calibrated on the frequency tables "gender × working hours" and "gender × education," which can be estimated from the larger and therefore more accurate data blocks 2 and 3, respectively. Unless the block weights of data blocks 2 and 3 are already calibrated on "gender," both tables in the weighting model of the target table must be estimated by calibrating on the register count of "gender" to obtain consistency with the register.[8] As a consequence, the target table will also be consistent with the register count of "gender." In the Appendix, the results of the reweighting procedure for this example are shown. The example shows that repeated weighting leads not only to numerically consistent estimates, but also to more accurate estimates. Since the regression estimator is only asymptotically unbiased, one might initially fear that repeated application of it will lead to an excessively growing bias and an

[8] In fact, the one-way tables "working hours" and "education" can be estimated consistently from blocks 2 and 3 without reweighting, that is, using the block weights. The two-way tables in the weighting model of the target table should therefore be calibrated not only on "gender" from the register, but also on the corresponding one-way table "working hours" or "education," respectively.

accumulating error. However, as long as the sample size is sufficiently large such that the regression estimator is asymptotically unbiased, one can intuitively understand that repeated weighting leads to more accurate estimates. 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 block weights $w_i$.

## 4. Points of Attention with Consistent Weighting in General

With repeated weighting, a fully consistent set of tables can be estimated. The weights that are used for each estimate are either the block weights $w_i$ (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] The method has been applied in several research projects on real data at Statistics Netherlands (see for instance, Statistics Netherlands 2000). It was observed that the method of repeated weighting works well in the sense that relatively simple and well-defined table sets can be estimated consistently from the Social Statistical Database (SSD). Nonetheless, it is clear that the method is not without complications. In this section, we mention some of these complications that require special attention and, in some cases, further development of the theory of repeated weighting.

In the SSD, sample surveys are linked to registers. 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. However, a requirement is 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 the two 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 the two 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 is at present putting considerable effort into resolving the issue. One problem, for instance, is that lack of harmonization prevents one from using the Integrated System on Social Surveys data on education together with the Labor Force Survey data to estimate tables related to education in the Structure of Earnings table set. 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,

[9] Although the reweighting may yield estimates with lower variances, the method is in the first place applied for cosmetic purposes, and should therefore have no large influence on the actual estimates.

since one is no longer completely free to choose different categories for similar variables, depending on what one is interested in 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.

A second point that requires some attention is related to the estimating process itself. Even when cross-tabulations are estimated according to the rules given in Section 3, there is no unique estimate for tables that are estimated via repeated weighting. More precisely, the adjusted weights for each table may differ since they depend on the weighting model used. The weighting model, in turn, depends on the tables in the table set that have already been estimated. As an example, consider two tables $T_A$ and $T_B$ that have to be estimated from the same data block. Suppose that both tables need reweighting because they cannot be estimated consistently with the block weights $w_i$. Assume also that these tables have some variables in common. If table $T_A$ is estimated first, then the margin with respect to the common variables with table $T_B$ will occur in the weighting model for table $T_B$, and the other way around if table $T_B$ is estimated first. It is clear that the resulting estimates will 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 margins of a table are estimated. If, for instance, the three-way frequency table "gender × working hours × education" is to be estimated, first the one-way tables "gender," "working hours," and "education" are estimated. Subsequently, the two-way tables "gender × working hours," "gender × 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. But even though the order problem can be solved by completely fixing the order, 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.

A third 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 weighting model. 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. One way to deal with this problem is to combine several categories in the variables where the problem occurs. Owing 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 variable. The first option leads to loss of information and the second option will be difficult to implement in the process of repeated

weighting, because it may be difficult to find cell combinations that solve all empty-cell problems and at the same time satisfy the required hierarchy.

The use of synthetic estimators may be another way to treat empty-cell problems. In analogy with pseudo-Bayes estimators (see Bishop et al. 1975), one removes the survey zeros by filling the empty cells in the target table which cause the estimation problems with a small "ghost value" $\varepsilon$. These ghost values lead to a small change $\delta\hat{\mathbf{T}}$ in the matrix $\hat{\mathbf{T}}$ with estimated cell counts of the target table. The matrix $\delta\hat{\mathbf{T}}$ is, like the matrix $\hat{\mathbf{T}}$, a $(P \times P)$ diagonal matrix, with values $\delta\hat{\mathrm{T}}_{pp} = \varepsilon$ when the empty cell $p$ is filled with a ghost value, and zero otherwise. The ghost values $\varepsilon$ also make a small contribution to the restrictions $\vec{r}$ on which needs to be calibrated. Defining $\hat{\mathbf{T}}^* = \hat{\mathbf{T}} + \delta\hat{\mathbf{T}}$ and $\vec{r}^* = \vec{r} + \mathbf{L}\delta\hat{\mathbf{T}}\vec{i}_P$, it is easily seen that the synthetic estimator

$$\hat{\vec{t}}_y^S = \hat{\mathbf{T}}^*\mathbf{L}'(\mathbf{L}\hat{\mathbf{T}}^*\mathbf{L}')^{-1}\vec{r}^* - \delta\hat{\mathbf{T}}\vec{i}_P \tag{6}$$

satisfies the calibration restrictions, i.e.,

$$\mathbf{L}\hat{\vec{t}}_y^S = \vec{r}$$

Thus, by adding a small value to empty cells in the target table, the estimation problems are avoided. Of course, the estimated table will be somehow "artificial," and may even lead to negative cell counts if there is no other solution for the interior of the table, given the restrictions. Nevertheless, the influence of the ghost values on the bias of cells, which have sufficient contributing records, is small, since survey zeros are most likely to occur in rare domains and the corrections are of the order of the population size in these domains. After estimating all tables in the table set, artificial cell counts may be combined with other cells, or left out completely from a publication. Note that care should be taken in picking the empty cells to be filled. For instance, structural zeros should always remain empty. Preferably, as little as possible should be changed in the original table, which means that as few empty cells as possible should be filled with a ghost value. The value of $\varepsilon$ itself can be the same for all cells, but also more advanced methods can be used, such as taking $\varepsilon$ cell dependent and proportional to some *a priori* distribution.

A fourth point is related to edit rules between variables. 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 in the SSD. For example, it could easily happen that the number of people having a driver's license in some small area exceeds the number of people who are 18 and older (see Kroese and Renssen 2000). In The Netherlands, no person younger than 18 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 by including the age variable in the cross-tabulation on possession of a driver's license.

A special case of edits is related to quantitative variables. Suppose that the total income per income class is to be estimated, and the number of people per income class has been estimated independently. Then the average income in any income class should be higher than the average income in all lower income classes. This can be guaranteed by adding the frequency table to the weighting model of the income table (see Renssen et al. 2001). In general, when estimating a table on some quantitative variable like "income" or "hours

worked," the underlying frequency table always has to be included in the weighting model. Note that in this way, not only can more than one count variable occur in the weighting model, but also the simplified form of the regression estimator can always be used.

A problem related to edit rules arises when the 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 women in these households should equal the total number of females in the population. When the persons register contains a key that identifies to which household each person belongs, the general integrated estimation procedures from Lemaître and Dufour (1987) can, in principle, be used to ensure such consistency.[10] However, these techniques are not yet incorporated in the method of repeated weighting.

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 cross-tabulations 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 original block weights according to the distance function used, and they can even become negative. A large variability in the adjusted weights leads to larger variances. So, although the mean squared 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. 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. This breakdown point of the repeated weighting estimator is a topic for further research.

A last complication with the consistent estimates from the SSD that we want to mention is related to the estimation of variances of the estimates. Remember that weights of the surveys are determined first, which are then reweighted to correct for nonresponse and to reduce the variance. Subsequently, these weights are adjusted again to estimate tables that are not yet consistent. An approximated variance estimator for the regression estimator can be readily derived (see Särndal et al. 1992). However, this variance estimator is only valid when the population totals of the auxiliary variables are known. In the case of repeated weighting, the restrictions on which a target table must be calibrated are often estimates themselves. These estimates are usually less detailed margins of the target table, and are estimated also by calibration on even less detailed margins, and so forth. This quickly leads to a large tree of tables which all contribute in some way to the estimation of some target table. The calculation of the variance will therefore be correspondingly complicated. However, in the case of independent, and record-wise nonoverlapping

[10] In The Netherlands, for the vast majority of cases, households can be derived from the information available in the Municipal Base Administration (MBA). The household position of the remaining persons is imputed using household information from large-scale household surveys linked to the MBA. The persons can then be linked uniquely to households.

surveys, with small sampling fractions and identical sampling frames, an approximated variance formula can be derived (see Knottnerus 2003). If the above conditions are not met, a rough approximation for the variance of the table estimates can still be made (see Houbiers et al. 2003).

## 5. Concluding Remarks

In this article, we have given an overview of how Statistics Netherlands envisions improving the quality of its estimates by using register data. In order to do so, Statistics Netherlands is constructing a large Social Statistical Database where registers are linked to each other as well as to survey data. Using the combined registers as auxiliary information, Statistics Netherlands hopes to obtain from this database accurate and reliable estimates related to social statistics. In addition to accurate and reliable estimates, Statistics Netherlands wishes to publish – as far as possible – one single figure for each statistical concept. To achieve numerical consistency between estimates, the method of repeated weighting is being developed. It is not very likely that the ultimate goal of consistency between all possible estimates from the Social Statistical Database will be reached soon, if at all, but repeated weighting seems a suitable method to estimate well-defined table sets consistently. As such, the method should be seen as a tool in the toolbox of estimation methods. 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. Although there are still some unsolved problems (for instance, how to deal with edit rules), the method is successfully applied in relatively simple cases such as the Structure of Earnings Survey and independently, a table set concerning the search behavior on the labor market of people on social welfare. At present, the Social Statistical Database and the method of repeated weighting are also used to meet Eurostat's demands regarding the 2001 Census (see Van der Laan 2000).

To facilitate the calculations once rectangular data blocks are extracted from the social statistical database, a software package called VRD has been developed. After entering all relevant meta data about the categories of variables, the hierarchical relations between different levels of each variable, and the composition of the rectangular data blocks, the user can indicate which tables should be estimated. These tables are subsequently estimated, each table from the most suitable data block. However, the software package is still under construction. The basic functionalities are built in, but more advanced functionalities, which allow for e.g., synthetic estimators in the case of empty-cell problems, or a general functionality to deal with edit rules, are not yet included. Variances of the estimates can, however, still be calculated under the assumptions mentioned in the previous section.

## Appendix: Numerical Example of Repeated Weighting

In this appendix, we give a fictitious example from the Structure of Earnings Survey (SES). The frequency table "gender × working hours × education" is estimated, both without and with repeated weighting. The target population is "jobs," so in this frequency table the number of jobs is estimated against some background variables, in particular, gender and educational level (seven categories) of the person having a certain job, and the

*Table 1. Estimate of target table from data block 4, using the corresponding block weights (unit = 1,000)*

| Gender × working hours | Male | | | Female | | |
|---|---|---|---|---|---|---|
| Level of education | Part-time | Full-time | Total male | Full-time | Part-time | Total female |
| Primary or less | 92.2 (12.8) | 277.4 (21.6) | 369.6 (24.7) | 173.3 (16.7) | 52.1 (10.3) | 225.4 (18.7) |
| Lower secondary general | 103.5 (13.7) | 144.0 (11.3) | 247.5 (17.2) | 193.3 (13.8) | 85.6 (12.4) | 278.9 (17.9) |
| Lower secondary vocational | 87.7 (10.3) | 478.0 (26.5) | 565.7 (27.9) | 213.1 (15.1) | 61.7 (10.0) | 274.8 (17.6) |
| Upper secondary general | 74.0 (11.2) | 149.7 (14.7) | 223.7 (17.8) | 154.8 (15.8) | 76.6 (10.2) | 231.5 (18.2) |
| Upper secondary vocational | 166.4 (16.3) | 1,236.7 (34.1) | 1,403.1 (35.5) | 672.6 (25.4) | 363.0 (22.5) | 1,035.6 (29.3) |
| Vocational college | 123.9 (11.1) | 482.4 (21.5) | 606.3 (23.7) | 305.4 (15.2) | 175.5 (9.9) | 480.9 (17.8) |
| University or more | 64.4 (6.7) | 267.0 (14.4) | 331.3 (15.8) | 85.6 (8.2) | 86.5 (10.3) | 172.1 (12.8) |
| Total | 712.1 (26.5) | 3,035.1 (26.5) | 3,747.2 | 1,798.1 (28.4) | 901.1 (28.4) | 2,699.2 |

*Table 2. Estimate of "gender × working hours" from data block 2, using the block weights (unit = 1,000)*

| Gender × working hours | Male Part-time | Full-time | Total male | Female Part-time | Full-time | Total female |
|---|---|---|---|---|---|---|
| Total | 701.1 | 3,046.2 | 3,747.2 | 1,827.9 | 871.3 | 2,699.2 |

*Table 3. Estimate of "gender × education" from data block 3, using the block weights (unit = 1,000)*

| Level of education × gender | Total male | Total female |
|---|---|---|
| Primary or less | 357.2 (5.2) | 209.8 (3.9) |
| Lower secondary general | 267.3 (4.5) | 290.5 (4.4) |
| Lower secondary vocational | 595.8 (6.3) | 320.6 (4.6) |
| Upper secondary general | 213.2 (4.2) | 205.4 (3.9) |
| Upper secondary vocational | 1,374.0 (8.3) | 1,006.9 (6.9) |
| Vocational college | 616.0 (6.3) | 500.1 (5.5) |
| University or more | 323.8 (4.9) | 166.0 (3.5) |
| Total | 3,747.2 | 2,699.2 |

hours worked per week in that job ($<35$ hours corresponds to part-time, $\geq 35$ hours corresponds to full-time). Referring to Figure 1, the frequency table must be estimated from data block 4, which consists of approximately 50,000 records. The table estimate is shown in Table 1. To obtain this estimate, the block weights were used to inflate from the data block to the population. In the SES, the block weights of data blocks 2, 3, and 4 are all calibrated on "gender," so the estimated total numbers of jobs occupied by males and females exactly coincide with the register totals, that is, 3,747.2 thousand for the males and 2,699.2 thousand for the females.[11]

The numbers in brackets in Table 1 give the standard errors of the corresponding totals. These standard errors are estimated using Taylor linearization of the regression estimator (see Särndal et al. 1992). As can be seen, the standard errors are quite large. However, the margin "gender × working hours" (last row in Table 1) can be estimated much more accurately from data block 2, which contains the records of approximately half (!) the population. Similarly, the margin "gender × education" (total male/female columns in Table 1) can be estimated more accurately from data block 3, which consists of approximately 100,000 records. The resulting estimates are shown in Tables 2 and 3. Except for the total number of males and females, the estimates of these tables clearly differ from the corresponding margins in Table 1.

The standard errors for the table "gender × education" are also shown; again they are obtained by Taylor linearization of the regression estimator. The standard errors are obviously much smaller than the corresponding ones in Table 1. In principle, the estimates for the table "gender × working hours" also have a small variance. Owing to the size of data block 2, they are ignored here.

[11] In fact, all three data blocks are calibrated on "gender × age class + business class." Owing to rounding, there are small differences between row and column totals and the sum of the interior cells.

*Table 4. Estimate of target table from data block 4, using repeated weighting (unit = 1,000)*

| Gender × working hours | Male | | | Female | | |
|---|---|---|---|---|---|---|
| Level of education | Part-time | Full-time | Total male | Part-time | Full-time | Total female |
| Primary or less | 87.3 (10.6) | 269.9 (11.2) | 357.2 (5.2) | 163.0 (9.2) | 46.8 (8.8) | 209.8 (3.9) |
| Lower secondary general | 110.2 (9.2) | 157.1 (9.5) | 267.3 (4.5) | 203.9 (9.7) | 86.6 (9.4) | 290.5 (4.4) |
| Lower secondary vocational | 90.4 (9.4) | 505.3 (10.7) | 595.8 (6.3) | 250.6 (8.9) | 70.0 (8.3) | 320.6 (4.6) |
| Upper secondary general | 69.2 (8.9) | 143.9 (9.3) | 213.2 (4.2) | 139.6 (8.8) | 65.8 (8.5) | 205.4 (3.9) |
| Upper secondary vocational | 159.0 (13.0) | 1,215.0 (14.8) | 1,374.0 (8.3) | 664.0 (14.9) | 342.8 (14.5) | 1,006.9 (6.9) |
| Vocational college | 123.3 (9.6) | 492.7 (10.8) | 616.0 (6.3) | 322.4 (9.3) | 177.7 (8.9) | 500.1 (5.5) |
| University or more | 61.5 (6.2) | 262.2 (7.3) | 323.8 (4.9) | 84.4 (6.8) | 81.6 (6.8) | 166.0 (3.5) |
| Total | 701.1 | 3,046.2 | 3,747.2 | 1,827.9 | 871.3 | 2,699.2 |

Using repeated weighting, the target table can be estimated consistently with the more accurate margins from data blocks 2 and 3. In terms of the notation in Section 3, the vector $\vec{r}$ has 18 components, given by the $4 + 14$ restrictions in Tables 2 and 3. The matrix $\hat{\mathbf{T}}$ contains the 28 "gender × working hours × education" estimates from Table 1 on its diagonal. The zero-one matrix $\mathbf{L}$ has $18 \times 28$ elements and these can be determined easily from Table 1. For instance, the elements $l_{j1} (j = 1, \ldots, 18)$ in the first column of $\mathbf{L}$ are all zero, except for $l_{11}$ and $l_{51}$, which are equal to 1, indicating that the first cell of the target table "Male, Part time, Primary or less" contributes to the first "Male, Part time" and the fifth "Male, Primary or less" of the 18 restrictions. Note that owing to the linear dependence of the restrictions, the matrix $\mathbf{L}\hat{\mathbf{T}}\mathbf{L}'$ is singular, so that the generalized inverse, instead of the normal inverse, must be calculated. As long as the rank of $\mathbf{L}\hat{\mathbf{T}}\mathbf{L}'$ is larger than or equal to the number of independent restrictions in $\vec{r}$, this has no effect on the final repeated weighting estimate (see Renssen and Martinus 2002).[12] The result is shown in Table 4. Indeed, the more accurate estimates of "gender × working hours" and "gender × education" from Tables 2 and 3 are reproduced in the margins of the target table.

The estimated standard errors of the repeated weighting estimates are also shown. They are smaller than the corresponding ones in Table 1, showing that repeated weighting leads not only to numerical consistency, but also to more accurate estimates owing to a better use of auxiliary information. A detailed derivation of these variance estimates goes far beyond the scope of this article, but roughly speaking, these standard errors are calculated by repeated Taylor linearization of the regression estimator (see Knottnerus 2003 and Houbiers et al. 2003).

## 6. References

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. (2004). Calibration of Tables of Estimates. Research paper, BPA-no M336-04-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, 615–620.

Deville, J.-C. (1988). Estimation Linéaire et Redressement sur Informations Auxiliaires d'Enquêtes par Sondage. Essais en l'Honneur d'Edmond Malinvaud A.Monfort and J.J.Laffond (eds). Paris: Economica, 915–927.

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

Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H., and Snijders, V. (2003). Estimating Consistent Table Sets: Position Paper on Repeated Weighting. Discussion paper 03005. Statistics Netherlands: Voorburg/Heerlen, http://www.cbs.nl/en/publications/articles/general/discussion-papers/discussion-papers.htm.

[12] Note that survey zeros cause the rank of the matrix $\mathbf{L}\hat{\mathbf{T}}\mathbf{L}'$ to drop below the number of independent restrictions, indicating that survey zeros create estimation problems. We discuss this issue in Section 4.

Knottnerus, P. and Wiegert, R. (2002). DACSEIS deliverable 7.1: Questionnaire 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. New York: Springer-Verlag.

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. Proceedings of IASS Satellite Conference on Small Area Estimation. Riga: Latvia, 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. Proceedings of ICES II, Buffalo NY, American Statistical Association, 831–840.

Lemaître, G. and Dufour, J. (1987). An Integrated Method for Weighting Persons and Families. Survey Methodology, 13, 199–207.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. Journal of Official Statistics, 15, 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 Matching Techniques in Calibration Estimation. Survey Methodology, 24, 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.

Renssen, R.H. and Martinus, G.H. (2002). On the Use of Generalized Inverse Matrices in Sampling Theory. Survey Methodology, 28, 209–212.

Silva, P.L.D.N. and Skinner, C.J. (1997). Variable Selection for Regression Estimation in Finite Populations. Survey Methodology, 23, 23–32.

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

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

Särndal, C.-E. and Wright, R.L. (1984). Cosmetic Form of Estimators in Survey Sampling. Scandinavian Journal of Statistics, 11, 146–156.

Thomson, I. and Kleive Holmøy, A.M. (1998). Combining Data from Surveys and Administrative Record Systems: The Norwegian Experience. International Statistical Review, 66, 201–221.

Van der Laan, P. (2000). The 2001 Census in The Netherlands: Integration of Registers and Surveys. Proceedings of the Insee-Eurostat seminar on censuses after 2001, Paris, 39–52.

Received November 2002
Revised October 2003