Discussion Paper

Divide-and-Conquer solutions for estimating large consistent table sets

2016 | 19

Jacco Daalmans

Content

Summary

When several frequency tables need to be produced from multiple data sources, there is a risk of numerical inconsistent results. This means that different estimates are produced for the same cells or marginal totals in multiple tables. As inconsistencies of this kind are often not tolerated, there is a clear need for compilation methods for achieving numerical consistent output. Statistics Netherlands developed a Repeated Weighing (RW) method for this purpose. The scope of applicability of this method is however limited by several known estimation problems. This paper presents two new Divide-and-Conquer (D&C) methods that can be used as an alternative for RW. The two D&C methods break down the estimation problem as much as possible into independently estimated parts, rather than the dependently estimated parts that are distinguished in RW. In this way estimation problems are as much as possible prevented; a result that was confirmed by an application to the Dutch 2011 Census. Thus, we arrive at the conclusion that the two newly developed D&C methods can be much more easily implemented than the existing RW method.

Keywords

Repeated weighting; Quadratic Programming, Census, Weighting, Consistency; Constrained optimization.

1. Introduction

Statistical outputs are often interconnected. It may happen that a cell in one table is also published in another table, or that two tables share a common marginal total. In such cases certain relationships might be expected to be fulfilled, i.e. that the same values are published for common outputs in different publications. Otherwise, there may be confusion about the 'true' value and there may be a risk of cherry-picking, i.e. users who choose results that are most convenient for themselves. When compiling statistics numerical consistent results are however not automatically achieved. Inconsistencies may emerge due to differences in data sources and compilation methods. As inconsistencies are often not tolerated, there is a clear need for compilation methods for achieving numerical consistent output. In this report we consider the problem of numerical consistent estimation of interrelated contingency tables. Contingency tables display a multivariate frequency distribution, for instance Dutch population by age and sex. An important example of a multiple table statistical output in the Netherlands is the Dutch virtual census. For the census, dozens of detailed contingency tables need to be produced with many overlapping variables. Numerical consistent results are required by the European Census acts and a number of implementing regulations (European Commission, 2008). A distinction can be made between a traditional and a virtual census. A traditional census is based on a complete enumeration of the population based on a questionnaire. In this approach consistency is automatically present. Statistics Netherlands belongs to a minority of countries that conducts a virtual census. In a virtual census estimates are produced from already available data that are not collected for the census. The Dutch virtual census is for a large part based on integral information from administrative sources. For a few variables not covered by integral data sources, supplemental sample survey information is used. Because of incomplete data, census compilation relies on estimation. Because of the different data sources that are used numerically inconsistent results would be inevitable if standard estimation techniques were applied (de Waal, 2015 and 2016). To prevent inconsistency, Statistics Netherlands developed a method called "Repeated Weighting" (RW), see e.g. Renssen and Nieuwenbroek (1997), Nieuwenbroek *et al.* (2000), Houbiers *et al.* (2003), Knottnerus and Van Duin (2006). In RW the problem of consistently estimating a number of contingency tables with overlapping variables is simplified by splitting the problem into dependent sub problems. In each of these sub problems a single table is estimated. Thus, a sequential estimation process is obtained. The implementation of RW is however not without its problems (see Houbiers *et al.* 2003, Daalmans, 2015 and Subsection 2.4 below). In particular, there are problems that are directly related to the sequential approach. Most importantly, RW does not always succeed in estimating a consistent table set, even when it is clear that such a table set exists. After a certain number of tables have been estimated, it may become impossible to estimate a new one consistently with all previously estimated ones. This problem

seriously limits future application possibilities of repeated weighting. For the Dutch 2011 Census several ad-hoc solutions were designed after long trial-and-error. For any future application, it is however not guaranteed that numerical consistent estimates can be produced. Other problems with the sequential approach of RW are order-dependency of estimation and that a suboptimal solution may be obtained. Because of these problems there is a clear need for improving methodology. We present two Divide-and-Conquer algorithms that can be used as alternatives for RW. These algorithms break down the problem of estimating a large consistent table set into a number of sub problems that can preferably be independently estimated. In each step parts of a table set are estimated, but contrary to RW these parts are not the same as individual tables, but a combination of cells from different tables. Because the previously mentioned estimation problems do not occur, or at least have a smaller impact, the new approach can be much more easier implemented than RW. The problem specific solutions that are necessary for RW can be avoided. This paper is organised as follows. In Section 2, we describe the RW method. Section 3 presents an alternative quadratic programming (QP) formulation for this problem. Section 4 explains a simultaneous weighting approach. Two new Divide-and-Conquer methods that are introduced in Section 5. Results of a practical application are given in Section 6 and Section 7 concludes this paper with a discussion.

2. Repeated Weighting

In this section we explain the RW method. Subsection 2.1 describes prerequisites. The main properties of the method are given in Subsection 2.2. A more technical description follows in Subsection 2.3 and Subsection 2.4 deals with known complications of the method.

2.1 Prerequisites

Although RW can be applied to contingency and continuous data, this paper deals with contingency tables only, which are also often called frequency tables.
We assume that multiple prescribed tables need to be produced with overlapping variables. If there were no overlapping variables, it would not be any challenge to produce numerical consistent estimates.
Further, it is assumed that the target populations are the same for each table. This means for example that all tables necessarily have to add up to the same grand-total. All data sources relate to the same target population. There is no under- or overcoverage: the target population of the data sources coincides with the target population of the tables to be produced.
For each target table a predetermined data set has to be available from which that table is compiled. As explained in Houbiers *et al.*(2003), these data sets may contain records from one or multiple data source(s). Sometimes multiple choices are possible: one target table may be estimated from several (combined) data sources. In that case, using the data source(s) with the most units usually yields most accurate results.
Two types of data sets will be distinguished: data sets that cover the entire target population and data sets that cover a subset of that population. As the first type is often obtained from (administrative) registers and the latter type from statistical sample surveys, these data sets will be called registers and sample surveys from now on.
Theoretically, it is possible to use data based on a complete coverage for a subpart of the population and a sample for the other part of the population. For ease of explanation we will however not consider such 'mixed' data in this paper.
It will be assumed that all register-based data sets are already consistent at the beginning of RW. That means that all common units in different data sets have the same values for common variables. Subsection 2.2 explains why this assumption is important. To ensure that the assumption is satisfied in practise, a so-called micro-integration process has to be applied prior to repeated weighting, see e.g. Bakker (2011) and Bakker *et al.* (2014).
For sample surveys data sets it is required that weights are available for each sample survey unit. These are weights that are meant to be used to draw inferences for a population. For example, a weight of 12.85 means that one unit in the sample survey counts for 12.85 units in the population. To obtain weights for sample surveys, one usually starts with the sample weight, i.e. the inverse of the probability of selecting a

unit in the sample. Often, these sample weights are adjusted to take selectivity or non-response into account. Resulting weights will be called starting weights, as these are weights that are available at the beginning of repeated weighting.

2.2 Non-technical description

Below we expose the main ideas of RW. First we will explain how a single table is estimated. The estimation method depends on the type of the underlying data set. Tables that are derived from a register data source can simply be produced by counting from the register. This means that for each cell in the table, it is counted how much the corresponding categories (e.g. 28 year old males) occur. There is no estimation involved, because registers are supposed to cover the entire target population. The fact that register-based data are not adjusted explains why registers need to be already consistent at the beginning.

Below, we focus on tables that need to be estimated from a sample survey. These tables have to be estimated consistently. This basically means two things: common marginal totals in different tables have to be identically estimated and all marginal totals for which known register values exist have to be estimated consistently with those register values.

In the RW-approach consistent estimation of a table set is simplified by estimating tables in sequence. The main idea is that each table is estimated consistently with all previously estimated tables. When estimating a new table, it is determined first which marginal totals the table has in common with all registers and previously estimated tables. Then, the table is estimated, such that:

1) Marginal totals that have already been estimated before are kept fixed to their previously estimated values;
2) Marginal totals that are known from a register are fixed to their known value.

To illustrate this idea, we consider an example in which two tables are estimated:

Table 1: age × sex × educational attainment

Table 2: age × geographic area × educational attainment

A register is available that contains age, sex and geographic area. Educational attainment is available from a sample survey. Because educational attainment appears in Table 1 and 2, both tables need to be estimated from that sample survey. To achieve consistency, Table 1 has to be estimated, such that its marginal totals age × sex aligns with the known population totals from the register. For Table 2 it needs to be imposed that the marginal total age × geographic area complies with the known population totals from the register and that the marginal total age × educational attainment is estimated the same as in Table 1.

Each table is estimated by means of the generalised regression estimator (GREG), see Särndal *et al.* (1992), an estimator that belongs to the class of calibration estimators, see e.g. Deville and Särndal *et al.* (1992). By using this technique, it can be assured that a certain table is consistently estimated with all previously estimated tables and with all known totals from registers. Thus, repeated weighting comes down to a repeated application of the GREG-estimator.

Repeated weighting can be considered a macro-integration method, as adjustments are made at aggregate level; the data of individual units are not corrected. But, in

addition to the functionality of many other macro-integration methods, repeated weighting also establishes a link between the microdata and the reconciled tables. For each table, so-called corrected (or adjusted) weights are obtained that can be used to derive reconciled tables from microdata. For data sets that underlie estimates for multiple tables, corrected weights are usually different for each table. Since each reconciled table can be obtained by multiplying the underlying microdata units by corrected weights, categories of variables that do not occur in a sample survey will by definition have a zero value in all table estimates based on that survey. On the one hand this is a very desirable property, as it precludes the possibility of non-zero counts for categories that cannot exist in practice, for example five year-old professors. On the other hand, it is also the source of estimation problems, which will be explained in Subsection 2.4.

2.3 Technical description

In this subsection, repeated weighting is described in a more formal way. Below we will explain how a single table is estimated from a sample survey.
Aim of the repeated weighting estimator (RW-estimator) is to estimate the P cells of a frequency table $Y_1, \dots, Y_P$. We will use vector notation to express the elements of a table. The estimates are made from a sample survey, of which initial, strictly positive weights w_i are available for all n records. Each record in the microdata contributes to exactly one of the cells of a table. A dichotomous variable y_{ip} will be used which is one if record i contributes to cell p and zero otherwise.
A simple population estimator is given by

$$\hat{\boldsymbol{t}}_y^w = \sum_{i=1}^{n} w_i \boldsymbol{y}_i$$

where $\boldsymbol{y}_i$ is a P-vector, containing the elements y_{ip} for $p = 1, \dots, P$. The estimator $\hat{\boldsymbol{t}}_y^w$ is obtained by aggregation of starting weights of the data set used for estimation. Initially, we assume that all elements $\hat{\boldsymbol{t}}_y^w$ are strictly larger than zero, meaning that for each cell at least one record is available that contributes to that cell. This assumption is purely made for ease of explanation. We will consider the case with zero valued cell estimates at the end of this subsection.
The so-called initial table estimate $\hat{\boldsymbol{t}}_y^w$ is independent of all other tables and registers and is not necessarily consistent with other tables. To realize consistency, a population estimate needs to be calibrated on all marginal totals that the table has in common with all registers and with all previously estimated tables. These marginal totals are denoted by the J-vector $\boldsymbol{r}$.
There is a relationship between the cells of a table and its marginal totals: a marginal total is a collapsed table that is obtained by summing along one or more dimensions. Each cell contributes to a specific marginal total or it does not. The relation between the P cells and the J marginal totals is expressed in an $(J \times P)$ − aggregation matrix $\boldsymbol{L}$. An element l_{jp} is 1 if cell p of the target table contributes to marginal total j and zero otherwise.
A table estimate $\hat{\boldsymbol{t}}_y$ is consistent if it satisfies

$$\boldsymbol{L}\hat{\boldsymbol{t}}_y = \boldsymbol{r}. \qquad (1)$$

Usually, initial estimates $\hat{\boldsymbol{t}}_y^w$ do not satisfy (1), otherwise no adjustment would be necessary.

Therefore, our aim is to find a table estimate $\hat{\boldsymbol{t}}_y^*$ that is in some sense close to $\hat{\boldsymbol{t}}_y^w$ and that satisfies all consistency constraints. The well-established technique of least-square adjustment can be applied to find such an adjusted estimate. In this approach, a consistent table estimate $\hat{\boldsymbol{t}}_y^*$ is obtained as a solution of the following minimization problem

$$\begin{aligned} &\min_{\hat{\boldsymbol{t}}_y^*} \ \left(\hat{\boldsymbol{t}}_y^* - \hat{\boldsymbol{t}}_y^w\right)' \mathbf{W}^{-1} \left(\hat{\boldsymbol{t}}_y^* - \hat{\boldsymbol{t}}_y^w\right), \\ &\text{such that: } \boldsymbol{L}\hat{\boldsymbol{t}}_y^* = \boldsymbol{r}. \end{aligned} \tag{2}$$

where $\mathbf{W}$ is a symmetric, non-singular weight matrix.

Despite that several alternative methods can be applied as well (see e.g. Deville and Särndal, 1992 and Little and Wu,1991), the Generalised Least Squares (GLS) problem in (2) has a long and solid tradition in official statistics. It is applied in many areas. An example, in the field of macro-economics, is the reconciliation of National Accounts data. The formulation of the corresponding minimal adjustment problem as a GLS is known as the method of Stone (e.g. Stone *et al*. 1942, Sefton and Weale 1995, Wroe *et al.* 1999, Magnus *et al.* 2000, United Nations 2000, and Bikker *et al.* 2011).

A closed-form expression for the solution of the problem in (2) can be obtained by the Lagrange Multiplier method (see e.g. Mushkudiani *et al.* 2014). This expression is given by

$$\hat{\boldsymbol{t}}_y^{opt} = \hat{\boldsymbol{t}}_y^w + \boldsymbol{W}\boldsymbol{L}'(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}')^{-1}\left(\boldsymbol{r} - \boldsymbol{L}\hat{\boldsymbol{t}}_y^w\right). \tag{3}$$

When estimating a single table the RW-solution corresponds to the GREG-estimator. The GREG-estimator is obtained as special case of (3) in which $\mathbf{W}$ is set to $\widehat{\boldsymbol{T}}$, where $\widehat{\boldsymbol{T}} = \text{Diag}\left(\hat{\boldsymbol{t}}_y^w\right)$, a diagonal matrix with the entries of $\hat{\boldsymbol{t}}_y^w$ along its diagonal (see Deville and Särndal, 1992 and Mushkudiani *et al*., 2014). Thus, we obtain the following expression for the RW-estimator.

$$\hat{\boldsymbol{t}}_{\mathrm{y}}^{\mathrm{RW}} = \hat{\boldsymbol{t}}_{\mathrm{y}}^{\mathrm{w}} + \widehat{\boldsymbol{T}}\boldsymbol{L}'\left(\boldsymbol{L}\widehat{\boldsymbol{T}}\boldsymbol{L}'\right)^{-1}\left(\boldsymbol{r} - \boldsymbol{L}\hat{\boldsymbol{t}}_y^w\right), \tag{4}$$

In writing (4), it is assumed that the inverse of square matrix $\boldsymbol{L}\widehat{\boldsymbol{T}}\boldsymbol{L}'$ is properly defined. In practise, this is however not always true. When the constraint set in (1) contains any redundancies, i.e. constraints that are implied by other constraints, $\boldsymbol{L}\widehat{\boldsymbol{T}}\boldsymbol{L}'$ will be singular. In that case, it may still be possible to apply (4) by using a generalised inverse (see e.g. Ben-Israel and Greville, 2003).

As an alternative to minimizing adjustment at cell level, the RW solution can also be obtained by adjustment of underlying weights. Deville and Särndal (1992) show that a set of corrected weights w_{ip}^* can be derived such that the RW table estimate $\hat{\boldsymbol{t}}_{\mathrm{y}}^{\mathrm{RW}}$ can be obtained by weighting the underlying microdata. That is, such that:

$$\left(\hat{\boldsymbol{t}}_y^{RW}\right)_p = \sum_{i=1}^{n} w_{ip}^* y_{ip}. \tag{5}$$

Besides reconciled table estimates, RW also provides measures to estimate the precision of these estimates. Variances of table estimates can be estimated. We refer to Houbiers et al. (2003) and Boonstra (2004) for mathematical expressions.

In the beginning of the subsection we assumed that all initial cell estimates in $\hat{\boldsymbol{t}}_y^w$ are strictly positive. We will now consider the more generic case in which $\hat{\boldsymbol{t}}_y^w$ may include zero valued initial estimates. Zero valued cells are cells to which no micro unit contributes. There are no weights associated with those cells. Because RW is a weight adjustment method, it follows that RW does not adjust zero valued initial estimates. The expression for the RW-estimator in (4) may however still be used in case of zero

valued cell estimates, as it can easily be derived that initial estimates of zero remain zero in (4)[1].
However, in presence of zero-valued initial estimates, the so-called empty cell problem may occur. This happens if there is a constraint imposing a sum of variables that each has a zero initial estimate to align with a nonzero value in r. Because in RW zero values cannot be adjusted achieving consistency is impossible. The RW estimator in (4) is undefined because $\mathbf{L}\widehat{\mathbf{T}}\mathbf{L}'$ will be singular, as it includes an all zeroes row. Consequently, the originally proposed RW-method cannot be applied in that case.
The empty cell problem can be tackled by the epsilon method: a technical solution described by Houbiers (2004), based on the pseudo-Bayes estimator of Bishop et al. (1975) for log-linear analysis. The epsilon method means that zero-valued estimates in an initial table are replaced by small, artificial, non-zero "ghost" values, which were set to one for all empty cells in the 2011 Census tables. In other words, it was assumed a priori that each empty cell is populated by one fictitious person.

2.4 Problems with repeated weighting

Below we summarise complications of RW, besides the above-mentioned empty cell problem. Problems that are inherent to the sequential way of estimation are described first, then other complications are given.

Problem 1. Impossibility of consistent estimation
A first problem of RW is that, after a number of tables have been estimated, it may become impossible to estimate a new one. Earlier estimated tables impose certain consistency constraints on a new table, which reduces the degree of freedom for the estimation of that new table. When a number of tables have already been estimated it may become impossible to satisfy all consistency constraints at the same time. The problem is also known in literature. It has been described by Cox (2003) for the estimation of multi-dimensional tables with known marginal totals.

Example
One wants to estimate the table country of citizenship × industry of economic activity × educational attainment. Citizenship and industry are observed in a register, educational attainment comes from a survey. According to the register there are: 10 persons from Oceania and 51 persons working in the mining industry. The combination Oceania and mining industry is observed for four persons. The following marginal totals are derived from previously estimated tables

1 This follows because the relevant rows in $\widehat{\mathrm{T}}\mathrm{L}'\left(\mathrm{L}\widehat{\mathrm{T}}\mathrm{L}'\right)^{-1}$ contain zeros only.

2.4.1 Population by Citizenship and Education

Citizenship	Education	Count
Oceania	Low	1
Oceania	High	9

2.4.2 Population by Industry and Education

Industry	Education	Count
Mining	Low	49
Mining	High	2

By combining both tables it can be seen that the combination Oceania & mining industry can occur three times at most; there cannot be more than two highly educated people and one lowly educated person. This contradicts results from the register that states that there are four “mining” persons from Oceania.

Problem 2. Suboptimal solution
In the RW-approach the problem of estimating a set of coherent tables is split into a number of sub problems, in each of which one table is estimated. Because of the sequential approach, a suboptimal overall solution may be obtained, that deviates more from the data sources than necessary.

Problem 3. Order dependency
The order of estimation of the different tables matters for the outcomes. Besides that ambiguous results are not desirable as such, it can be expected that there is a relationship between the quality of the RW-estimates and the order of estimation, as tables that are estimated at the beginning of the process do not have to satisfy as many consistency constraints as tables that are estimated later in the process.

In addition to the aforementioned problems, there are also some other problems that are not directly caused by sequential estimation.
A first problem is that although RW achieves consistency between estimates for the same variable in different tables, the method does not support consistency rules between different variables (so-called ‘edit rules’). An example of such a rule is that the number of people who have never resided abroad cannot exceed the number of people born in the country concerned.
A second complication is that RW may yield negative cell estimates. In many practical applications, such as the Dutch Census, negative values are however not allowed.
A third complication is the previously mentioned empty cell problem. As mentioned in Subsection 2.3, this problem occurs when estimates have to be made without underlying data.

3. Alternative methods

This section demonstrates that the consistent estimation problem can alternatively be solved by available techniques from Operations Research (OR).

3.1 Repeated weighting as a QP problem

The repeated weighting estimator in (4) may also be obtained as a solution of the following quadratic programming problem (QP).

$$\min_{\hat{\boldsymbol{t}}_y^*} \sum_{i:\,(\hat{\boldsymbol{t}}_y^w)_i>0} \frac{1}{\left|(\hat{\boldsymbol{t}}_y^w)_i\right|}\left((\hat{\boldsymbol{t}}_y^*)_i - (\hat{\boldsymbol{t}}_y^w)_i\right)^2,$$
$$\text{such that:} \qquad (6)$$
$$\boldsymbol{L}\hat{\boldsymbol{t}}_y^* = \boldsymbol{r},$$
$$(\hat{\boldsymbol{t}}_y^*)_i = 0 \quad \text{for } i \text{ with } (\hat{\boldsymbol{t}}_y^w)_i = 0.$$

The objective function minimizes squared differences between reconciled and initial estimates. The constraints are the same as in RW. The last mentioned type of constraint ensures that the zero initial valued estimates are not adjusted.

Main advantage of the QP-approach is its computational efficiency. Unlike the closed-form expression of the RW estimator (4), Operations Research methods do not rely on matrix inversion. Therefore, very efficient solution methods are available (e.g. Nocedal and Wright, 2006). Operations Research methods are available in efficient software implementations ('solvers'), that are able to deal with large problems. Examples of well-known commercial solvers are: XPRESS (FICO, 2009) , CPLEX (IBM, 2015) and Gurobi (Gurobi Optimization Inc., 2016). Bikker *et al*. (2013) apply such solvers for National Accounts balancing; an application that requires solving a quadratic optimization problem of approximately 500,000 variables.

A second advantage of the QP-approach is that it can still be used in case of redundant constraints. Contrary to the WLS-approach, there is no need to remove redundant constraints, or to apply sophisticated techniques like generalised inverses.

A third advantage is that QP can be more easily generalised than WLS to include additional requirements. Inequality constraints can be included in the model to take account of non-negativity requirements and edit rules (see Subsection 2.4). The empty cell problem can be dealt with by the following slight modification of the objective function

$$\min_{\hat{\boldsymbol{t}}_y^*} \textstyle\sum_{i=1}^{P} \frac{1}{(\hat{\boldsymbol{t}}_y^{w*})_i}\left((\hat{\boldsymbol{t}}_y^*)_i - (\hat{\boldsymbol{t}}_y^w)_i\right)^2,$$
$$\text{such that: } \boldsymbol{L}\hat{\boldsymbol{t}}_y^* = \boldsymbol{r}. \qquad (7)$$

where $\hat{\boldsymbol{t}}_y^{w*} = \max\left(\left|\hat{\boldsymbol{t}}_y^w\right|, 1\right)$. The solution in (7) is less radical than replacing each initial estimate with one, the solution that was applied for the 2011 Dutch census.

Disadvantages of the QP-approach are that the method does not provide means to derive corrected weights and to estimate variances of reconciled tables. However, because of the equivalence of the QP and the WLS formulation of the problem, it

follows that, although corrected weights are not obtained in a solution of a QP-problem, these weights do exist from a theoretical point of view.

3.2 Other approaches

The quadratic programming formulation in (6) is commonly applied. A more general formulation of the objective function is given by $\min_{\hat{\boldsymbol{t}}^*} \sum_{i=1}^{P} \frac{1}{d_i} \left(\hat{t}_i^* - \hat{t}_i\right)^2$.

Stone *et al.* (1942) prove that the most precise results are obtained when weights d_i are set to the reciprocal of the variances of the initial estimates. In practice, (estimates of) these variances are often not available. In the absence of any information about data reliabilities, three weight definitions are often mentioned: $d_i^{(0)} = 1$, $d_i^{(1)} = |\hat{t}_i|$ and $d_i^{(2)} = (\hat{t}_i)^2$, see e.g. Boonstra (2004). The model in (6) is obtained after choosing $d_i^{(1)}$. There is an ongoing debate about these alternatives in literature.

The alternative $d_i^{(0)}$ minimizes the sum of squared absolute adjustments. It has the advantage that it can still be applied when initial values of zero occur. However for many practical applications it may be considered undesirable that the amount of adjustment is unrelated to the initial cell values.

The second alternative $d_i^{(1)}$ results in adjustments that are proportionate to the magnitude of the cells. It leads to results that closely approximate the well-known IPF-method (Stephan, 1942).

The third alternative, $d_i^{(2)}$, minimizes the sum of relative adjustments. It assumes that different cells have the same coefficients of variation (CV), meaning that standard errors are proportionate to the magnitude of the cells.

Statistics Netherlands used to apply weights $d_i^{(2)}$ for reconciling National Accounts (see Bikker *et al.*, 2013), a choice that is also preferred in Di Fonzo and Marini (2009). However, recently Statistics Netherlands changed this into $d_i^{(1)}$. There are several reasons for preferring $d_i^{(1)}$ over $d_i^{(2)}$.

Firstly, in estimating counts of a certain subpopulation small values are often associated with a small number of sample survey observations and therefore these are often not very precisely measured. When using $d_i^{(2)}$ it may happen that relatively small values are hardly adjusted, which does not conform with the relatively low reliability of measurement.

Secondly, $d_i^{(1)}$ has the desirable property that, relative importance of a group of variables does not change after (dis)aggregation. When aggregating two or more cells to a single total, the share of the weight of the total is the same as the share of the sum of the weights of its components. This property is especially important in this paper, because in Subsection 5.2 a Divide-and-Conquer method is presented based on aggregation of part of the variables.

Thirdly, there are examples in literature of undesirable results when choosing $d_i^{(2)}$, see e.g. Fortier and Quenneville (2006). It may happen that relative sizes change. Consider an example with four variables, having initial values of 10, 10, 10 and 20, respectively. These four variables have to align with a total of 20. Under $d_i^{(1)}$ and $d_i^{(2)}$ the result are: 4, 4, 4, 8 and 5.7, 5.7, 5.7, 2.9, respectively. In the first case relative adjustment are the same, i.e. -60% for each entry. Hence, relative sizes are

preserved. In the latter case relative adjustment of the fourth variable is much larger than for the first three entries, leading to a result in which the largest value becomes the smallest one.
We will use $d_i^{(1)}$ in the remainder of this paper, because of the equivalence with the WLS approach and because of the above-mentioned three reasons.
As an alternative to the QP-approach, the well-known Iterative Proportionate Fitting (IPF) (Deming and Stephan, 1942) can be applied, a method that was originally developed for the 1940 U.S. Census and that is popularly applied for a wide range of applications. In this paper we choose a QP-approach, since a QP-based model has a longer and more solid tradition in Official Statistics and can be more easily generalised to deal with additional requirements like inequality constraints and solutions for the empty cell problem.

4. Simultaneous approach

In this section we argue that the three problems mentioned in Subsection 2.4 ("Impossibility of consistent estimation", "Suboptimal solution" and "Order dependency") that are inherent to the sequential way of estimation can be circumvented in an approach in which all tables are estimated simultaneously.
The QP-model in (6) can be easily generalized for the consistent estimation of a table set.
That is, a consistent table set can be obtained as a solution to the following problem:

$$\min_{\hat{t}^{SW}} \sum_{i:\, \hat{t}_i^w > 0} \frac{1}{\hat{t}_i^w} (\hat{t}_i^{SW} - \hat{t}_i^w)^2,$$
such that:
$$\boldsymbol{L}\hat{\boldsymbol{t}}^{SW} = \boldsymbol{r}.$$
$$\left(\hat{\boldsymbol{t}}^{SW}\right)_i = 0, \text{for } i \text{ with } \left(\hat{\boldsymbol{t}}_y^w\right)_i = 0 \tag{8}$$

In this formulation $\hat{\boldsymbol{t}}^{SW} = \left(\hat{\boldsymbol{t}}_{1,\ldots,}^{SW} \hat{\boldsymbol{t}}_N^{SW}\right)'$, a vector containing estimates for the cells of all N tables, similarly $\hat{\boldsymbol{t}}^{w} = \left(\hat{\boldsymbol{t}}_{1,\ldots,}^{w} \hat{\boldsymbol{t}}_N^{w}\right)'$, a vector of initial estimates. The subscript SW stands for simultaneous weighting, as opposed to RW, which stands for repeated weighting.
The objective function minimises a weighted sum of squared differences between initial and reconciled cell estimates for all tables. The constraints impose marginal totals of estimated tables to be consistent with known population totals from registers and estimated tables to be mutually consistent. The former means that for each table all marginal totals with known register totals are consistently estimated with those register totals. The latter means that for each pair of two distinct tables all common marginal totals have the same estimated counts. These constraints impose a sum of cells in one table to be equal to a sum of cells in another table, where the value of that sum is not known in advance. For comparison, in RW marginal totals of one table need to have the same value as known marginal totals from earlier estimated tables. Analogous to the RW-model in (6), the SW-model in (8) can easily

be extended to take account of additional requirements, like non-negativity of estimated cell values, edit rules and the empty cell problem.
It can be easily seen that Problems 1, 2 and 3 in Subsection 2.4 do not occur if all tables are estimated simultaneously. An optimal solution for the entire set of tables is obtained if it exists. From a practical point it is more attractive to solve one problem in SW rather than several problems in RW. A SW-approach may however not always be feasible in practise. A large estimation problem needs to be solved consisting of many variables and constraints. The capability of solving such large problems may be limited due to computer memory size. Because SW applications may still be infeasible even for modern computers, we focus on ways of splitting the problem up into a number of smaller sub problems. To prevent estimation problems, our purpose is to split the problem as much as possible into independent sub problems.

5. Divide-and-Conquer algorithms

In this section two so-called Divide-and-Conquer (D&C) algorithms will be presented for estimating a set of coherent frequency tables. These algorithms recursively break down a problem into sub problems that can each be more easily solved than the original problem. The solution of the original problem is obtained by combining the sub problem solutions.

5.1 Splitting by common variables

The main idea of our first algorithm is that an estimation problem, with one or more common register variable(s) can be split into a number of independent sub problems. Each of those sub problems belongs to one (combination of) category(ies) for the common register variable(s). For example, if sex were included in all tables of a table set, a table set can be split into two independent sets: one for men and one for women.
In practice, it is often not the case that a table set has one or more common register variables in each table. Common variables can however always be created by adding variables to tables, provided that a data source is available from which the resulting, extended tables can be estimated. In our previous example, all tables that do not include sex can be extended by adding this variable to the table. In this way, the level of detail increases, meaning that more cells need to be estimated than in the original problem. This increase in detail may lead to less precise results at the required level of publication. However, at the same time, the possibility is created of splitting a problem into independent sub problems. Since all ‘added’ variables are used to split

the problem, one can easily understand that the number of cells in each of these sub problems cannot exceed the total number of cells of the original problem.
For any practical application the question arises which variable should be chosen as "splitting" variables. Preferably, this should be variables that appear in most tables, e.g. sex and age in the Dutch 2011 Census, as this choice leads to the smallest total number of cells to be estimated.
The approach is especially useful for a table set with many common variables, because in that case the number of added cells remains relatively limited.
The proposed algorithm has the advantage over Repeated Weighting that the sub problems that are created can be solved independently. For this reason there are no problems with "impossibility of estimation" (Problem 1 in Subsection 2.4) and "order-dependency of estimation" (Problem 3 in Subsection 2.4). Problem 2 "Suboptimal solution" is not necessarily solved. This depends on the need of adding additional variables to create common variables. If a table set contains common register variables in each table and the estimation problem is split using these common variables, an optimal solution is obtained. However, if common variables are created by adding variables to tables, extended tables are obtained, for which the optimal estimates do not necessarily comply with the optimal estimates for the original tables.

5.2 Aggregation and disaggregation

A second divide-and-conquer algorithm consists of creating sub problems by aggregating categories of one or multiple variables. In the first stage, categories are aggregated (e.g. estimating 'educational attainment' according to two categories rather than the required eight). In a second stage, table estimates that include the aggregation variable(s) are further specified according to the required definition of categories.
Since the disaggregation into required categories can be carried out independently for each aggregated category, a set of independent estimation problems is obtained in the second stage.
The following example clarifies the idea. Consider the following tables

- Table 1: educational attainment × age;
- Table 2: educational attainment × sex;
- Table 3: sex × occupation.

The required categories for educational attainment are: 1,…,8. Two aggregated categories I and II are defined; category I comprises the original categories 1,…,4 and category II the other categories 5,…,8.
In the first stage, the three tables are estimated using aggregated categories for educational attainment. Then, in the second stage, tables 1 and 2 are re-estimated using original categories for educational attainment. In this way, two independent estimation problems are obtained, one for educational attainment categories 1, 2, 3 and 4, the other one for categories 5, 6, 7 and 8.
When estimating tables in the second step, it needs to be ensured that results are consistent with the tables that are estimated in the first stage.
In previous example one variable was aggregated, educational attainment. It is however also possible to aggregate multiple variables. In that case a multi-step

method is obtained, in which in each stage after Stage 1, one of the variables is disaggregated.
Because of these dependencies of the estimation processes in different stages, it cannot be excluded that the three problems of Section 2.5 occur. However, the problems may have a lower impact than in Repeated Weighting. This is because of a lower degree of dependency between different estimation problems. In RW each estimated table may be dependent on all earlier estimated tables, whereas in the proposed D&C approach, estimation of a certain sub problem only depends on one previously solved problem.

6. Application

In this section we present results of a practical application of the proposed D&C methods to the Dutch 2011 Census tables. Our aim is to test the feasibility of the methods, as well as to compare results with the officially published results that are largely based on RW. Subsection 6.1 describes backgrounds of the Dutch Census. Subsection 6.2 explains the setup of the tests and Subsection 6.3 discusses results.

6.1 Dutch 2011 Census

According to the European Census implementing regulations, Statistics Netherlands was required to compile sixty high-dimensional tables for the Dutch 2011 Census, for example, the frequency distribution of the Dutch population by age, sex, marital status, occupation, country of birth and nationality. Census tables contain demographic, housing and commuting variables. The tables are very detailed, comprising five, six, sometimes seven, and even nine dimensions. An example of a cell in one of the tables: the number of 36 year-old male widowed managers, born in Portugal with the Polish nationality. Since the sixty tables contain many common variables 'standard weighting' does not lead to consistent results.
Several data sources are used for the Census, but after micro-integration, two combined data sources are obtained: one based on a combination of registers and the other one is a combination of sample surveys. From now on, when we refer to a Census data source, a combined data source is meant after micro-integration. The 'register' data sources cover the full population (in 2011 over 16.6 million persons) and include all relevant Census variables except 'educational attainment' and 'occupation'. For the 'sample survey' data sources it is the other way around, it covers all relevant Census variables, but it is available for a subset of 331,968 persons only.
Repeated weighing is applied to the tables that need to be estimated from a sample survey. These are 42 tables with person variables that include ´educational attainment´ and/or 'occupation'. The target population of these tables consists of the registered Dutch population, with the exception of people younger than 15 years. Young children are excluded because the two sample survey variables 'educational

attainment' and 'occupation' are not relevant for these people. Hence, all required information for these people are not estimated by using repeating weighting, but these are directly compiled by counting from registers.
The total number of cells in the 42 tables amounts to 1,047,584, the number of cells within each table ranges from 2,688 to 69,984.
The following measures were taken to prevent estimation problems:
To prevent "impossibility of consistent estimation" (Problem 1 in Subsection 2.4) a specific order of estimation was determined, after long trial-and-error. The occurrence of the problem was further avoided by reducing the total number of estimated tables. This was done by merging a few tables into a large table.
To solve the empty cell problem, the aforementioned "epsilon method" was applied, meaning that all initial zero cell estimates were replaced by one.
To obtain nonnegative cell estimates, iterative proportional fitting was used as an estimation method rather than weighted least squares.
Order-dependency and sub-optimality of results (Problems 2 and 3 in Subsection 2.4) were accepted. For a more extensive description we refer to Schulte Nordholt *et al.*(2014).

6.2 Setup

Below we explain how the two D&C algorithms were applied to the 2011 Dutch Census.

Setup 1 - Splitting by common variables
In this setup, the original table set is split into 48 table sets, by using geographic area (12 categories), sex (2 categories), and employment status (2 categories)[2] as splitting variables. Each of the 48 table sets contains a subset of the 42 Census tables, determined by the categories of the splitting variables. For each of the combined categories of splitting variables - one of them is for instance: North Holland & male & employed - a table set is estimated independently from other table sets.
The three splitting variables are however not present in all 42 Census tables. In 13 tables geographic area does not occur and in one table sex is absent. Tables that do not include the three splitting variables were extended by incorporating missing variables. As a result, the total number of cells in the 42 tables was increased from 1,047,584 to 4,556,544.
The largest optimization problem in this procedure consists of 79,315 variables and 137,493 constraints[3].

[2] Employment status is not an official Census variable. It can be obtained by aggregating occupation activity into two categories: employed / unemployed.
[3] One may notice that 48*79,315 is less than 4,556,544, the total number of cells. An explanation for this is that cells whose value necessarily have to be zero – which are cells that need to align with a known population count of zero -are not included in the optimization problem. These cells are not included in the number 79,315.

Setup 2 - Aggregation and disaggregation
In this setup educational attainment (8 categories) and occupation (12 categories) were selected for aggregation of categories. Initially, both variables are aggregated into two main categories, that each contain half of the categories of the original variables. Thereafter, results were obtained for the required categories for the two aggregation variables.
Five optimization problems are defined in this procedure. In the first problem a table set is estimated based on aggregated categories for educational attainment and occupation. In each of the following stages either one of the two aggregated categories for educational attainment or occupation is disaggregated into required categories.
The largest optimization problem consists of 183,432 variables and 361,830 constraints. Less sub problems are defined than in "Setup 1 - Splitting by common variables", resulting in a larger problem size for each sub probem.

Results for the two D&C algorithms are derived using the QP-approach that is described in Sections 3 and 4. Apart from the method of splitting up the estimation problems into sub problems, the most relevant difference with the method used for estimating the Dutch 2011 Census is about the way the empty cell problem is tackled. In the Dutch 2011 Census all initial cell estimates of zero were replaced with one (see also Subsection 2.3), whereas the two D&C methods adopt a less rigorous approach, consisting of a slight adjustment of the QP objective function as shown in (7).

6.3 Results

In this subsection we compare results of the two D&C methods with the RW-based method as applied to the official 2011 Census. All practical tests were conducted on a 2.8 GHZ computer with 3.00 GB of RAM.
Our first conclusion is that all optimization problems that were defined by the two D&C approaches were successfully solved; no estimation problems were experienced. Thus, we arrive at our main conclusion that the D&C approaches do not suffer from the important drawback of RW that at some point in estimation it may become impossible to estimate a new table consistently with all previously estimated tables. This is an important advantage for practitioners as it brings about considerable time savings in the implementation of the method.
We now continue with a comparison of the reconciliation adjustments. The criterion used to compare degree of reconciliation adjustment is based on the QP objective function in (7), a sum of weighted squared differences between initial and reconciled estimates, given by

$$\sum_{i=1}^{P} \frac{1}{\left(\hat{\boldsymbol{t}}_y^{w*}\right)_i} \left(\left(\hat{\boldsymbol{t}}_y^{*}\right)_i - \left(\hat{\boldsymbol{t}}_y^{w}\right)_i \right)^2 , \tag{9}$$

where $\hat{\boldsymbol{t}}_y^{w}$ is a vector with initial estimates, $\hat{\boldsymbol{t}}_y^{*}$is a vector with reconciled estimates, $\left(\hat{\boldsymbol{t}}_y^{w*}\right)_i$= max $\left(\left|\left(\hat{\boldsymbol{t}}_y^{w}\right)_i\right|, 1\right)$ and the summation is over all relevant categories.

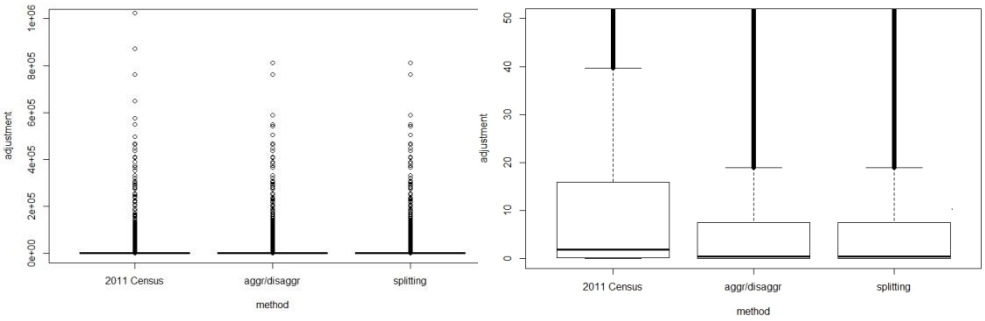
Table 6.3.1 compares total adjustment, as defined according to (9), based on all cells in all 42 estimated tables.

6.3.1 Total adjustment by three methods

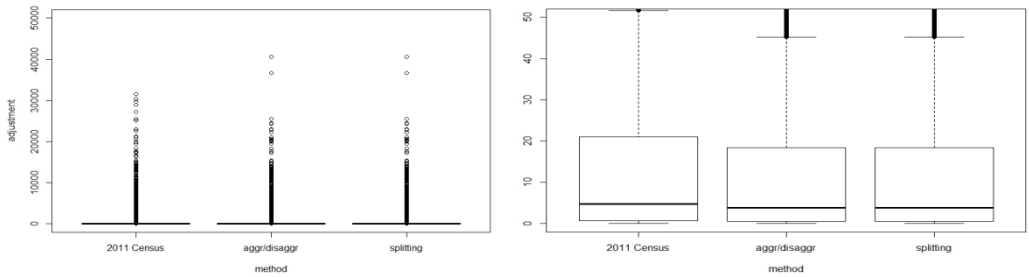
Method	Total adjustment	
	All cells	Cells with initial estimate larger than zero
Dutch 2011 Census	1.1E+08	1.3E+07
Splitting by common variables	8.8E+07	1.2E+07
Aggregation and Disaggregation	7.0E+07	1.2E+07

The RW method, as applied to the 2011 Census, actually leads to a suboptimal result, as the amount of reconciliation adjustment is larger than for the D&C methods. The result that "Aggregation and Disaggregation" method gives rise to a better solution than "Splitting by common variables" can be explained by the lower amount of sub-problems that are defined in the chosen setups.
However, if we only compare cells with larger than zero initial estimates, differences between three methods become very small. This shows that the way how original estimates of zero are processed is more important in explaining the differences in results than the way how the estimation problem is broken down into sub problems.
The boxplots in 6.3.1 and 6.3.2 compare adjustment at the level of individual cells. It can be seen that the amount of relatively small corrections is larger for the two D&C methods than for the RW-based method used for officially published Census tables. Differences in results are however smaller again, if zero initial estimates are not taken into account.

6.3.2 Boxplots of adjustments to all cells of 42 Census tables.The right panel zooms in on the lower part of the left panel.



6.3.3 Boxplots of adjustments to cells with nonzero initial estimates. The right panel zooms in on the lower part of the left panel.



In summary, besides avoidance of estimation problems, the two newly proposed D&C methods have the advantage of slightly smaller reconciliation adjustment for the Dutch 2011 Census estimation. The largest improvement of results can be attributed to a better treatment of initial estimates of zero.

7. Discussion

When several frequency tables need to be produced from multiple data sources, the problem may arise that results are numerically inconsistent. That is, that different results are obtained for the same marginal totals in different tables.
To solve this problem, Statistics Netherlands developed a Repeated Weighting (RW) method for the consistent estimation of a set of tables. This method was applied to the 2001 and 2011 Dutch census. However, the scope of applicability of this method is limited by several known estimation problems. In particular, the sequential way of estimation causes problems. As a result, estimation of the 2011 Census was troublesome. A suitable order of estimation was found after long trial and error.
This paper presents two D&C methods that can be more easily applied than RW. Contrary to RW, the two D&C methods break down the estimation problem as much as possible into independently estimated parts, rather than the dependently estimated parts that are distinguished in RW. In this way estimation problems may be prevented. One of the two newly developed methods partitions a given table set according to categories of variables that are contained in each table. The other method is based on aggregation and disaggregation of categories. An application to 2011 Census tables showed that estimation problems that were experienced with RW are actually avoided. Thus, the key message of this paper is when estimating a coherent set of tables, there can be smarter ways of breaking down the problem than estimating single tables in sequence.
However, if a simultaneous estimation method is computational feasible, such an approach is to be preferred. Most importantly, because estimation problems are avoided that are inherent to a sequential process. A second reason is that, if the problem is solved as a whole, an optimal solution is obtained with minimal adjustment from the data sources. A third reason is that from a practical point of

view solving one (or few) optimization problem(s) is much easier than solving many problems.
The newly developed methods seem to be feasible for the upcoming Dutch 2021 Census. According to the current plans, 32 tables with educational attainment and/or occupation are required to be compiled, variables that were estimated from sample surveys in the previous 2011 Census. The total number of cells in all these tables amounts to 720,288, a lower number than in the previous 2011 Census tables. If a simultaneous estimation method is not feasible at the time of 2021 Census estimation, the D&C method based on partitioning by common variables seems to be most appropriate. According to the current planned definition of tables, the 32 tables with one or two sample survey variable(s), educational attainment and occupation, can be easily subdivided according to the combined categories of geographic area (12 categories) and sex (2 categories). All 32 tables contain sex and geographic area is missing in two tables only. The proposed D&C method can however still be applied after extending those two tables with geographic area. A total number of 764,640 cells is obtained in the resulting set of tables. These cells can be allocated to 24 independently estimated sub problems. For comparison, in the 2011 Census a total number of 4,556,544 cells were estimated that were subdivided into 48 independently estimated problems. From this it follows that the optimization problems for the 2021 Census can be expected to have a smaller size than the problems for the previous 2011 Census. As the newly proposed D&C method turned out to be feasible for the 2011 Census, it can also be expected to be feasible for the 2021 Census.

Acknowledgements

The author would like to thank Ton de Waal, Tommaso di Fonzo, Reinier Bikker, Nino Mushkudiani and Eric Schulte Nordholt for their helpful comments on previous versions of this paper.

8. References

Bakker, B.F.M. (2011). *Micro integration*. Statistical Methods (201108). Statistics Netherlands, The Hague/Heerlen. http://www.cbs.nl/NR/rdonlyres/DE0239B4-39C6-4D88-A2BF-21DB3038B97C/0/2011x3708.pdf

Bakker, B.F.M., J. van Rooijen and L. van Toor (2014). The system of social statistical datasets of Statistics Netherlands: an integral approach to the production of register-based social statistics. *Statistical journal of the IAOS*, 30, 411-424.

Ben-Israel A. and T.N.E. Greville (2003). *Generalized inverses. Theory and applications*, Second Edititon, Springer Verlag, New York.

Bikker, R.P., J. Daalmans and N. Mushkudiani (2011), *Macro Integration – Data Reconciliation*. Statistical Methods Series, Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/90BAF023-74AE-4286-B7BB-1EFEC20D701E/0/2011x3704.pdf

Bikker R.P., J. Daalmans and N. Mushkudiani (2013), Benchmarking Large Accounting Frameworks: A Generalised Multivariate Model, *Economic Systems Research*, 25, 390-408.

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

Boonstra, H.J. (2004), *Calibration of Tables of Estimates*. Report, Statistics Netherlands.
Cox L.H. (2003). On properties of multi-dimensional statistical tables. *Journal of Statistical Planning and Inference*, 117, 251–273.

Daalmans, J. (2015), *Estimating Detailed Frequency Tables from Registers and Sample Surveys*. Discussion paper, Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/D4769B63-0D25-405D-8134-E11817D642EE/0/-201503DPestimatingdetailedfrequencytablesfromregistersandsamplesurveys.pdf

de Waal T. (2015), *General Approaches for Consistent Estimation based on Administrative Data and Surveys*, Discussion paper, Statistics Netherlands . https://www.cbs.nl/-/media/imported/documents/2015/37/2015-general-approaches-to-combining-administrative-data-and-surveys.pdf

de Waal T. (2016), Obtaining numerically consistent estimates from a mix of administrative data and surveys*, Statistical Journal of the IAOS*, 32, 231-243.

Deming W. and F. Stephan (1940), On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal totals are Known*. Annals of Mathematical Statistics*, 11, 427-444.

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

Di Fonzo, T. and M. Marini, Simultaneous and Two-step Reconciliation of Systems of Time Series: methodological and practical issues, (2010), *Journal of the Royal Statistical Society Series C*, 60, 143-164.

European Commission (2008). *Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses*. Official Journal of the European Union, L218, 14-20.

FICO (2009), *FICO (TM) Optimization suite, Xpress-Optimizer Reference manual*, Release 20.00, Fair Isaac cooperation, Warwickshire.

Fortier S. and B. Quenneville (2006), *Reconciliation and Balancing of Accounts and Time Series: From Concepts to a SAS Procedure*, In Joint Statistical Meeting 2009 Proceedings, Business and Economic statistics session. Alexandria, American Statistical Association, 130-144.

Gurobi Optimization Inc. (2016), *Gurobi Optimizer Reference Manual*.

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

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen and V. Snijders (2003), *Estimating Consistent Table Sets: Position Paper on Repeated Weighting*. Discussion paper, Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/6C31D31C-831F-41E5-8A94-7F321297ADB8/0/discussionpaper03005.pdf

IBM (2015*). ILOG CPLEX V 12.6 User's Manual for CPLEX*. IBM Corp.

Knottnerus P. and C. van Duin (2006). Variances in Repeated Weighting with an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, 565–584.

Little R.J.A. and M. Wu (1991), Models for Contingency Tables with known margins when Target and Sampled Populations Differ, *Journal of the American Statistical Association*, 86, 87-95.

Magnus, J.R., J.W. van Tongeren and A.F. de Vos (2000), National Accounts Estimation using Indicator Ratios, *The Review of Income and Wealth*, 3, p. 329-350.

Mushkudiani N., J. Daalmans and J. Pannekoek (2014), Macro-integration for Solving large Data Reconciliation Problems, *Austrian Journal of Statistics*, 43, 29-48.

Nieuwenbroek, N.J., R.H. Renssen & L. Hofman (2000). *Towards a generalized weighting system*. In: Proceedings, Second International Conference on Establishment Surveys, American Statistical Association, Alexandria VA.

Nocedal J. and S.J. Wright (2006), *Numerical Optimization*, Second Edition, Springer Verlag, New York.

Renssen, R.H. and N.J. Nieuwenbroek (1997), Aligning Estimates for Common Variables in two or more Sample Surveys. *Journal of the American Statistical Association*, 90, 368-374.

Särndal, C.E., B. Swensson and J. Wretman (1992). *Model assisted survey sampling*, Springer Verlag, New York.

Schulte Nordholt, E., J. van Zeijl, L. Hoeksma (2014). *The Dutch Census 2011: analysis and methodology*, Statistics Netherlands, The Hague,

http://www.cbs.nl/NR/rdonlyres/5FDCE1B4-0654-45DA-8D7E-807A0213DE66/-0/2014b57pub.pdf

Stephan, F. F. (1942), Iterative Method of Adjusting Sample Frequency Tables when Expected Margins Are Known, *The Annals of Mathematical Statistics*, 13, 166-178.

Stone, R, J.E. Meade and D.G. Champernowne (1942), The Precision of National Income Estimates. *The Review of Economic Studies*, 9, 111-125.

Sefton, J. and M.R. Weale (1995*), Reconciliation of national income and expenditure: balanced estimates for the United Kingdom, 1920-95*. Cambridge University Press, Cambridge.

United Nations, Statistics Division (2000), *Handbook of National Accounting: Use of Macro Accounts in Policy Analysis. Studies Methods*, United Nations, New York.

Wroe D., P. Kenny, U. Rizki and I. Weerakoddy (1999*), Reliability and Quality Indicators for National Accounts Aggregates.* Office for National Statistics (ONS). Document CPNB 265-1 for the 33rd meeting of the GNP Committee. http://ec.europa.eu/eurostat/documents/64157/4374310/40-RELIABILITY-AND-QUALITY-INDIC-NATIONAL-ACCOUNTS-AGGREGATES-1999.pdf/56dad551-849d-4ce1-8069-179a52558268

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

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

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

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