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

Jacco Daalmans

Acknowledgement and disclaimer
This work has partly been funded by the European Union ("Improvement of the use of administrative sources (ESS.VIP ADMIN WP6 Pilot studies and applications)", Grant Agreement number 07112.2015.002-2015.353)

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of the European Commission or of Statistics Netherlands. Neither the European Commission nor Statistics Netherlands are responsible for any use that may be made of the information this paper contains.
Index

1. Introduction 4

2. Repeated Weighting 6
   2.1 Prerequisites 6
   2.2 Non-technical description 7
   2.3 Technical description 8
   2.4 Dutch 2011 Census 9
   2.5 Problems with sequential estimation 10
   2.6 Example 11

3. Quadratic programming 12
   3.1 Repeated weighting as a QP problem 12
   3.2 Weights 12

4. Simultaneous weighting 14
   4.1 Formal description 14
   4.2 Discussion 15

5. Divide-and-conquer algorithms 16
   5.1 Splitting by common variables 16
   5.2 Aggregation and disaggregation 17

6. Practical applications 19
   6.1 Setup 19
   6.2 Results 20

7. Discussion 24

8. References 27

Appendix A. Estimation problems 30
   A.1 Description 30
   A.2 Known solutions for Repeated Weighting 30
   A.3 QP-Solutions for estimating a single table 32
   A.4 QP-Solutions for simultaneous estimation 33

Appendix B. Weights in general QP-problems 35

Appendix C. Auxiliary tables in a splitting approach 36

Appendix D. Aggregation and disaggregation 38
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 in practise, 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 statistical 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 already existing sample survey information is used. Because of the use of incomplete data, census compilation relies on estimation.

If standard estimation techniques were applied, numerically inconsistent results would be inevitable because of the different data sources that are used to estimate the different tables (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 consistent 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 2.5 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 applied 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.
Statistics Netherlands was awarded a Eurostat grant for this purpose. The work was carried out as part of the project "Improvement of the use of administrative sources (ESS.VIP ADMIN WP6 Pilot studies and applications)". This report presents the final results of the project.

We will present two Divide-and-Conquer algorithms that can be used as alternatives for RW. These algorithm break down the problem of estimating a large consistent table set into a number of smaller 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. In the new approach previously mentioned estimation problems do not occur, or at least have a smaller impact than in RW.

Thus, the new approach is much more easier to implement. A consistent table set can be obtained, if it is actually possible to define independent estimation problems. Hence, there is no need for determining problem-specific solutions that are often necessary in RW when estimating detailed tables.

Although some specific recommendations will be given for the upcoming 2021 Dutch Census, results in this report are more generally useful for all readers involved with the compilation of interrelated frequency tables of any kind of subject. The key message is that when estimating a coherent table set that cannot be estimated as a whole, there can be smarter ways of breaking down the problem than estimating single tables in sequence.

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 that will be the basis of two new Divide-and-Conquer methods that are explained in Section 6. Results of a practical application are given in Section 7 and Section 8 concludes this report with a discussion.
2. Repeated Weighting

In this section we explain the RW method. The assumptions of the method are presented in Subsection 2.1. Subsection 2.2 describes the main properties of the method. A technical description is given in Subsection 2.3. Subsection 2.4 describes the application of the method to the Dutch 2011 Census. In Subsection 2.5 a number of complications of the method are given. An example of one of these problems is given in Subsection 2.6.

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 over-coverage: 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 (2003) et al., 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 of 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.

Further, it is assumed that so-called edit rules are fulfilled within each data set (register or sample survey). Edit rules are relations between different variables, for instance the rule that place of work cannot be known for someone who does not work. In the remainder of this report (Appendix A.2) it will become clear why this assumption is necessary.

The two above mentioned assumptions are not automatically satisfied in practise. To obtain data sets that do satisfy these two assumptions 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 will be 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 consistently estimated. 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 (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 5 year-old professors. On the other hand, it is also the source of estimation problems, which will be demonstrated in Appendix A.
Besides reconciled table estimates, RW also provides means to estimate precision of these estimates. Variances of table estimates can be estimated.

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 $Y_1, \ldots, Y_p$. 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

$$\hat{t}_y^w = \sum_{i=1}^n w_i y_i$$

where $y_i$ is a $P$-vector, containing the elements $y_{ip}$ for $p = 1, \ldots, P$. The estimator $\hat{t}_y^w$ is obtained by aggregation of starting weights of the data set used for estimation.

Initially, we assume that all elements $\hat{t}_y^w$ 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 $\hat{t}_y^w$ 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 $r$.

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 $L$. An element $l_{jp}$ is 1 if cell $p$ of the target table contributes to marginal total $j$ and zero otherwise.

A table estimate $\hat{t}_y$ is consistent if it satisfies

$$L \hat{t}_y = r.$$  \hspace{1cm} (2.1)

Usually, initial estimates $\hat{t}_y^w$ do not satisfy (2.1), otherwise no adjustment would be necessary. Therefore, our aim is to find a table estimate $\hat{t}_y^*$ that is in some senses close to $\hat{t}_y^w$ 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 $\hat{t}_y^*$ is obtained as a solution of the following minimization problem

$$\min_{\hat{t}_y^*} \left( \hat{t}_y^* - \hat{t}_y^w \right)^T W^{-1} \left( \hat{t}_y^* - \hat{t}_y^w \right),$$

such that: $L \hat{t}_y^* = r$.

(2.2)

where $W$ 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 (2.2) 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 (2.2) can be obtained by the Lagrange Multiplier method (see e.g. Mushkudiani et al. 2014). This expression is given by

\[
\hat{\mathbf{y}}_{\text{RW}}^{\text{pt}} = \hat{\mathbf{y}}_{\text{w}} + \mathbf{W} \mathbf{L}^\prime (\mathbf{LW} \mathbf{L}^\prime)^{-1} (\mathbf{r} - \mathbf{L} \hat{\mathbf{y}}_{\text{w}}),
\]

When estimating a single table the RW-solution corresponds to the GREG-estimator. The GREG-estimator is obtained as special case of (2.3) in which \( \mathbf{W} \) is set to \( \hat{\mathbf{r}} \), where \( \hat{\mathbf{r}} = \text{Diag}(\hat{\mathbf{y}}_{\text{w}}) \), a diagonal matrix with the entries of \( \hat{\mathbf{y}}_{\text{w}} \) along its diagonal (see Deville and Särndal, 1992 and Mushkudiani et al., 2014). Thus, we obtain the following expression for the RW-estimator.

\[
\hat{\mathbf{y}}_{\text{RW}}^{\text{w}} = \hat{\mathbf{y}}_{\text{w}} + \mathbf{L}^\prime (\mathbf{L} \hat{\mathbf{r}} \mathbf{L}^\prime)^{-1} (\mathbf{r} - \mathbf{L} \hat{\mathbf{y}}_{\text{w}}),
\]

In writing (2.4), it is assumed that the inverse of square matrix \( \mathbf{L} \hat{\mathbf{r}} \mathbf{L}^\prime \) is properly defined. In practise, this is however not always true. When the constraint set in (2.1) contains any redundancies, i.e. constraints that are implied by other constraints, \( \mathbf{L} \hat{\mathbf{r}} \mathbf{L}^\prime \) will be singular. In that case, it may still be possible to apply (2.4) 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 \( \mathbf{w}^\prime \) can be derived such that the RW table estimate \( \hat{\mathbf{y}}_{\text{RW}} \) can be obtained by weighting the underlying microdata. That is, such that:

\[
\left( \hat{\mathbf{y}}_{\text{RW}}^{\text{w}} \right)_p = \sum_{i=1}^{n} \mathbf{w}_{ip} \mathbf{y}_{ip}.
\]

As noted in last subsection it is possible to estimate the variance of the RW estimator. 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 \( \hat{\mathbf{y}}_{\text{w}} \) are strictly positive. We will now consider the more generic case in which \( \hat{\mathbf{y}}_{\text{w}} \) 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 as a weight adjustment method, it follows that RW does not adjust zero valued initial estimates.

The expression for the RW-estimator in (2.4) 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 (2.4)

\[1\]. 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 \( \mathbf{r} \). Because in RW zero values cannot be adjusted achieving consistency is impossible. The RW estimator in (2.4) is undefined because \( \mathbf{L} \hat{\mathbf{r}} \mathbf{L}^\prime \) will be singular, as it includes an all zeroes row. Consequently, the originally proposed RW-method cannot be applied. We will come back to this problem later in Appendix A.

### 2.4 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. All Census 2011 tables are contingency tables, i.e. tables that display the number of times combinations of characteristics occur in the population, 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.

\[1\] This follows because the relevant rows in \( \mathbf{L} \hat{\mathbf{r}} \mathbf{L}^\prime \) contain zeros only.
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 obtained as a combination of sample surveys. When we refer to a Census data source from now on, 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 includes all relevant Census variables except ‘educational attainment’ and ‘occupation’. For the sample survey data source 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 is 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.

2.5 Problems with sequential estimation
Below we summarise complications that are inherent to the sequential way of estimation. These are problems that cannot be generally solved in a RW-context. After that we will describe several ad-hoc solutions that were applied in the 2011 Dutch Census. The text is for a large part based on Chapter 7 of Schulte Nordholt et al. (2014).

Problem 1. Impossibility of consistent estimation
A first problem arising in the estimation process 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 above-mentioned problem is very cumbersome in practise, as its occurrence becomes clear after some tables have already been estimated. The problem cannot be anticipated. When the problem occurs, all tables that have already been estimated need to be discarded. The problem that is described here is known in literature: it has been described by Cox (2003) for the estimation of multi-dimensional tables with known marginal totals. We will present an example of this problem below in Subsection 2.6.

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 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.
Problem 1 “Impossibility of consistent estimation” was circumvented in the 2011 Dutch Census by estimating tables in a specific order. A suitable order was found after long trial and error. Furthermore, a couple of tables with many overlapping variables were combined into one large table, consisting of the union of all variables of the original tables. Estimation problems were avoided, as a result of the reduced number of tables.

For Problem 3 “Order dependency” Houbiers et al. (2003) proposed a solution based on a so-called splitting up procedure. This means that all lower-dimensional marginal totals of all tables are estimated, before the tables themselves are estimated. If, for example, a three-dimensional table is to be estimated, first all one and two-dimensional marginal totals are estimated. A disadvantage is that a large amount of marginal totals is obtained for high-dimensional tables. The estimation of these tables can be problematic, in particular Problem 1 “Impossibility of consistent estimation” can be expected to occur often. For this reason it was decided not to apply the splitting up solution for the Dutch 2011 Census. Order-dependency of results was accepted.

In Appendix A.1 four other types of estimation problems are described that are not caused by the sequential approach of estimation. These problems are: nonnegativity requirements of estimated population counts, edit rules that need to be fulfilled, the empty cell problem and unacceptably large adjustment at aggregate level.

2.6 Example
As mentioned in Subsection 2.5, it may happen that RW does not succeed in finding consistent estimates. Below we will show a fictitious example of this problem.

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

<table>
<thead>
<tr>
<th>Citizenship</th>
<th>Education</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oceania</td>
<td>Low</td>
<td>1</td>
</tr>
<tr>
<td>Oceania</td>
<td>High</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Industry</th>
<th>Education</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mining</td>
<td>Low</td>
<td>49</td>
</tr>
<tr>
<td>Mining</td>
<td>High</td>
<td>2</td>
</tr>
</tbody>
</table>

We will show that it is impossible to realize consistency with respect to these marginal totals and the register at the same time.

Firstly, we consider the lowly educated people: there is one person from Oceania and 49 “mining” persons with a low educational level. This implies that the combination “Oceania”, “Mining Industry” & “Low educational level” can occur once at most.

Secondly, we consider the highly educated people: nine persons from Oceania and two mining workers have a high educational level. It follows therefore that the combination “Oceania”, “Mining Industry” & “High educational level” can occur twice at most.

By combining both results, 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, which states that there are four “mining” persons from Oceania. Thus, it will be impossible to satisfy all required conditions at the same time. Earlier estimated tables exclude the possibility of estimating new tables.
3. Quadratic programming

This section demonstrates that the consistent estimation problem for which a solution can be obtained with RW, can alternatively be formulated as a quadratic programming (QP) problem that can be efficiently solved by means of Operations Research techniques.

3.1 Repeated weighting as a QP problem

As an alternative for the weighted least squares (WLS) formulation in (2.4), the repeated weighting estimator may also be obtained as a solution of a quadratic programming problem (QP).

The QP formulation is given by

\[ \min_{\hat{t}_y} \sum_{i: (\hat{w}_y)_i > 0} \frac{1}{(\hat{e}_y)_i} \left( (\hat{e}_y)_i - (\hat{w}_y)_i \right)^2 \]

such that:

\[ L\hat{t}_y = r, \]
\[ (\hat{e}_y)_i = 0, \quad \text{for } i \text{ with } (\hat{w}_y)_i = 0. \]

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 (2.4), Operations Research methods do not rely on matrix inversion. Therefore, very efficient solution methods are available (e.g. Nocedal and Wright, 2006). Bikker et al. (2013) apply such methods for National Accounts balancing; an application that requires solving a quadratic optimization problem of approximately 500,000 variables. 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).

Another 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. Nonnegativity requirements and edit rules can be taken into account by including inequality constraints. The empty cell problem can be dealt with by a slight modification of the objective function. The reader is referred to Appendix A.3 for more explanation.

Disadvantages of the QP-approach are that the method does not provide means to derive corrected weights and to estimate variances of reconciled tables.\(^2\)

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. Hence, it can be concluded that there is still a link between the microdata and the reconciled results at macro level.\(^3\)

3.2 Weights

The QP-problem in (3.1) can be more generally written as:

\(^2\) When additional solutions are implemented in the WLS-approach to deal with the empty cell problem or non-negativity of results the two mentioned properties do not necessarily apply for WLS either.

\(^3\) Provided that additional solutions for the additional requirements in Appendix A are not implemented.
\[
\min_{\hat{\mathbf{t}}} \sum_{i=1}^{p} \frac{1}{d_i} (\hat{t}_i - \hat{t}_i^*)^2,
\]

such that:
\[
\mathbf{L}\hat{\mathbf{t}}^* = \mathbf{r},
\]
\[
(\hat{t}_i^*)_i = 0, \quad \text{for } i \text{ with } (\hat{t}_i^w)_i = 0
\]

where \( d_i \) is a weight. Several choices for weights \( d_i \) are possible. Notice, that (3.1) is obtained after setting \( |t_i| \).

A vast amount of literature is available describing data reconciliation problems that can be formulated in the form (3.2). Stone et al. (1942) prove that the most precise results are obtained when weights are set to the reciprocal of the variances of the initial estimates. In practice, estimates of these variances are however often not available. In the absence of any information about data reliabilities, three alternative weight definitions may be used (see e.g. Boonstra, 2004): \( d_i^{(0)} = 1 \), \( d_i^{(1)} = |t_i| \) and \( d_i^{(2)} = (t_i)^2 \).

In literature there is an on-going debate about these alternatives. In the remainder of this report \( d_i^{(1)} \) will be chosen. Several reasons for this, besides equivalence with the RW estimator, are presented in Appendix B. Notice however, that the chosen criterion \( d_i^{(1)} \) does not lead to minimizing percentage adjustment. One can easily derive that choosing criterion \( d_i^{(2)} \) corresponds to this often-used criterion.
4. Simultaneous weighting

This section presents a Simultaneous Weighting (SW) method, of which an outline is already given in Mushkudiani et al. (2014). In SW tables are estimated simultaneously rather than in sequence. Although this is not always feasible in practice, due to problem size, we will explore this method here because it will be used as the basis for the Divide-and-Conquer methods that are presented in Section 5.

4.1 Formal description

The prerequisites for Simultaneous Weighting (SW) are the same as for Repeated Weighting, see Subsection 2.1. Simultaneous weighting starts with initial cell estimates for each table and a set of consistency constraints. The method produces consistent table estimates that are in some sense as close as possible to initial table estimates.

This is done such that all marginal totals of estimated tables are consistent with known population totals from registers and such that all estimated tables are 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.

Analogous to Repeated Weighting, the problem of reconciling multiple tables can be formulated as a generalised least squares problem. This problem is given by

$$\min \mathbf{t}^\top (\hat{\mathbf{t}}^{SW} - \hat{\mathbf{t}}^w) \mathbf{T}^{-1} (\hat{\mathbf{t}}^{SW} - \hat{\mathbf{t}}^w),$$

subject to:

$$L\hat{\mathbf{t}}^{SW} = r,$$

$$\left(\hat{\mathbf{t}}^{SW}\right)_i = 0, \text{ for } i \text{ with } (\hat{\mathbf{t}}^w)_i = 0$$

where $\hat{\mathbf{t}}^{SW} = (\hat{t}^{SW}_1, \ldots, \hat{t}^{SW}_N)^\top$, a vector containing SW estimates for the cells of all $N$ tables, similarly $\hat{\mathbf{t}}^w = (\hat{t}^w_1, \ldots, \hat{t}^w_N)^\top$, a vector of initial estimates and $\mathbf{T}$ is defined by $\text{Diag}(\hat{\mathbf{t}}^w)$.

The objective function minimises a weighted sum of squared differences between initial and reconciled cell estimates for all tables. The linear constraints contain the mathematical formulation of the above-mentioned consistency constraints.

The constraints for consistency between table estimates and registers impose sums of cell from one table to be equal to known population totals. This kind of constraint also occurs in RW, but in RW only one table at a time is considered, whereas in SW all tables are taken into account at the same time. Constraints for mutual consistency of different tables 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.

The problem obtained from (4.1) is a standard GLS problem, for which the solution is given by

$$\hat{\mathbf{t}}^{SW} = \hat{\mathbf{t}}^w + \mathbf{T} L (L^\top L)^{-1} (r - L \hat{\mathbf{t}}^w),$$

Notice that the solution is almost the same as the RW estimator in (2.4). The only difference is that estimates are obtained for all tables at once, rather than for individual tables. Thus, Simultaneous Weighting has similar properties as Repeated Weighting. One of the key properties, that all table estimates can be obtained by weighting underlying microdata, is still valid. Moreover, it is still possible to derive variances of SW estimates.

A Simultaneous Weighting solution can also be obtained as a solution of a QP-problem. The QP-formulation is given by
\[
\min_{\hat{t}^{SW}} \sum_{i,j \geq 0} \frac{1}{\hat{t}^{w}_i} (\hat{t}^{SW}_{i,j} - \hat{t}^{w}_{i,j})^2,
\]

such that:
\[
L \hat{t}^{SW} = r,
\]
\[
(\hat{t}^{SW})_i = 0, \text{ for } i \text{ with } (\hat{t}^{w})_i = 0
\]

where \(M\) denotes the total number of cells in all tables.

Appendix A.4 shows that the QP-approach can easily be generalized to include solutions for certain additional requirements that are often observed in practice. Firstly, inequality constraints can be included to take account of non-negativity requirements of estimated counts and to deal with edit rules. Secondly, the objective function can be modified to deal with the empty cell problem. Thirdly, a two-step procedure can be applied to enforce small adjustment at aggregate level (e.g. low-dimensional marginal totals of tables). This solution means that in a first step high-dimensional auxiliary tables are estimated and in a second step target tables are estimated, consistently with the auxiliary tables that are estimated in the first step.

### 4.2 Discussion

In previous subsection we considered a Simultaneous Weighting approach that searches for an optimal solution for all tables of a given set. It is clear that this approach does not suffer from three drawbacks mentioned in Subsection 2.5 (“Impossibility of consistent estimation”, “Suboptimal solution” and “Order dependency”). Moreover, from a practical point it is more attractive to solve one (or two) problem(s) rather than several problems in RW.

In practise, a SW-approach may however not always be feasible. A large estimation problem needs to be solved consisting of many variables and constraints. The capability of solving such large problems is currently still limited due to computer memory size. This is also the reason why Repeated Weighting was developed in the late 1990s. However, due to increase in computer power, estimation problems that used to be infeasible, do not always have to be a problem anymore.

Because SW applications may still be infeasible even for modern computers, we still 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.

The previous subsection shows that the SW-solution can be obtained by solving either a WLS-problem or a QP-problem. Alternatively, the well-known Iterative Proportionate Fitting (IPF) (Deming and Stephan, 1942) can be applied. We will concentrate on the QP-approach from now on. The reason is that this approach has considerable computational advantages over WLS and is more flexible than IPF, in generalising the model to deal with additional requirements mentioned in Appendix A.
5. Divide-and-conquer algorithms

As an alternative for simultaneous estimating of a set of interrelated tables, so-called Divide-and-Conquer (D&C) algorithms can be used. These algorithms recursively break down a problem into a number of sub problems that can be more easily solved than the original problem. The solution of the original problem is obtained by combining the sub problem solutions.

Subsections 5.1 and 5.2 sketch two divide-and-conquer algorithms for estimating a set of coherent frequency tables. These algorithms construct sub problems that can preferably be independently estimated, whereas RW creates sub problems with many dependencies.

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 as 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 2.5) and “order-dependency of estimation” (Problem 3 in Subsection 2.5). Problem 2 “Suboptimal solution” is however 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 categories). 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, the estimation problem for the second stage can be split into a number of independent problems: one for each aggregated category.

The following example clarifies the idea. Suppose that three tables are estimated.

Table 1: Educational attainment × Age;
Table 2: Educational attainment × Sex;
Table 3: Sex × Occupation.

The required categories for Educational attainment are: 1, 2, 3, 4, 5, 6, 7 and 8. We now obtain a new variable Educational attainment* by aggregating the categories of the original variable into two main categories I and II. Category I comprises 1, 2, 3 and 4, Category II contains the original categories 5, 6, 7 and 8.

In the first stage we estimate the tables:

Table 1*: Educational attainment* × Age;
Table 2*: Educational attainment* × Sex;
Table 3: Sex × Occupation.

The size of this estimation problem is smaller than the size of the original problem because of the aggregation of the categories for Educational attainment.

In the second stage the original tables, i.e. Tables 1 and 2, are estimated. Table 3 does not have to be estimated again because it was already obtained in Step 1. When estimating Tables 1 and 2 it needs to be assured that the results are consistent with the earlier estimated Tables 1* and 2*. Hence it follows that the estimation problem for the second stage depends on the problem for the first step. The joint estimation of Tables 1 and 2 can be separated in two independent problems: one for educational attainment categories 1, 2, 3 and 4; the other one for categories 5, 6, 7 and 8. In each of the two sub problems approximately half of all cells of Tables 1 and 2 are estimated. The size of the estimation problems in the second step is smaller than the size of the original problem, because of the exclusion of tables that do not include the aggregation variable Educational attainment and because of the splitting of the problem into independent sub problems.

In previous example one variable ('occupation') was aggregated. It is however also possible to aggregate multi variables in the first stage. In that case a multi-step method is obtained, in which in each stage after Stage 1, one of the variables is disaggregated.

The proposed procedure identifies multiple stage, in each of which one or multiple estimation problem(s) need(s) to be solved. The estimation problem(s) in one stage depend on the results of a previous stage. Because of these dependencies 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.

Firstly, because the number of estimation problems may be smaller than the number of estimated tables in RW. Secondly, because of a lower degree of dependency among 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.

In practice, the question may arise how to define aggregated categories from the originally defined categories. As aggregation is purely applied to reduce problem size, the aggregated
categories do not have to have any logical meaning. Hence, any choice can be made. However, to reduce problem size as much as possible, it is advisable that each aggregated category contains about the same number of original categories.
6. Practical applications

In this section we present results of practical applications in which the proposed Divide-and-conquer (D&C) methods are applied for the consistent estimation of 42 tables from the 2011 Census. 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. We will compare reconciliation adjustment at cell level and at a more aggregated level of two-dimensional marginal totals. All practical tests were conducted on a 2.8 GHZ computer with 3.00 GB of RAM.

6.1 Setup

A setup based on “splitting by common variables” will be compared with another setup based on “aggregation and disaggregation”.

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) as 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. Each of the 48 table sets contains a subset of the 42 Census tables, determined by the categories of the splitting variables.

It was decided to use Geographic area, Sex and Employment status as splitting variables, because these register variables are included in most tables. 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.

To obtain target tables 48 QP-problems were solved, one for each combination of splitting variable categories. The objective function and constraints for these problems are based on the functionalities described in Appendix A.4. This means that the objective function takes account of the empty cell problem, that inequality constraints are defined to include non-negativity requirements and edit rules in the model and that auxiliary tables are estimated prior to target tables to minimize adjustment at aggregated level. As explained in Appendix C, a more complicated method was necessary to estimate auxiliary tables than for RW and SW.

The largest optimization problem in this procedure consists of 79,315 variables and 137,493 constraints. All optimization problems were successfully solved.

Additionally to the aforementioned, another setup was considered in which Employment Status was used as the only splitting variable. Then, two problems are obtained: one for employed and another one for unemployed people. There is no need to extend the required tables with additional variables because “Employment status” is already included in each table.

---

4 Employment status is not an official Census variable. It can be obtained by aggregating Occupation activity into two categories: employed / unemployed. Officially, the variable does not appear in all required 42 Census tables. Since the two categories employed / unemployed can be derived from different variables - Occupation activity is just one of them - there is a risk of inconsistent results. It may happen that the number of employed people according to one Census table differs from the same number derived from another variable. To prevent this problem, it was decided to include Employment status in all tables.

5 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. Thus, these cells are not included in the number 79,315.
Consequently, using “Employment status” as the only splitting variable will yield an optimal solution, whereas in the previous setup with three splitting variables a suboptimal solution for the original problem will be obtained.

The approach with one splitting variable turned out to be infeasible. One of the resulting optimization problems was too large to store in computer memory. This problem consists of 496,840 variables and 818,485 constraints, but especially problematic is the large number of 17,979,268 nonzero coefficients within the set of constraints.

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 (4 categories for educational attainment and 6 for occupation). Thereafter, results were obtained for the all required categories for the two aggregation variables. This was done according to a five step procedure that is explained in Appendix D.

The largest optimization problem consists of 183,432 variables and 361,830 constraints. All optimization problems were successfully solved. Compared to “Setup 1 - Splitting by common variables” less sub problems are defined, leading to a larger problem size for each sub problem.

6.2 Results
In this subsection we will compare results of the two divide-and-conquer methods with the RW-based official 2011 Census results. The criterion used to compare degree of reconciliation adjustment is based on the QP objective function in (A.9), a sum of weighted squared differences, given by

\[
\sum \frac{1}{(\hat{\tau}_i^m)} (\hat{y}_i - (\hat{y}_i^m))^2
\]

where \(\hat{x}_i^m\) is a vector with initial estimates, \(\hat{y}_i\) is a vector with reconciled estimates, \(\hat{y}_i^m = \max(\{\hat{x}_i^m\}, 1)\) and the summation is over all relevant categories. This criterion is used because all methods under consideration aim at minimizing this or a similar criterion.

Weighted squared difference will be compared at cell level and at a more aggregate level of two-dimensional totals of tables. These marginal totals include at least one variable that need to be estimated from a sample survey. The totals are given by a combination of one sample survey variable (Educational attainment or Occupation) and one register variable (see Appendix E) or a combination of two sample survey variables (Educational attainment and Occupation). A list of these totals is displayed in Table 6.3 below. We do not compare results for combination of register variables because the results are necessarily the same as the value observed in the register, regardless the reconciliation method used.

Table 6.1 compares total adjustment, as defined according to (6.1), based on all cells in all 42 estimated tables.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total adjustment</th>
<th>All cells</th>
<th>Cells with initial estimate larger than zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch 2011 Census</td>
<td>1.1E+08</td>
<td>1.3E+07</td>
<td></td>
</tr>
<tr>
<td>Splitting by common variables</td>
<td>8.8E+07</td>
<td>1.2E+07</td>
<td></td>
</tr>
<tr>
<td>Aggregation and Disaggregation</td>
<td>7.0E+07</td>
<td>1.2E+07</td>
<td></td>
</tr>
</tbody>
</table>
This table shows that, when considering all cells, the RW method as applied to the 2011 Census, actually leads to a suboptimal result, as the amount of reconciliation adjustment obtained with the two D&C methods is less than in the census publications. 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 can be explained from the way how original estimates of zero are processed rather than by the way how the estimation problem is broken down into sub problems. In the Dutch 2011 Census initial cell estimates of zero were replaced by one, to avoid technical estimation problems (see also Appendix A.4). The two alternative methods rely on a different QP-approach that does not require this replacement of zero initial estimates. Hence, these methods result in much smaller adjustment of zero estimates.

The boxplots in Figures 6.1 and 6.2 compare adjustment at the level of individual cells. It can be seen that the two D&C methods lead to more relatively small corrections than the RW method used for officially published Census tables. Differences in results are however smaller again, if zero initial estimates are not taken into account.

![Boxplots of adjustments to all cells of 42 Census tables. The right panel zooms in on the lower part of the left panel.](image1)

![Boxplots of adjustments to cells with nonzero initial estimates. The right panel zooms in on the lower part of the left panel.](image2)

Below we will compare reconciliation adjustment at the more aggregated level of 23 two-dimensional totals that are shown in Table 6.3. These are totals that include one register variable and one sample survey variable, or both sample survey variables. Total adjustment for all two-dimensional totals are displayed in Table 6.2. There is no distinction between cells with zero and nonzero initial estimates in this table, because nonzero initial estimates barely occur at the highly aggregated level of two-dimensional totals.
Table 6.2. Total adjustment by three methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Total adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch 2011 Census</td>
<td>1.28E+06</td>
</tr>
<tr>
<td>Splitting by common variables</td>
<td>1.27E+06</td>
</tr>
<tr>
<td>Aggregation and Disaggregation</td>
<td>1.27E+06</td>
</tr>
</tbody>
</table>

Table 6.2 shows that the two D&C methods lead to slightly lower adjustment than the RW-based method used for Census 2011 compilation; a similar result as for the nonzero initial estimates in Table 6.1.

Results for the underlying two-dimensional totals are shown in the boxplots in Figure 6.3 and in Table 6.3. Notice that in Table 6.3 results of the two-dimensional tables are presented in order of estimation at the 2011 Census. The first row belongs to the total that is estimated first, the last row belongs to the last estimated total.

Table 6.3 confirms the expected problems with order-dependency that were already mentioned in Subsection 2.5. Repeated Weighting reconciliation adjustments are relatively small for tables that are estimated at the beginning of the process and relatively large for tables estimated at the end. For 8 out of 10 first estimated tables 2011 Census adjustments are smaller than those of the two D&C methods. However, when considering the last 13 estimated tables, RW outperforms the other two methods in 4 cases only.

As shown in Table 6.2, total adjustment for the two D&C method is smaller than for RW. Differences in adjustments are however not evenly spread across underlying two-dimensional totals. Despite the larger total adjustment, RW performs better than the two D&C methods for a majority of two-dimensional totals. In most cases, these are totals that do not require large adjustment, regardless the reconciliation method used. There are even few two-dimensional totals with zero RW-adjustments, e.g. EDU x OCC. Initial weights for the microdata were constructed in such a way that no adjustment is necessary. On the contrary, D&C methods perform better than the RW-based method for a relatively small number of marginal totals that require relatively large adjustment, e.g. YAE x EDU and ROY x OCC.

The desirability of the above-mentioned results for two-dimensional totals is questionable. In our practical application the purpose is to minimize total adjustment for a number of two-dimensional totals, without differentiating between different totals. Some two-dimensional totals have a larger influence on total correction of all two-dimensional marginal totals than others. However, if this result is not desirable, a solution is to adjust weights in the optimization problem, such that the adjustment of the different totals will be more equal, or to obtain a result in which low-dimensional totals that are considered the most important are reconciled less.
Table 6.3. Total adjustment of sets of two-dimensional marginal tables (see appendix F for the meaning of the abbreviations for variable names)

<table>
<thead>
<tr>
<th>Total adjustment</th>
<th>2011 Census</th>
<th>Splitting</th>
<th>Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDU x OCC</td>
<td>0</td>
<td>142</td>
<td>112</td>
</tr>
<tr>
<td>GEO x EDU</td>
<td>124</td>
<td>158</td>
<td>152</td>
</tr>
<tr>
<td>GEO x OCC</td>
<td>1,516</td>
<td>1,559</td>
<td>1,559</td>
</tr>
<tr>
<td>SEX x EDU</td>
<td>345</td>
<td>384</td>
<td>382</td>
</tr>
<tr>
<td>SEX x OCC</td>
<td>4,814</td>
<td>4,878</td>
<td>4,885</td>
</tr>
<tr>
<td>AGE x EDU</td>
<td>18,811</td>
<td>18,802</td>
<td>18,705</td>
</tr>
<tr>
<td>AGE x OCC</td>
<td>18,690</td>
<td>18,795</td>
<td>18,770</td>
</tr>
<tr>
<td>CAS x EDU</td>
<td>0</td>
<td>68</td>
<td>65</td>
</tr>
<tr>
<td>CAS x OCC</td>
<td>0</td>
<td>73</td>
<td>72</td>
</tr>
<tr>
<td>IND x EDU</td>
<td>29,008</td>
<td>28,230</td>
<td>28,293</td>
</tr>
<tr>
<td>IND x OCC</td>
<td>31,816</td>
<td>30,336</td>
<td>30,500</td>
</tr>
<tr>
<td>YAE x EDU</td>
<td>271,075</td>
<td>266,027</td>
<td>265,972</td>
</tr>
<tr>
<td>YAE x OCC</td>
<td>274,375</td>
<td>272,047</td>
<td>272,050</td>
</tr>
<tr>
<td>COC x EDU</td>
<td>42,177</td>
<td>42,113</td>
<td>41,969</td>
</tr>
<tr>
<td>COC x OCC</td>
<td>55,351</td>
<td>55,451</td>
<td>55,391</td>
</tr>
<tr>
<td>HFST x EDU</td>
<td>275,858</td>
<td>268,389</td>
<td>268,389</td>
</tr>
<tr>
<td>POB x EDU</td>
<td>1,047</td>
<td>901</td>
<td>886</td>
</tr>
<tr>
<td>POB x OCC</td>
<td>17,192</td>
<td>16,823</td>
<td>16,814</td>
</tr>
<tr>
<td>SIE x EDU</td>
<td>3,274</td>
<td>3,325</td>
<td>3,321</td>
</tr>
<tr>
<td>SIE x OCC</td>
<td>3,414</td>
<td>3,456</td>
<td>3,454</td>
</tr>
<tr>
<td>LPW x EDU</td>
<td>990</td>
<td>1,010</td>
<td>1,007</td>
</tr>
<tr>
<td>LPW x OCC</td>
<td>1,026</td>
<td>1,036</td>
<td>1,010</td>
</tr>
<tr>
<td>ROY x OCC</td>
<td>233,485</td>
<td>232,975</td>
<td>232,980</td>
</tr>
</tbody>
</table>

Underlined means method with lowest adjustment. Marginal totals are presented in order of estimation in the 2011 Census, i.e. EDU x OCC is estimated first and ROY x OCC is estimated last. The marginal totals HFST x OCC and ROY x EDU are not shown. These totals were not estimated because these totals are not contained in any of the 42 estimated Census tables.
7. Discussion

This paper considers the problem of numerical consistent estimation of a set of interrelated frequency tables from different data sources. Several adjustment methods are considered that achieve consistency by minimal adjustment of inconsistent estimates that are directly derived from the data sources. It is argued that minimum adjustment is obtained in a simultaneous approach, in which all tables are reconciled at once.

Because of large problem size, Simultaneous Weighting is however not always feasible. Computer memory size may still limit capabilities of solving such problems. Additional complications are that it may be difficult to interpret the results and to detect causes of model misspecification.

Because of these reasons, a Repeated Weighting method was developed at Statistics Netherlands in the 1990s. In RW the process of obtaining consistent table estimates is broken down in sequential sub problems, in each of which a single table is estimated. Repeated Weighting has however important drawbacks that can be directly explained from the sequential approach. The most important problem is that the method does not always yield a consistent table set, even if such a consistent tables set does exist. When estimating a certain table, it may be impossible to satisfy all consistency demands imposed by all earlier estimated tables.

Problems with the impossibility of obtaining consistent estimates occurred several times when estimating the detailed Dutch 2011 tables. After several weeks of trial and error a number of problem specific solutions were found to solve these estimation problems; an approach which is highly undesirable for any future applications.

In this paper an alternative method is presented that does not suffer from the estimation problem experienced with RW. This method splits the estimation of a set of tables into independently estimated parts, rather than the dependently estimated parts that are distinguished in RW.

A so-called Divide-and-Conquer algorithm (D&C) is presented that partitions a given table set according to the categories of variables that are contained in each table. Common variables may however not always be available in a given table set, but these can be artificially created by adding missing variables to tables, provided that a data source is available from which the resulting extended tables can be estimated. The proposed solution increases the level of detail of tables, which comes at the cost of less precise estimates and potentially more reconciliation adjustment than necessary. Thus, the proposed solution is especially suitable for table sets with many common variables that do not need to be extended much.

As an alternative to the aforementioned method, a second D&C method is proposed based on aggregation of categories for part of the variables. In a first step part of the variable categories are combined into less detailed categories. In the following steps required tables are estimated, according to the prescribed variable categories. In each of those steps one fewer variable is aggregated than in the previous step. Because the optimization problems that are obtained in this way cannot be independently estimated, it cannot be excluded that similar estimation problems occur as in RW. These problems can however be expected to have a lower impact. Firstly, because less sub problems may be distinguished as in RW. Secondly, because the degree of dependency is lower than in RW. 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. The newly proposed D&C methods are much easier to implement, because estimation problems, as experienced with RW can be avoided. Thus, the complicated process of designing ad-hoc solutions for those problems is removed.
Both D&C methods were tested on 2011 Census tables. It turned out that the estimation problems were actually avoided. Furthermore, slightly smaller reconciliation adjustments were observed than in the officially published tables that were based on RW, both at cell level and at a more aggregate level of two-dimensional totals.

At this time, the newly developed methods seem to be feasible for the upcoming 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, at the time of 2021 Census estimation, it is feasible to estimate all tables simultaneously, this is to be preferred. Most importantly, the estimation problems experienced with Repeated Weighting are guaranteed avoided. Secondly, from a practical point of view, it is easy to obtain estimates for all required tables by solving one (or few) optimization problem(s), rather than solving dozens of problems in the current RW approach. Thirdly, an optimal solution is obtained with minimal reconciliation adjustment. From a practical point of view, it is very likely that Simultaneous Weighting is actually possible at Statistics Netherlands when estimating the 2021 Census tables, because the current 32 bit desktops at Statistics Netherlands are planned to be replaced with 64 bit desktops at the end of this year.

If simultaneous table estimation is not possible, 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. Geographic Area is missing in two tables only. After extending two tables with Geographic Area a total of 764,640 cells is obtained; a slightly larger than number than the number of 720,288 cells in the original tables. The 764,640 cells can be allocated to 24 independent estimation problems, one for each combination of categories for Geographic area and Sex. For comparison, in 2011 Census application, a total number of 4,556,544 cells were estimated in 48 independently estimated problems. From this it follows that the optimization problems that are expectedly obtained for the 2021 Census have a smaller size than the problems that were solved for the 2011 Census. Consequently, the proposed method can be expected to be feasible for estimating the 2021 census tables.

This paper proposes to use a different estimation technique than in the original RW papers. Here, it is proposed to obtain a solution by means of solving QP-problems, whereas the original RW papers make use of closed-form expressions for the solution of the reconciliation problem. The main advantage of the QP-approach is better computational performance. Another advantage is that it can be more easily generalized to take additional requirements into account, like non-negativity of estimated counts, (edit rule) relations between different variables and solutions for the empty cell problem. A disadvantage of the QP-approach is however that it does not provide means to estimate variances of reconciled estimates. In the original RW papers closed-form expressions are derived for these variances. Since these expressions rely on matrix inversion, using these expressions in practice is however only feasible for tables that are not too large. In the 2011 census, estimated variances were obtained for certain low-dimensional “key” tables only. Implementation of the newly proposed D&C-methods could mean that new solutions need to be developed for estimating precision of results. At the current moment it is not clear whether variances are required for the 2021 census.

A final remark on the software that was used to implement the methods in this paper. For any large scale application of QP problems it is advisable to use a commercial solver for
mathematical optimization problems, like CPLEX, XPRESS or GUROBI. Besides good computational performance, these software also provide extensive additional facilities, e.g. to detect errors in model specification. We used XPRESS in this paper; because this software is already available at Statistics Netherlands. It is currently used for National Accounts reconciliation, a problem that gives rise to similar optimization problems as the Dutch virtual census. The currently developed prototype software may be used again to test the feasibility of the proposed methods on realistic test data for the upcoming 2021 Census.

In the beginning of the project the open source R-package “RSPA” (van der Loo, 2012) was considered, a package that is especially developed for large adjustment problems. This package was however not chosen, because it cannot be used for QP problems with linear combination of variables in the objective function; a problem that is described in (A.9).
8. References


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.


van der Loo, M. (2012), *rspa: Adapt numerical records to (in)equality restrictions with the Successive Projection Algorithm*. R package version 0.1-5. Available at: [http://cran.r-project.org/web/packages/rspa/index.html](http://cran.r-project.org/web/packages/rspa/index.html).

Appendix A. Estimation problems

Subsection 2.5 presents three RW problems that are directly related to the sequential estimation. In addition to this Appendix A.1 below presents four other complications of RW that are not directly caused by the sequential approach. Known solutions for these problems are summarized in Appendix A.2. In Appendices A.3 and A.4 it is explained that the problems can be easily solved in a Quadratic Programming (QP) context.

A.1 Description

Below we summarise complications that are experienced when applying RW in practice that are not directly caused by the sequential approach of RW.

Problem a. Edit rules
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. The reason underlying this rule is that a person who has never lived abroad was by definition born in the country concerned. This rule not only applies to the overall population, but also to all possible sub-populations (e.g. 23 year-old married men).

Problem b. Negative cell estimates
Repeated weighting may yield negative values as cell estimates. In many practical applications, such as the Dutch Census, negative values are however not allowed.

Problem c. Empty cell problem
The empty cell problem occurs when estimates have to be made without underlying data. It is caused by sampling effects: a characteristic known to exist in the population is not covered by the sample survey from which a table is estimated. A fictitious example of this is the estimation of a table comprising population by Geographic area, Branch of economic activity and Educational attainm

Problem d. Large correction of low-dimensional marginal totals
Although RW aims at minimizing adjustment of the cells of tables, there may still be relatively large corrections of low-dimensional marginal totals. This can happen, for example, if all corrections of the underlying cells are in the same direction (upward or downward). An illustrative case is when components of a marginal total all have a zero initial estimate and non-negativity constraints apply. In that case, all components may be slightly adjusted in upward direction, but adjustment of the marginal total, i.e. the sum of all adjustments, can be unacceptably large.

A.2 Known solutions for Repeated Weighting
In this subsection solutions are presented for the practical problems mentioned in Appendix A.1. In particular, we will mention the solutions that were applied for the 2011 Dutch Census.
Additionally, we will review some solutions that are known from literature.

Problem a (“Edit rules”) was solved in the 2011 Census by extending all tables containing a variable involved with a certain edit rule with all other variables appearing in that edit rule. In the previous mentioned example that someone who never resided abroad was necessarily born in the country of residence can be solved by adding ‘country of birth’ to all tables that include ‘ever/never resided abroad’. This will prevent violation of edit rules because all tables are estimated from one data set and all edit rules are assumed to be satisfied within each data set (see Subsection 2.1). A drawback however is that the already detailed Census tables are further extended. For the 2011 Census, the problem size was increased from 1,047,584 to 1,633,504 cells.

Problem b (“Negative cell estimates”) was solved in the 2001 Census by using iterative proportional fitting (IPF) as estimation technique, rather than weighted least squares (WLS). The IPF algorithm is a recursive, minimal adjustment method that proportionally fits sample observations to known marginal totals. Contrary to WLS, IPF necessarily leads to nonnegative estimates. The algorithm is generally attributed to Deming and Stephan (1940), who applied the method to the 1940 American census, but the method goes by many names, depending on the field and the context (Pritchard, 2008). It has been pointed out by Stephan (1942) that in most cases IPF very closely approximates the WLS formulation in (2.4). IPF also has a number of drawbacks: it is not clear how variances of estimates can be computed and similar to RW, zero values will never be adjusted to nonzero, even when this is desirable. A second approach to prevent negative cells estimates, besides using IPF, is to bound corrected weights that are obtained by repeated weighting. Several methods are described in the literature for the GREG-estimator, see e.g. Deville and Särndal (1992) and Huang and Fuller (1978). A comparative study of these and other methods is presented in Park and Fuller (2005).

Problem c (“Empty cell problem”) was tackled in the 2011 Dutch Census 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 was 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. The underlying microdata were not adjusted. Although the epsilon method solves the empty cell problem, the method has also drawbacks. Firstly, results may be affected by the artificial replacement of zero values. In particular, non-zero counts may be obtained for categories that are not likely to exist in the population, so-called structural zeroes. (e.g. 15 year-old professors). This problem was solved by estimating low-dimensional ‘auxiliary tables’ before all target tables, see the explanation at the end of this section.

Secondly, the connection between microdata and final table estimates gets lost. When applying the epsilon method, it is no longer possible to derive corrected weights such that final table estimates can be obtained by weighting the underlying microdata.

Thirdly, it is no longer possible to derive variances of estimates.

Several other solutions for the empty cell problem are given literature (see e.g. Beckman et al. (1996) and Guo and Bhat (2007)) but these are not appropriate for the Dutch Census, because these either lead to inconsistent results, or to a loss of details in the results.

Problem d (“Large correction of low-dimensional marginal totals”) was tackled in the Dutch 2011 Census by estimating so-called auxiliary tables before all other tables. Auxiliary tables are not very detailed, typically a combination of one register and one sample survey variable (e.g. education by sex, education by age). These tables were chosen, such that when estimating them, the empty cell problem does not occur. After auxiliary tables are estimated, all target
tables follow. As all target tables have to be estimated consistently with the auxiliary tables, there is no deviation from the data sources at the low-dimensional level of the auxiliary tables. Estimation of auxiliary tables may also prevent illegal nonzero counts for cells that cannot have any different result than a zero count. Because auxiliary tables can be estimated without the epsilon method, structural zeroes will be preserved in these tables. Consequently, the problem of illegal nonzero results, i.e. a nonzero estimate for characteristics that cannot occur, will not occur for all cells covered by an auxiliary table.

A.3 QP-Solutions for estimating a single table
In this subsection we will show that the QP formulation in Subsection 3.1 can easily be extended to solve the additional requirements mentioned in Appendix A.1. Problem a (“Edit rules”) and Problem b (“Nonnegative cell estimates”) can be dealt with by adding inequality constraints. Problem c (“empty cell problem”) can be solved by modification of the minimization function. Problem d can be dealt with by adding low-dimensional auxiliary tables before all target tables.

Non-negativity
Non-negativity of reconciled cell estimates can be imposed by adding the following inequality constraints.

\[ \hat{t}_{y} \geq 0 \quad y \in (1, \ldots, N). \]  \hspace{1cm} (A.1)

Edit rules
Edit rules require that marginal totals of a certain table to be bounded (from above or below) by known marginal totals of an earlier estimated table. This constraint can be written in the following form:

\[ C_{yu} \hat{t}_{y} \leq u, \]  \hspace{1cm} (A.2)

\[ C_{yl} \hat{t}_{y} \geq l. \]  \hspace{1cm} (A.3)

In this formulation \( u \) and \( l \) are upper and lower bounds that are used in the different edits, further \( C_{yu} \) and \( C_{yl} \) are aggregation matrices for deriving marginal totals of Table \( y \) for which known upper respectively lower bounds are available. The constraints in (A.2)-(A.3) are linear inequalities that can easily be dealt with in QP problems.

Empty cell problem
For the empty cell problem we will consider a solution based on Boonstra (2004). The solution consists of replacing the minimisation function (3.2) with

\[ \min_{\hat{t}_y} \sum_{i=1}^{n_y} \left( \frac{1}{|\hat{t}_y|} \right) \left( \hat{t}_y - \hat{t}_y \right)^2 \]

such that: \( L \hat{t}_y = r. \)  \hspace{1cm} (A.4)

where \( \hat{t}_y^{\text{max}} = \max \left( |\hat{t}_y|, 1 \right). \) The denominator of each term of the objective function is replaced by 1, if the original value would be zero. As opposed to RW, zero valued initial estimates do not necessarily remain zero.

This solution has similarities with the epsilon method, but it is not the same. The solution that is proposed here is less radical, in the sense that it only changes the denominator of each term of the objective function, while the epsilon method also changes \( \hat{t}_y \) in the nominator. In the context of Subsection 3.2, the currently proposed solution basically means that the weight \( d_i^{(0)} \) is used instead of \( d_i^{(1)} \), but only when necessary to avoid division by zero problems.
Large correction at aggregate level

For this problem a similar solution can be applied as the one used for the 2011 Dutch Census: estimating low-dimensional auxiliary tables before all other tables.

A.4 QP-Solutions for simultaneous estimation

In this subsection we will show that the QP formulation for SW in (4.1) can easily be generalised to include solutions for the practical problems described in Subsection A.1.

Non-negativity

Analogous to Subsection A.3, non-negativity constraints and edit rules can be modelled as inequality constraints. Non-negativity of cell estimates can be enforced by imposing:

\[ \hat{t}^{SW} \geq 0. \]  \hspace{1cm} (A.5)

Edit rules

In a SW-context edit rules require that marginal totals of a certain table to be bounded (from above or below) by marginal totals of the same or another table. The following inequality constraints can be defined to take edit rules into account

\[ C_{e_1} \hat{t}^{SW} \leq C_{e_2} \hat{t}^{SW}, \]  \hspace{1cm} (A.6)

where \( C_{e_1} \) and \( C_{e_2} \) are aggregation matrices, for deriving marginal totals involved with a set of edit rules.

Empty cell problem

For the “empty cell” problem we propose a similar solution as the one proposed in previous subsection. This means that objective function (4.3) is replaced with

\[ \min_{\hat{t}^{SW}} \sum_{i=1}^{M} \frac{1}{\hat{t}_i^w} (\hat{t}_i^{SW} - \hat{t}_i^w)^2, \]

such that: \( \hat{L} \hat{t}^{SW} = r \),

where \( \hat{t}_i^w = \max(\{\hat{t}_i^w, 1\}) \).

Large correction at aggregate level

For the problem “Large correction of low-dimensional marginal totals” a two-step solution will be proposed. The first step aims at minimizing as much as possible adjustments at an aggregate level. The second step derives cell estimates that are consistent with aggregate totals that are derived in Step 1.

First, predetermined low-dimensional aggregate totals are estimated. In the first step the following model is solved:

\[ \min_{\hat{t}^{SW}} \sum_{i=1}^{M} \frac{1}{\hat{t}_i^w} (\hat{C}_a \hat{t}^{SW}_i - \hat{C}_a \hat{t}_i^w)^2, \]

such that: \( \hat{L} \hat{t}^{SW} = r \),

\[ (\hat{C}_a \hat{t}_i^w)_i = 0 \quad \text{for all } i \text{ with } (\hat{C}_a \hat{t}_i^w)_i = 0, \]

where \( \hat{C}_a \) is an aggregation matrix for deriving the low-dimensional aggregates from target tables cells.

The objective function minimizes adjustment at the level of the low-dimensional aggregate totals. The constraints in (A.8) are the same as before. These constraints are defined to achieve consistency among different tables and between tables and known population totals from
registers. Moreover, inequality constraints for nonnegative demands and edit rules may also be included.

In a second step all target tables are estimated. The following model is solved for that purpose

\[
\min \sum_{i=1}^{M} \frac{1}{\hat{t}_i^{sw}} (\hat{t}_i^{sw} - \hat{t}_i^{w})^2,
\]

such that: \( L \hat{t}^{sw} = r^* \).

This model is basically the same as (A.7), but the set of constraints \( L \hat{t}^{sw} = r \) does not only include the original constraints \( L \hat{t}^{sw} = r \), but also constraints that enforce consistency with the marginal totals that are estimated in the first step. The latter constraints are defined by

\[
C_i \hat{t}_i^{w} = \hat{t}_i^{opt}
\]

where \( \hat{t}_i^{opt} \) denotes the aggregate totals that are obtained as a solution of the problem (A.8). The solution that is proposed here is a two-step procedure. We have however seen before that in another multi-step procedure – Repeated Weighting – that the set of constraints that is obtained after a certain step may become infeasible (Problem 1 of Section 2.5). This problem may in principle also occur in a two-step procedure. In the second step it may become impossible to comply with all consistency constraints that are imposed by the aggregate totals obtained in the first step. The above mentioned problem is however avoided in the proposed procedure.

This is done in the estimation of aggregate totals in Step 1, where constraints at the underlying cell level are taken into account. As a result, not only aggregate totals are obtained in Step 1, but also a set of underlying cell values that satisfy all consistency constraints. Hence, at the beginning of Step 2, it is already guaranteed that at least one set of consistent cell estimates exists that fits to the aggregated totals that are derived in Step 1.
Appendix B. Weights in general QP-problems

In Subsection 3.2 three alternative weight definitions for the general QP-problem

\[
\min \sum_{i=1}^{P} \frac{1}{d_i} (\hat{t}_{i*} - \hat{t}_i)^2,
\]

such that: \( \hat{t} \hat{t}^* = r \).

are considered: \( d_i^{(0)} = 1, d_i^{(1)} = |t_i| \) and \( d_i^{(2)} = (t_i)^2 \). Below we will discuss properties of these three alternatives and it will be explained why \( d_i^{(1)} \) is preferred in this paper.

The choices \( d_i^{(0)} \) and \( d_i^{(2)} \) lead to minimal sums of squared absolute and relative adjustments, respectively. The alternative \( d_i^{(0)} \) 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.

Alternative \( d_i^{(1)} \) leads to results that closely approximates IPF (see Stephan, 1942); adjustments are proportionate to the magnitude of the cells, which is intuitively easily understood.

Alternative \( d_i^{(2)} \) is preferred in Di Fonzo and Marini (2009) because it implies that different cells have the same coefficients of variation (CV), meaning that standard errors are proportionate to the magnitude of the cells. Moreover, \( d_i^{(2)} \) minimizes percentage differences: a criterion that is often used and well-understood.

Statistics Netherlands used to apply weights \( d_i^{(2)} \) for reconciling National Accounts (see Bikker et al., 2013), but recently 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. The relative share of total adjustment that is attributed to small values is larger in \( d_i^{(1)} \) than in \( d_i^{(2)} \). 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 Section 6 a Divide-and-Conquer method is presented (the so-called “Aggregation Disaggregation method”) that is 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, which may be undesirable for a practical application.
Appendix C. Auxiliary tables in a splitting approach

Subsection 5.1 describes a Divide-and-conquer (D&C) algorithm based on partition of tables according to common variables (the so-called “Splitting by common variables algorithm”). It will be explained that when applying the splitting algorithm the originally proposed method for estimating auxiliary tables will not always lead to the desired result of low adjustment at aggregate level. First we will explain this problem, thereafter we will give a solution.

The main idea of “Splitting” method is that a table set is partitioned on the basis of categories for common variables, i.e. variable that appear in each table. When using this approach, auxiliary tables need to be subdivided as well, because these auxiliary tables are imposed as required marginal totals for the partitioned target tables. Consequently, auxiliary tables necessarily need to contain all splitting variables. When several splitting variables are used, this implies that auxiliary tables have a certain level of detail. Using detailed auxiliary tables may however not avoid large adjustment at a more aggregated level. To solve this problem the following three-step procedure will be proposed for the estimation of auxiliary tables:

Step 1 Estimation of lowly detailed auxiliary tables, i.e. tables that do not necessarily contain all splitting variables. These tables are at the level of aggregation for which low adjustment from the data source is strived for.

Step 2 Estimation of more detailed auxiliary tables that contain all splitting variables, such that these tables are consistent with the less detailed tables that are estimated in Step 1.

Step 3 Minimal adjustment of the estimated tables in Step 2 to ensure that target tables can be consistently estimated with the resulting auxiliary tables.

In each of the three steps one or multiple QP-problem(s) is (are) solved.

In Step 1, the variables of the optimization problem are the cells of the lowly detailed auxiliary tables. The objective function minimizes weighted squared differences between initial estimates and newly obtained totals. The minimization is under the constraints that the lowly detailed auxiliary tables are: mutually consistent (i.e. that common marginal totals are identically estimated), consistently estimated with known register totals, that all estimated values are nonnegative and edit rules are satisfied (see also Appendix A.4).

For the optimization model in Step 2, the variables of the optimization model are the cells of the detailed auxiliary tables. The objective function minimizes weighted squared differences between initial estimates and newly obtained detailed auxiliary tables. The constraints are the same as the constraints used for Step 1, but additional constraints are defined to ensure that the marginal totals of the detailed auxiliary tables comply with the less detailed tables that are available from Step 1.

In Step 3 multiple QP-optimization problems are solved; one for each combined category of the splitting variables. Each optimization problem contains variables for cells of the detailed auxiliary tables and for the underlying target table cells. The objective function minimizes weighted squared differences between estimates for the detailed auxiliary tables that are obtained in Step 2 and the newly obtained detailed auxiliary tables.

The constraints do not only take account of the auxiliary tables, but these also take account of underlying target table cells. The constraints impose that all target tables cells: are mutually consistent, comply with known register totals, are nonnegative, fulfill edit rules and comply with the newly estimated detailed auxiliary tables. See Appendix A.4 for a more technical description of these constraints.
In the application mentioned in Subsection 6.1 three splitting variables are defined: Geographic area, Sex and Employment status. The application aims at minimizing reconciliation adjustment at the level of two-dimensional totals, i.e. all combinations of one register variable and one sample survey variable, and the combination of two sample survey variables. In the first step auxiliary tables are estimated for the above-mentioned two-dimensional totals. In the second step these auxiliary tables are extended to include the three splitting variables. As a result five-dimensional auxiliary tables are obtained with:
- three splitting variables (Geographic area, Sex and Employment status);
- one sample survey variable (Occupation or Education attainment)
- one register variable (other than the three splitting variables, see Appendix E).
One additional auxiliary table is defined, with:
- three splitting variables (Geographic area, Sex and Employment status);
- two sample survey variables (Occupation and Education attainment)
It turned out that using five-dimensional auxiliary tables only, i.e. applying Step 3 without Steps 1 and 2, lead to relatively large (undesirable) adjustment at the level of 2-dimensional totals. That was the reason why the extended method was developed.
Appendix D. Aggregation and disaggregation

This appendix explains the setup for a practical application of the “aggregation and disaggregation” method that was explained in Section 6.

The two variables “educational attainment” (8 categories, denoted by 1,...,8) and “occupation” (12 categories, denoted by 1,...,12) were aggregated into two main categories, that each contain half of the categories of the original variables (4 categories for educational attainment and 6 for occupation). These main categories will be called EDU-I, EDU-II, OCC-I and OCC-II.

The first main category EDU-I comprises the “educational attainment” categories 1,...,4, the second category EDU-II contains the other categories 5,...,8, OCC-I contains the “occupation” category 1,...,6 and OCC-II the other categories 7,...,12.

A five step procedure was applied, consisting of the following steps:

Stage 1: Aggregating Educational attainment and Occupation

In this stage all 48 target tables are simultaneously estimated, using aggregated categories for Educational attainment and Occupation.

Stage 2: Disaggregating Educational attainment

In Stages 2a and 2b two SW-problems are solved to obtain estimates for the original categories for Educational attainment. In each problem only those tables are estimated that include Educational attainment (26 tables for the 2011 Census). Categories for Occupation are aggregated and the original required categories for Educational attainment are used. The problems for underlying categories of EDU-I and EDU-II can be independently solved. This is done in Stages 2a and 2b.

- Stage 2a: EDU-I
  A SW estimation problem is solved that only includes underlying categories for EDU-I. (underlying categories for EDU-II are not included in the optimization model)

- Stage 2b: EDU-II
  A SW estimation problem is solved that only includes underlying categories for EDU-II.

Stage 3: Disaggregating Occupation

In Stages 3a and 3b two SW-problems are solved to obtain estimates for the original categories for Occupation. In those SW problems only those tables are estimated that include Occupation (26 tables for the 2011 Dutch Census). The original required categories for Educational attainment and Occupation are used in those estimation problems. For the estimation of the underlying subcategories for OCC-I and OCC-II two independent problems are solved in Stages 3a and 3b.

- Stage 3a: OCC-I
  A SW problem is solved that only includes categories for OCC-I.

- Stage 3b: OCC-II
  A SW problem is solved that only includes categories for OCC-II.

The above described steps give rise to five SW problems.

In the practical application described in Subsection 6.1 a QP-approach is followed as described in Appendix A.4.

In addition to the constraints mentioned in Appendix A.4 additional constraints are defined to ensure that tables that are estimated at a certain stage comply with more aggregated tables that were estimated in a previous stage. For tables that contain Educational Attainment and Occupation, results of Stages 3a and 3b need to be consistent with the result of Stages 2a and 2b, respectively. For tables with Occupation, but without Educational attainment, estimation results of Stages 3a and 3b need to comply with the results obtained in Stage 1. The above-mentioned consistency demands can be translated into linear equality constraints, which
impose a sum of cells from a certain table to be equal to a known total from an earlier estimated table.
Appendix E. List of variables

The following variables appear in the demographic part of the 2011 Dutch Census.

- Educational attainment (EDU)
- Occupation (OCC)
- Age (AGE)
- Current activity status (CAS)
- Country of citizenship (COC)
- Place of Usual residence / Geographical area (GEO)
- Household status (HST)
- Industry / branch of economic activity (IND)
- Locality / Size of locality (LOC)
- Location place of work (LPW)
- Country place of birth (POB)
- Place of usual residence one year prior to the census (ROY)
- Sex (SEX)
- Status in employment (SIE)
- Year of arrival in the country (YAE)

The first two variables are observed from a sample survey. The other variables are observed from registers. Therefore, these variables are called sample survey variables and register variables throughout this paper.