Macro-integration techniques with applications to census tables and labour market statistics

Nino Mushkudiani, Jacco Daalmans and Jeroen Pannekoek

Discussion paper (201201)

Statistics Netherlands

The Hague/Heerlen, 2012

Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/'12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/'10–2011/'12	crop year, financial year, etc. 2009/'10 to 2011/'12 inclusive

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

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201201 X-10

Macro-integration techniques with applications to census tables and labour market statistics

Nino Mushkudiani, Jacco Daalmans and Jeroen Pannekoek

Summary: Macro-integration is widely used for the reconciliation of macro figures, usually in the form of large multi-dimensional tabulations, obtained from different sources. Traditionally these techniques have been extensively applied in the area of macro-economics, especially in the compilation of the National Accounts. Methods for macro-integration have developed over the years and have become very versatile techniques for solving integration of data from different sources at a macro level. In this paper we propose applications of macro-integration techniques in other domains than the traditional macro-economic applications. In particular, we present two possible applications for macro-integration methods: reconciliation of tables of a virtual census and combining estimates of labour market variables.

Keywords: Macro-integration, Data reconciliation, Census, labour market data

1 Introduction

Macro-integration is widely used for the reconciliation of macro figures, usually in the form of large multi-dimensional tabulations, obtained from different sources. Traditionally these techniques have been extensively applied in the area of macro-economics, especially in the compilation of the National Accounts, for example to adjust input-output tables to new margins (see, e.g. Stone et al. 1942). Combining different data at the macro level, while taking all possible relations between variables into account, is the main objective of reconciliation or macro-integration. Combining different data sources also makes possible to detect and correct flaws in data and to improve the accuracy of estimates. The methods for macro-integration have developed over the years and have become very versatile techniques for solving integration of data from different sources at a macro level. In this paper we propose several new applications of macro-integration techniques in other domains than the traditional macro-economic applications.

Currently, at Statistics Netherlands, methods for macro-integration are applied to match quarterly and yearly values of the National Accounts. The multivariate Denton method (see Bikker and Buijtenhek 2006) was extended for this application with ratio restrictions, soft restrictions, inequalities and variances. Besides a large number of variables, this model can also handle a very large number of restrictions (see Bikker et al. 2010). Another application of macro-integration at SN uses a Bayesian approach to deal with inclusion of inequality constraints, for integration of international trade statistics and transport statistics (see Boonstra et al. 2010).

In this paper we investigate the application of macro-integration techniques in the following areas:

- Reconciliation of tables for the Census 2011;
- Early estimates for the labour market variables.

The paper is organized as follows: in Section 2 we will give a short outline of macro-integration methods used in this paper, including the extended Denton method. In Section 3, we describe the Census 2011 data problem for SN and the use of macro-integration for it. In Section 4, we will do the same for the early estimates of labour market variables. The conclusions can be found in Section 5

2 Methods

2.1 The macro-integration approach

We consider a set of estimates in tabular form. These can be quantitative tables such as average income by region, age and gender or contingency tables arising from the cross-classification of categorical variables only, such as age, gender, occupation and employment. If some of these tables have certain margins in common and if these tables are estimated using different sources, these margins will often be inconsistent. If consistency is required, a macro-integration approach can be applied to ensure this consistency.

The macro-integration approach to such reconciliation problems is to view them as constrained optimization problems. The totals from the different sources that need to be reconciled because of inconsistencies are collected in a vector $\mathbf{x}$ $(x_i : i = 1, \ldots, N)$. Then a vector $\widehat{\mathbf{x}}$, say, is calculated that is close to $\mathbf{x}$, in some sense, *and* satisfies the constraints that ensure consistency between the totals. For linear constraints, the constraint equations can be formulated as

$$\mathbf{C}\widehat{\mathbf{x}} = \mathbf{b}, \tag{1}$$

where $\mathbf{C}$ is a $c \times N$ matrix, with c the number of constraints and $\mathbf{b}$ a c-vector. These linear constraints include equality constraints that set the corresponding margins of tables estimated from different sources equal to each other as well as benchmarking constraints that set the estimates of certain margins from all sources equal to some fixed numbers. The equality constraints are likely to apply to common margins that can be estimated from different sample surveys but cannot be obtained from a population register, while the benchmarking constraints are likely to apply when the common margins can be obtained from register data in which case the fixed numbers are the values for this margin obtained from the register.

Consider a class of penalty functions represented by $(\mathbf{x}-\widehat{\mathbf{x}})'\mathbf{A}(\mathbf{x}-\widehat{\mathbf{x}})$, a quadratic form of differences between the original and the adjusted vectors, here $\mathbf{A}$ is a symmetric, $N \times N$ nonsingular matrix. The optimization problem can now be formulated as:

$$\min_{\widehat{\mathbf{x}}} \ (\mathbf{x} - \widehat{\mathbf{x}})'\mathbf{A}(\mathbf{x} - \widehat{\mathbf{x}}), \quad \text{with} \ \ \mathbf{C}\widehat{\mathbf{x}} = \mathbf{b}.$$

In the case that $\mathbf{A}$ is the identity matrix, we will be minimizing the sum of squares of the differences between the original and new values:

$$(\mathbf{x} - \widehat{\mathbf{x}})'(\mathbf{x} - \widehat{\mathbf{x}}) = \sum_{i=1}^{N} (x_i - \widehat{x}_i)^2.$$

To solve this optimization problem, the Lagrange method can readily be applied. The Lagrangian is

$$L = (\mathbf{x} - \widehat{\mathbf{x}})'\mathbf{A}(\mathbf{x} - \widehat{\mathbf{x}}) - \boldsymbol{\lambda}'(\mathbf{C}\widehat{\mathbf{x}} - \mathbf{b}) \tag{2}$$

with $\boldsymbol{\lambda}$ a vector with Lagrange multipliers. For an optimum, we must have that the gradient of $L(\boldsymbol{\lambda}, \widehat{\mathbf{x}})$ with respect to $\widehat{\mathbf{x}}$ is zero. This gradient is:

$$\frac{\partial L}{\partial \widehat{\mathbf{x}}} = -2(\mathbf{x} - \widehat{\mathbf{x}})'\mathbf{A} - \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0}$$

and hence,

$$2(\mathbf{x} - \widehat{\mathbf{x}}) = -\mathbf{A}^{-1}\mathbf{C}'\boldsymbol{\lambda}. \tag{3}$$

By multiplying both sides of this equation with $\mathbf{C}$ and using equation (1) we obtain for $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = -2(\mathbf{C}\mathbf{A}^{-1}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{x} - \mathbf{b}),$$

where $\mathbf{C}\mathbf{A}^{-1}\mathbf{C}'$ is a square matrix that is nonsingular as long as there are no redundant constraints. Substituting this result in (3) leads to the following expression for $\widehat{\mathbf{x}}$:

$$\widehat{\mathbf{x}} = \mathbf{x} - \mathbf{A}^{-1}\mathbf{C}'(\mathbf{C}\mathbf{A}^{-1}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{x} - \mathbf{b}). \tag{4}$$

2.2 Comparison with the GREG-estimator

In survey methodology it is common to make use of known marginal totals of variables that are also measured in the survey by the use of calibration or generalized regression (GREG) estimation, (see Särndal et al. 1992). Following Boonstra 2004, we will compare in this subsection the GREG-estimator with the adjusted estimator given by (4) for the estimation of contingency tables with known margins.

The situation in which calibration or GREG-estimation procedures can be applied is as follows. There is a target variable y, measured on a sample of n units, for which the population total, x_y say, is to be estimated. Furthermore, there are measurements on a vector of q auxiliary variables on these same units for which the population totals are known. For the application of the GREG-estimator for the total of y, first the regression coefficients for the regression of y on the auxiliary variables are calculated. Let the measurements on y be collected in the n-vector $\mathbf{y}$ with elements $y_i, (i = 1, \ldots, n)$, and the measurements on the auxiliary variables in vectors $\mathbf{z}_i$ and let $\mathbf{Z}$ be the $n \times q$ matrix with the vectors $\mathbf{z}_i$ as rows. The design-based estimator of the regression coefficient vector $\boldsymbol{\beta}$ can then be obtained as the weighted least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\boldsymbol{\Pi}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Pi}^{-1}\mathbf{y}, \tag{5}$$

with $\mathbf{\Pi}$ a diagonal matrix with the sample inclusion probabilities π_i along the diagonal.

Using these regression coefficients the regression estimator for the population total of y is estimated by

$$\hat{x}_{y.greg} = \hat{x}_{y.ht} + (\mathbf{x}_{z.pop} - \hat{\mathbf{x}}_{z.ht})' \hat{\boldsymbol{\beta}}, \tag{6}$$

with $\hat{x}_{y.ht}$ and $\hat{\mathbf{x}}_{z.ht}$ the 'direct' Horwitz-Thompson estimators, $\sum_i y_i/\pi_i$ and $\sum \mathbf{z}_i/\pi_i$, for the population totals of y and $\mathbf{z}$, respectively and $\mathbf{x}_{z.pop}$ the known population totals of the auxiliary variables. The regression estimator $\hat{x}_{y.greg}$ can be interpreted as a 'weighting' estimator of the form $\sum_i w_i y_i$ with the weights w_i given by

$$w_i = \frac{1}{\pi_i}\left[1 + (\mathbf{x}_{z.pop} - \hat{\mathbf{x}}_{z.ht})' \left(\mathbf{Z}'\mathbf{\Pi}^{-1}\mathbf{Z}\right)^{-1} \mathbf{z}_i\right]. \tag{7}$$

From (7) two important properties of the GREG-estimator are directly apparent. Firstly, the weights depend only on the auxiliary variables and not on the target variable. This means that the GREG-estimators for different target variables can be obtained by the same weights as long as the auxiliary variables remain the same. Secondly, the GREG-estimates of the totals of the auxiliary variables, $\hat{x}_{z.greg} = \sum_i w_i \mathbf{z}_i$, are equal to their known population totals.

For multiple target variables, $\mathbf{y}_i = (y_{i1} \ldots y_{ip})$ the GREG-estimators can be collected in a $p-$vector $\hat{\mathbf{x}}_{y.greg}$ and (6) generalizes to

$$\hat{\mathbf{x}}_{y.greg} = \hat{\mathbf{x}}_{y.ht} + \mathbf{B}\left(\mathbf{x}_{z.pop} - \hat{\mathbf{x}}_{z.ht}\right), \tag{8}$$

with $\hat{\mathbf{x}}_{y.ht}$ the $p-$vector with Horvitz-Thompson estimators for the target variables and $\mathbf{B}$ the $p \times q$-matrix with the regression coefficients for each target variable on the rows. Generalizing (5), we have for the coefficient matrix $\mathbf{B} = \mathbf{Y}'\mathbf{\Pi}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{\Pi}^{-1}\mathbf{Z}\right)^{-1}$, where $\mathbf{Y}$ is the $n \times p$-matrix with the vectors of target variables, $\mathbf{y}_i$, on the rows.

Now, consider the case when the totals to be estimated are the cell-totals of a contingency table obtained by the cross-classification of a number of categorical variables. For instance, the target totals could be the numbers of individuals in the categories 1.Unemployed and 2.Employed of the variable Employment by age category and sex in some (sub)population. If we assume, for ease of exposition, that Age has only two categories, 1.Young and 2.Old and Sex has the categories 1.Male and 2.Female, then there are eight totals to be estimated, one for each cell of a $2 \times 2 \times 2$ contingency table. Corresponding to each of these eight cells we can define, for each individual, a zero-one target variable indicating whether the individual belongs to this cell or not. For instance $y_1 = 1$ if Employment = 1, Age = 1 and Sex = 1, and zero in all other cases and $y_2 = 1$

if Employment = 2, Age = 1 and Sex = 1, and zero in all other cases, etc. Each individual scores a 1 in one and only one of the eight target variables.

For such tables, some of the marginal totals are often known for the population and GREG-estimators that take this information into account are commonly applied. In the example above, the population totals of the combinations of Sex and Age could be known for the population and the auxiliary variables then correspond to each of the combinations of Sex and Age. The values for the individuals on these auxiliary variables are sums of values of the target variables. For instance, the auxiliary variable for Age = 1 and Sex = 1 is the sum of y_1 and y_2 and will have the value 1 for individuals that are young and male and either employed or unemployed and the value 0 for individuals that are not both young and male. Similarly, we obtain for each of the four Age $\times$ Sex combinations zero-one auxiliary variables as the sum of the corresponding target variables for Unemployed and Employed. In general, if there are p target variables and q auxiliary variables corresponding to sums of target variables, we can write the values of the auxiliary variables as

$$\mathbf{z}_i = \mathbf{C}\mathbf{y}_i, \tag{9}$$

with $\mathbf{C}$ the $q \times p$ constraint matrix (consisting of zeroes and ones) that generates the sums of the y_i values corresponding to the auxiliary variables. Since (9) applies to each row of $\mathbf{Z}$ and $\mathbf{Y}$, we can write $\mathbf{Z} = \mathbf{Y}\mathbf{C}'$ and so

$$\mathbf{B} = \mathbf{Y}'\mathbf{\Pi}^{-1}\mathbf{Y}\mathbf{C}'\left(\mathbf{C}\mathbf{Y}'\mathbf{\Pi}^{-1}\mathbf{Y}\mathbf{C}'\right)^{-1}. \tag{10}$$

In the case considered here, where the target variables correspond to cells in a cross-classification of categorical variables, this expression can be simplified as follows. The rows of $\mathbf{Y}$ contain a 1 in the column corresponding the cell to which the unit belongs and zeroes elsewhere. After rearranging the rows such that the units that belong to the same cell (score a one on the same target variable) are beneath each other, $\mathbf{Y}$ can be written as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_q} \end{pmatrix},$$

where n_j is the number of units scoring a one on target variable j and $\mathbf{1}_{\mathbf{n_j}}$ is a column with n_j ones. In this example there are no units that score on the third target variable. When this matrix is premultiplied by $\mathbf{Y}'\mathbf{\Pi}^{-1}$ we obtain $\mathbf{Y}'\mathbf{\Pi}^{-1}\mathbf{Y} = Diag(\hat{\mathbf{x}}_{y.ht})$ and $\mathbf{B}$ can be expressed as

$$\mathbf{B} = Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}'\left(\mathbf{C}Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}'\right)^{-1}. \tag{11}$$

Substituting this value for $\mathbf{B}$ in (8) and using $\mathbf{C}\hat{\mathbf{x}}_{y.ht} = \hat{\mathbf{x}}_{z.ht}$ we obtain

$$\hat{\mathbf{x}}_{y.greg} = \hat{\mathbf{x}}_{y.ht} + Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}' \left(\mathbf{C}Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}'\right)^{-1} (\mathbf{x}_{z.pop} - \mathbf{C}\hat{\mathbf{x}}_{y.ht}), \quad (12)$$

which is equal to (4) with the initial unadjusted vector ($\mathbf{x}$) equal to the Horwitz-Thompson estimators for the cell-totals, the weighting matrix ($\mathbf{A}^{-1}$) a diagonal matrix with the initial vector along the diagonal and the values of the constraints (b) equal to the known population totals of the margins of the contingency table that are used as auxiliary variables.

2.3 Extension to time series data

The optimization problem described in 2.1 can be extended to time series data of the form x_{it} $(i = 1, \ldots, N,\ t = 1, \ldots, T)$. In this case the total number of the variables x_{it} is $N \cdot T$ and the constraint matrix will have $N \cdot T$ columns. The number of rows will be equal to the number of constraints as before. The matrix A will be a symmetric, $NT \times NT$ nonsingular matrix.

For this data we want to find adjusted values $\widehat{x}_{it}$ that are in some metric ς (for example Euclidean metric) close to the original time series. For this purpose we consider the following objective function

$$\min_{\widehat{x}} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{w_{it}} \varsigma(\widehat{x}_{it}, x_{it}), \quad (13)$$

where w_{it} denotes the variance of the i^{th} time series at time t. We minimize this function over all $\widehat{x}_{it}$ satisfying the constraints

$$\sum_{i=1}^{N} \sum_{t=1}^{T} c_{rit} \widehat{x}_{it} = b_r, \quad r = 1, \ldots, C. \quad (14)$$

In (14), r is the index of the restrictions and C is the number of restrictions. Furthermore, c_{rit} is an entry of the restriction matrix and b_r are fixed constants. Most economic variables cannot have negative signs. To incorporate this (and other) requirement(s) in the model, inequality constraints are included. A set of inequalities is given by

$$\sum_{i=1}^{N} \sum_{t=1}^{T} a_{rit} \widehat{x}_{it} \leq z_r, \quad r = 1, \ldots, I, \quad (15)$$

where I stands for the number of inequality constraints.

In Bikker et al. 2010 this model was extended by soft linear and ratio restrictions. A soft equality constraint is different from the hard equality constraints (14), in that the constants b_r are not fixed quantities but are assumed to have a variance and an expected value. This means that the resulting $\widehat{x_{it}}$ need not

match the soft constraints exactly, but only approximately. A soft linear constraint similar to (14) is denoted as follows:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} c_{rit}\widehat{x}_{it} \sim (b_r, w_r), \quad r = 1, \ldots, C. \tag{16}$$

By the notation $\sim$ in (16) we define b_r to be the expected value of the sum $\sum_{i=1}^{N}\sum_{t=1}^{T} c_{rit}\widehat{x}_{it}$ and w_r its variance. In the case that ς is the Euclidean metric the linear soft constraints can be incorporated in the model by adding the following term to the objective function in (13):

$$+\sum_{r=1}^{C}\frac{1}{w_r}\left(b_r - \sum_{i=1}^{N}\sum_{t=1}^{T} c_{rit}\widehat{x}_{it}\right)^2. \tag{17}$$

Another important extension of the model in Bikker et al. 2010 is the ratio constraint. The hard and soft ratio constraints that can be added to the model, are given by

$$\frac{\widehat{x}_{nt}}{\widehat{x}_{dt}} = v_{ndt} \quad \text{and} \quad \frac{\widehat{x}_{nt}}{\widehat{x}_{dt}} \sim (v_{ndt}, w_{ndt}), \tag{18}$$

where $\widehat{x}_{nt}$ denotes the numerator time series, $\widehat{x}_{dt}$ denotes the denominator time series, v_{ndt} is some predetermined value and w_{ndt} denotes the variance of a ratio $\frac{\widehat{x}_{nt}}{\widehat{x}_{dt}}$. In order to add the soft ratio constraints to the objective function these are first linearized. The soft constraints in (18) can be rewritten as:

$$\widehat{x}_{nt} - v_{ndt}\widehat{x}_{dt} \sim (0, w^*_{ndt}). \tag{19}$$

The variance of the constraint will be different, we denote it as w^*_{ndt}. Soft linearized ratios are incorporated in the model in case when ς is a Euclidean metric, by adding the following term to the objective function

$$+\sum_{n,d=1}^{N}\sum_{t=1}^{T}\frac{(\widehat{x}_{nt} - v_{ndt}\widehat{x}_{dt})^2}{w^*_{ndt}}. \tag{20}$$

The extensions of the constraints that can be handled beyond the traditional linear (in)equality constraints, opens up a number of applications to reconciliation problems in several areas. An example of one such application is described in section 4

3 Reconciliation of census tables

In this section we describe the Dutch Census data and formulate the reconciliation of census tables as a macro-integration problem.

The aim of Census 2011 is to produce 60 hypercubes about demographics and occupation. For each of these hypercubes the figures should be produced for the

whole Dutch population, for each province and for each municipality. Consisting in the end from a great number of hypercubes. For this task, data from many different sources and different structures are combined. The majority of the variables are obtained from the Central Population Register (CPR), however quite a few other sources (sample surveys and registers) are used as well, such as for example the labour force survey (LFS).

Each table consists of up to 10 variables. We call these high dimensional crosstables hypercubes. Most of the variables are included in many hypercubes. The hypercubes have to be consistent with each other, in a sense that all marginal distributions that can be obtained from different crosstables are the same. Consistency is required for one dimensional marginals, e.g. the number of men, as well as for multivariate marginals, e.g. the number of divorced men aged between 25 and 30 year. Multivariate marginal crosstables are hypercubes as well.

In different hypercubes one variable may have a different category grouping (classification). For example, the variable age can be requested to be included in different hypercubes aggregated in different levels of detail: groups of ten years, five years and one year. Still, the marginal distributions of age obtained from different hypercubes should be the same for each level of aggregation.

In general, the data that are collected by Statistics Nederlands (SN) involve many inconsistencies; the cause of this varies: different sources, differences in population coverage, different time periods of data collection, nonresponse, or modeling errors.

Currently at SN, the method of repeated weighting (see Houbiers 2004) is used to combine variables from different sources and to make them consistent. Using repeated weighting, tables are reconciled one by one. Assuming that the tables 1 till t are correct, these figures are fixed. Then, the method of repeated weighting adjusts table $t+1$, so that all marginals of this table become consistent with the marginals of all previous tables, 1 till t. The method of repeated weighting was successfully used for the last census in 2001. However, the number of the tables has increased since and with the number of tables the number of restrictions also increased. As a consequence, it is not obvious that the method of repeated weighting will work for the Census 2011.

The method of macro-integration has some advantages over repeated weighting. Firstly, the method of macro-integration reconciles all tables simultaneously, meaning that none of the figures need to be fixed during the reconciliation process. By doing so, there are more degrees of freedom to find a solution than in the method of repeated weighting. Therefore a better solution may be found, which requires less adjustment than repeated weighting. Secondly, the results

of repeated weighted depend on the order of weighting the different tables, while the macro-integration approach does not require any order. Thirdly, the method of macro-integration allows inequality constraints, soft constraints and ratio constraints, which may be used to obtain better results.

A disadvantage of macro-integration is that a very large optimization problem has to be solved. However, by using up-to-date solvers of mathematic optimization problems, very large problems can be handled. The software that has been built at Statistics Netherlands for the reconciliation of National Accounts tables is capable of dealing with a large number of variables (500 000) and restrictions (200 000).

We should emphasize that reconciliation should be applied on the macro level. First the imputation and editing techniques should be carried out for each source separately on the micro level. The aggregated tables should then be produced, containing variables at the publication level. Furthermore, for each separate aggregated table, a variance of each entry in the table should be computed, or at least an indication of the variance should be given. For example, an administrative source will in general have the most reliable information, and hence have a very small or zero variance. Each aggregated table separately can be optimized to get variance as small as possible, by imputation or other means. During the reconciliation process, each entry of all tables will be adapted in such a way that the entries with the highest variance will be adapted the most, until all constraints are met.

The procedure that we propose here is as follows:

1. For each data source define the variables of interest;
2. Use imputation and editing techniques to improve data quality on a micro level;
3. Aggregate the data to produce the tables, and calculate the variances of each entry;
4. Use reconciliation to make the tables consistent. Calculate the covariance matrix for the reconciled table.

We have identified different kinds of reconciliation problems for census data:

I For some variables we will have different classifications, for example the variable Age can be in years, or five year intervals or ten year intervals. It is required that number of persons obtained from the hypercube with the variable Age with one year intervals for example from 10 till 20 years should add up to the number of persons of this age interval obtained from

any other hypercube, where Age is measured in five or ten years intervals. The objective function and the constraints can be modified in order to handle this problem.

II In the macro-integration approach presented in this paper, the reconciliation is carried out at the macro level. It is assumed that an initial estimate for each hypercube can be made. However the estimation of these hypercubes is not always straightforward. This is especially the case for hypercubes that include variables from different data sources: for example a register and a sample. An example of such a case is considered in Appendix A In this example we have three variables (Province, Gender and Age) obtained from the register and a sample that contains these three variables and the additional variable: Occupation. In Appendix A we combine these two data sets using macro-integration notations, to obtain an initial estimate of one hypercube. This reconciliation problem is very simple and macro-integration might not be the first choice to solve the problem. However it immediately becomes complex when the number of hypercubes and the number of variables in these hypercubes increase.

III A problem that has to be solved in any method is the lack of information. Part of the source information is based on samples. However, these samples may not cover each of the categories of the variables in the hypercubes. For instance, a sample may not include any immigrant from Bolivia, while this sample may be the only source for some of the variables in the census.

3.1 The objective function

We distinguish two steps while making the census hypercubes:

1. At first the hypercubes should be made from all available sources;
2. Then for all hypercubes we should make the same marginals equal;

Building of the census hypercubes from different sources could be carried out using many different methods, like weighting or post-stratification. In Appendix A we present a simple example of making a hypercube using two different data sources. In this section we will not discuss these methods. From the macro-integration point of view the second step of making the hypercubes is of our interest.

Using the notation from the previous section we can now apply the macro-integration method for reconciliation of the hypercubes by their common marginals.

In the previous section we defined the objective function (13) using an arbitrary metric. Here we use a Euclidean metric.

We introduce the notations especially for census data. For $j = 1, \ldots, N$, a hypercube is defined by $H^{(j)}$, and any of this marginal hypercube is defined by $M^{(j)}$. A variable in the hypercube $H^{(j)}$ is defined by $x_i^{(j)}$, where the subindex i is the identity of the variable, for example Province or Age and the super index (j) identifies the hypercube where the variable is in. For example, if we have two hypercubes $H^{(1)}$ and $H^{(2)}$, the variables from $H^{(1)}$ will be defined by $x_1^{(1)}, x_2^{(1)}, \ldots, x_m^{(1)}$, assuming that the hypercube $H^{(1)}$ consists of m variables. Suppose now that the hypercube $H^{(2)}$ consists of n variables and it has three variables $x_1^{(2)}, x_2^{(2)}$ and $x_4^{(2)}$ in common with the hypercube $H^{(1)}$. Denote the marginal hypercube of $H^{(1)}$ consisting of these variables by $M_{1,2,4}^{(1)}$:

$$M_{1,2,4}^{(1)} = x_1^{(1)} \times x_2^{(1)} \times x_4^{(1)}.$$

Reconciling the hypercubes $H^{(1)}$ and $H^{(2)}$ so that their common marginal hypercubes are the same will mean the finding of hypercubes $\widehat{H}^{(1)}$ and $\widehat{H}^{(2)}$ such that:

$$\varsigma(H^{(1)}, \widehat{H}^{(1)}) + \varsigma(H^{(2)}, \widehat{H}^{(2)}) \tag{21}$$

reaches its minimum under the condition that:

$$\widehat{M}_{1,2,4}^{(1)} = \widehat{M}_{1,2,4}^{(2)}. \tag{22}$$

In the case when the first marginal hypercube $M_{1,2,4}^{(1)}$ consists of the variables from a register, that are fixed and should not be reconciled, then instead of the condition in (22) we will have the following

$$\widehat{M}_{1,2,4}^{(2)} = M_{1,2,4}^{(1)}. \tag{23}$$

We can now define the objective function for the reconciliation of the hypercubes $H^{(j)}$, $j = 1, \ldots, N$. We want to find the hypercubes $\widehat{H}^{(j)}$, $j = 1, \ldots, N$ such that:

$$\min_{\widehat{H}} \sum_j \varsigma(H^{(j)}, \widehat{H}^{(j)}), \tag{24}$$

under the restriction that, all common marginal hypercubes are the same

$$\widehat{M}_{i,k,\ldots,l}^{(j_1)} = \cdots = \widehat{M}_{i,k,\ldots,l}^{(j_k)} \tag{25}$$

and the marginal hypercubes consisting of register data that are fixed, will not be changed:

$$\widehat{M}_{i,k,\ldots,l}^{(j_1)} = \cdots = M_{i,k,\ldots,l}^{(j_n)}. \tag{26}$$

If we transform the hypercube $H^{(j)}$ into a vector $\mathbf{h^{(j)}} = (\mathbf{h_1^{(j)}}, \ldots, \mathbf{h_{c_j}^{(j)}})'$ we can rewrite the objective function in (24) using the notation of the previous section. For all $\mathbf{h^{(j)}}$, $j = 1, ..., N$, we want to find vectors $\widehat{h}^{(j)}$, $j = 1, ..., N$ such that:

$$\min_{\widehat{\mathbf{h}}} \sum_{j=1}^{N} \sum_{i=1}^{c_j} \frac{1}{w_{ij}} \left(\widehat{h}_i^{(j)} - h_i^{(j)} \right)^2, \tag{27}$$

where w_{ij} is the weight of $h_i^{(j)}$.

3.2 Reconciliation of two hypercubes

Suppose we want to create two hypercubes, each with three variables. Hypercube one $H^{(1)}$ consists of variables Gender, Age and Occupation and the second hypercube, $H^{(2)}$ of the variables Gender, YAT (year of immigration) and Occupation. For convenience, we combine the original categories of these variables and consider the coding as it is presented in Table 1. From these variables the

Table 1. Categories of the variables

Gender	1	Male
	2	Female
Age	1	$<$ 15 years
	2	15-65 years
	3	$>$ 65 years
Occupation	0	Not manager
	1	Manager
YAT	0	Not immigrant
	1	Immigrated in 2000 or later
	2	Immigrated before 2000

only one that is observed in the survey is Occupation, the other three variables are obtained from the register and are therefore assumed to be fixed. The survey we use here is the LFS (labour force survey) and the register is the GBA (municipality data bases). As we mentioned already we assume that the figures obtained from GBA are exogenous, what means that these values should not be changed.

We aim to find the hypercubes $\widehat{H}^{(1)}$ and $\widehat{H}^{(2)}$ such that

$$\varsigma(H^{(1)}, \widehat{H}^{(1)}) + \varsigma(H^{(2)}, \widehat{H}^{(2)}) \tag{28}$$

is minimized under the restrictions that the marginal hypercubes of $\widehat{H}^{(1)}$ and $\widehat{H}^{(2)}$ coincide with the corresponding marginal hypercubes of the register. Hence we want to achieve that:

$$\widehat{M}^{(1)}_{\text{Gender, Age}} = M^{\text{register}}_{\text{Gender, Age}} \tag{29}$$

and

$$\widehat{M}^{(2)}_{\text{Gender, YAT}} = M^{\text{register}}_{\text{Gender, YAT}}. \tag{30}$$

In addition, the hypercubes should be reconciled with each other:

$$\widehat{M}^{(1)}_{\text{Gender, Occupation}} = \widehat{M}^{2}_{\text{Gender, Occupation}}; \tag{31}$$

Table 2. Hypercube 1

Sex	Age	Occup	0	I	II	III	IV	V
1	1	0	1761176	1501748	1501748	1501748	1501748	1501748
1	2	0	5181009	5065650	5065650	4924068	4907253	4916858
1	2	1	674373	507128	507128	648710	665525	655920
1	3	0	584551	831315	831315	1016430	1016072	1016276
1	3	1	13011	207889	20788	22774	23132	22928
2	1	0	1661478	1434236	1434236	1434236	1434236	1434236
2	2	0	5755370	5521997	5484427	5254234	5247781	5251467
2	2	1	241251	-37570	0	230193	236646	232960
2	3	0	534231	976868	986261	1370781	1370724	1370757
2	3	1	2037.85	399226	389833	5313	5370	5337

The first step before the actual reconciliation process is weighting up the sample to the population. The total number of GBA persons is $N_{GBA} = 16\,408\,487$ and the total number of LFS persons is $N_{LFS} = 104\,674$. The initial weight is

$$w = \frac{16\,408\,487}{104\,674}.$$

Table 3. Hypercube 2

Sex	YAT	Occup	0	I	II	III	IV	V
1	0	0	6723037	6505428	6505428	6378041	6362791	6371502
1	0	1	609945	444221	444221	571608	586858	578147
1	1	0	179174	213134	213134	291188	290865	291049
1	1	1	12697	98543	98543	20489	20812	20628
1	2	0	624524	680151	680151	773017	771417	772331
1	2	1	64741	172253	172253	79387	771417	772331
2	0	0	6965385	6889146	6879753	6870198	6864427	6867723
2	0	1	215699	184908	194301	203856	209627	206331
2	1	0	232472	253743	244350	319060	318945	319010
2	1	1	4232	70951	80344	5634	5749	5684
2	2	0	753222	790213	780820	869994	869369	869726
2	2	1	23357	105796	115189	26015	26640	26283

The results of the weighting are presented in Tables 2 and 3 under the column 0. Since we consider these figures as the starting figures before the reconciliation process, we call these model 0. These figures have marginals consistent with each other but not with the register data, see Table 4. For example, the total number of men is 8214119 from Table 2 and 3 and 8113730 in Table 4.

We applied the optimization solver XPRESS for the problem defined in (28-31) using the Euclidean distance for ς and applying the weight 1 for all figures. The results of this reconciliation are presented in Tables 2 and 3 under the column I. We observed negative figures after the reconciliation, therefore we added the restriction that all figures have to be nonnegative to the previous setting and applied the solver. Results of this optimization problem are presented in Tables 2 and 3 under the column II. Next we used weights equal to the initial value of each figure. The results of this execution are to be found under the column III in Tables 2 and 3. Applying more realistic weights led to different results, compared with models I and II, the figures with smaller values are adjusted less and the figures with bigger values are adjusted more.

Table 4. Register

Sex	Age	YAT	Total
1	1	0	1437385
1	1	1	48553
1	1	2	15810
1	2	0	4619201
1	2	1	255335
1	2	2	698242
1	3	0	893063
1	3	1	7789
1	3	2	138352
2	1	0	1369468
2	1	1	49026
2	1	2	15742
2	2	0	4502083
2	2	1	267916
2	2	2	714428
2	3	0	1202503
2	3	1	7752
2	3	2	165839

Since we want to preserve the initial marginal distribution of the variable Occupation, the next step is to add a ratio restriction. We only added one ratio restriction, that is the relation between the managers and non managers for the whole population. At first we added this restriction as a hard constraint and afterwards as a soft constraint to the model. The results of these reconciliation problems are presented in columns IV and V of Tables 2 and 3. For the soft restrictions the weight we choose is equal to 707405400, which is 100 times the largest register value. This value is found by trial and error. By choosing this value the ratio constraints significantly influences the results, but its effect is clearly less than that of a hard ratio constraint.

In Table 5 the ratios of the number of 'not manager' over the number of 'manager' is calculated for the models III, IV and V. The target value of the ratio is the ratio observed in LFS. As we could expect the value is best achieved in

Table 5. Ratio restriction

Model scenario	Ratio
Target value	16.631
Model outcome: no ratio (III)	17.091
Model outcome: hard ratio (IV)	16.631
Model outcome: soft ratio (V)	16.891

model IV, when the hard ratio restriction has to be fulfilled.

To compare the results of the models with each other we calculated the weighted quadratic difference between the reconciled values of models III, IV and V and the values of model 0, the hypercubes after the weighting, see Table 6.

Table 6. Weighted squared difference

Model scenario	Difference
Model 0 - Model III	1955390
Model 0 - Model IV	1956704
Model 0 - Model V	1956696

The weighted squared difference in Table 6 is calculated as follows

$$\sum_{j=1}^{2}\sum_{i=1}^{c_j}\frac{1}{w_{ij}}\left(\widehat{h}_i^{(j)}-h_i^{(j)}\right)^2, \tag{32}$$

here we sum over two hypercubes, $\widehat{h}_i^{(j)}$ are the reconciled figures of model III, IV or V and $h_i^{(j)}$ are the values of model 0. The weighted squared difference is smallest for model III, which implies that without the ratio restriction reconciled figures are closer to the original figures. We could anticipate this result since the ratio restriction (as any additional restriction would do) forces the original figures towards the distribution of the ratio and therefore the outcome of the model with the hard ratio restriction differs most from the initial values.

4 Early estimates for labour market

The second application of macro-integration methods that we want to study is making early estimates for the labour market variables. This is a very complex task, mainly caused by the variety of data sources, that contain variables with almost equal, but not exactly the same definitions. The sources include one or more labour market variables. The main data sources for labour market variables are the tax office data and the Dutch labour Force Survey (LFS). The tax office data are updated on a quarterly basis and LFS is a rotating panel design producing monthly figures. Among others we want to combine these two

data to construct the early estimates of statistics that are at the moment based only on LFS or tax register data. The difficulties that should be resolved are of a different nature:

- First of all we have series of data of different frequency (in time);
- Definitions of the variables in both sources are often very different;
- Population coverage is not the same;
- Other difficulties such as survey vs register data, nonresponse, etc.

Because of the different variable definitions of the different data sources the labour market data is more difficult to reconcile than the National Accounts data. It will not be possible to combine the labour market variables on a macro level, without first studying thoroughly the data structure.

Table 7. Population

Age	Sex		Total
	Woman	Man	
20 - <30	6600	6000	12600
30 - <40	6000	7200	13200
40 - <50	9600	8400	18000
$\geq$ 50	9000	7200	16200
Total	31200	28800	60000

We give here a simple example using the macro-integration method for combining two different sources. Suppose we have a labour force population of 60000 persons. We want to conduct a monthly labour forse survey to find out an unemployment rate. Suppose for simplicity that the auxiliary variables in the population register of our interest are: Sex and Age. We will know the distribution of these variables according to our register, see Table 7;

Table 8. Survey

Age	Sex		Total
	Woman	Man	
20 - <30	660	600	1260
30 - <40	600	720	1320
40 - <50	960	840	1800
$\geq$ 50	900	720	1620
Total	3120	2880	6000

Suppose for convenience that the total number of respondents in the survey we want to conduct is 6000. We divide 6000 over all cells of Age and Sex according to the register data, see Table 8; In the survey we observe two variables: whether a person has a job and if not whether she/he is registered at the unemployment

Table 9. Survey unemployment data

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20 - <30	35 (25)	34 (27)	36 (25)	35 (29)	37 (33)	33 (30)
30 - <40	40 (38)	35 (32)	42 (37)	36 (32)	42 (35)	37 (35)
40 - <50	60 (50)	56 (50)	58 (51)	56 (49)	61 (58)	58 (55)
$\geq$ 50	42 (30)	38 (25)	42 (31)	40 (31)	43 (35)	40 (38)

office (CWI). Suppose for simplicity that we do not have nonresponse and that the Table 9 of unemployment numbers, is the result of the survey for three months, January, February and March. In parenthesis are the numbers of respondents registered at CWI; From Table 9 we can estimate the number of

Table 10. Weighted unemployment data

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20 - <30	350 (250)	340 (270)	360 (250)	350 (290)	370 (330)	330 (300)
30 - <40	400 (380)	350 (320)	420 (370)	360 (320)	420 (350)	370 (350)
40 - <50	600 (500)	560 (500)	580 (510)	560 (490)	610 (580)	580 (550)
$\geq$ 50	420 (300)	380 (250)	420 (310)	400 (310)	430 (350)	400 (380)

unemployed persons in each group of the population, see Table 10. On the other hand from the unemployment office (CWI) we have the number of persons that were registered as unemployed at the end of the quarter, see Table 11. Suppose that we do not have the timeliness issues for the survey and register and both data are available for us at around the same time. The estimated registered

Table 11. Unemployment office data at the end of the first quarter

Age	Sex	
	Woman	Man
20 -<30	350	330
30 -<40	390	360
40 -<50	600	570
$\geq$ 50	370	395

unemployment figures from the survey and from the CWI are not consistent with each other. For example there are 330 women of age 20 -<30 registered according to the survey and 350 women according to the register at the end of March. Let us suppose that the register has a variance equal to zero, which means that 350 is a fixed figure.

Now, we want to achieve consistency between the labour force survey and the unemployment office register, in such a way that

1. The weighted survey data may be adjusted, while the CWI register data are fixed.

2. The numbers of persons registered at the CWI at the end of the third month in the reconciled survey data exactly matches the corresponding numbers in the unemployment register.

3. The ratios between the unemployment numbers and the numbers of persons registered at the CWI have to be as close as possible to their initial values as observed in the weighted labour force survey. For instance, for women of age 20-29 this ratio is 37/33 at the end of March. These ratios do not have to hold exactly, as they are observed in the sample.

4. All monthly changes of the number of unemployed persons are as close as possible to their initial value as observed in the weighted survey.

5. All monthly changes of the number of persons registered at the CWI are as close as possible to their initial value as observed in the weighted survey.

We will use a macro-integration model to reconcile the labour force survey with the register of the unemployment office. Define the weighted survey figures of unemployment by x_{ijt} and the number of persons registered at the CWI by y_{ijt}, where t stands for the month and i and j denote the entries of the matrix Age×Sex. The ratios between x_{ijt} and y_{ijt} will be denoted by d_{ijt} (i.e. $d_{ijt} = x_{ijt}/y_{ijt}$). Then, we want to find estimates $\widehat{x}_{ijt}$ of x_{ijt} and $\widehat{y}_{ijt}$ of y_{ijt} that satisfy the properties (1)-(5) listed above. The formulation of the model is

$$\begin{aligned} \min_{\widehat{y}\widehat{x}} \sum_{t=2}^{T} \sum_{ij} & \frac{((\widehat{x}_{ijt} - \widehat{x}_{ijt-1}) - (x_{ijt} - x_{ijt-1}))^2}{v_{1ij}} \\ + & \frac{((\widehat{y}_{ijt} - \widehat{y}_{ijt-1}) - (y_{ijt} - y_{ijt-1}))^2}{v_{2ij}} \\ + & \frac{(\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt})^2}{w^*_{ij}}, \end{aligned} \tag{33}$$

with

$$\widehat{y}_{ijt} = y^{CWI}_{ijk}, \quad \text{for all} \quad i, j \quad \text{and} \quad t = 3, k = 1. \tag{34}$$

where v_{1ij} denotes the variance of x_{ijt}, v_{2ij} the variance of y_{ijt} and w^*_{ij} is the variance of the linearized ratio $\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt}$.

The first term of (33) keeps the differences $\widehat{x}_{ijt} - \widehat{x}_{ijt-1}$ as close as possible to their initial values $x_{ijt} - x_{ijt-1}$ (the aforementioned property 4) and the second term does the same for y_{ijt} (property 5). The third term describes the soft

ratio restrictions for the relation between unemployment and registering at the CWI (property 3). They are similarly defined as the ratio constraints in (20). Here, we assume that the variances of the linearised ratios do not depend on t. The hard constraints in (34) ensure that the estimates of y_{ijt} of the last month ($t = 3$) are equal to the quarterly unemployment number of the first quarter ($k = 1$), as obtained from the CWI register y_{ijt}^{CWI} (property 2). Note that the quarterly unemployment numbers of the CWI y_{ijt}^{CWI} are included in the model as parameters only. They are not specified as free variables, because these figures are fixed (property 1).

The results of the model (33) - (34) where we have taken all variances $v_{1ij}, v_{2ij}, w_{ij}^{*}$ to be equal to 300, are shown in Table 12. These show that the number of persons registered at the CWI (the numbers between parenthesis) at the end of March are indeed consistent with the unemployment office register (as depicted in Table 11).

Table 12. Reconciled unemployment data

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20 -<30	375.1 (268.0)	375.2 (298.3)	385.0 (268.2)	393.7 (319.0)	393.7 (350.0)	363.8 (330.0)
30 -<40	445.2 (422.0)	360.8 (329.9)	466.3 (410.9)	467.1 (329.8)	467.1 (390.0)	380.7 (360.0)
40 -<50	622.4 (519.0)	581.9 (519.5)	602.0 (529.4)	631.5 (509.5)	631.5 (600.0)	601.5 (570.0)
$\geq$ 50	446.0 (318.8)	398.3 (262.5)	445.7 (329.2)	455.2 (323.7)	455.2 (370.0)	416.7 (395.0)

To illustrate the preservation of changes and ratios we focus, as an example, on the number of women in the age 20-29. Figure 1 shows that the initial monthly changes are preserved quite accurately and from Table 13 it can be seen that the same holds true for the ratios between the number of unemployed and CWI registered persons. Figure 1 also shows that the data reconciliation increases both the number of CWI registered persons and the number of unemployed people at each time period. The explanation is that number of CWI registered persons in the survey at the end of month 3 is below the register figure at the end of the first quarter. Since the survey has to exactly match the register and since all monthly changes of the survey have to be preserved as much as possible, all monthly survey figures on the number of CWI registered persons are increased. The same occurs to the number of unemployed persons, which can be explained from the preservation of the ratios between the number of unemployed and CWI registered people at each time period.

Now, suppose that we decrease the variance of the monthly changes from 300 to 100, but we do not change the variance of the ratios between unemployed and CWI registered persons. As a result, the initial quarterly changes are preserved better at the expense of the ratio between the number of unemployed and CWI

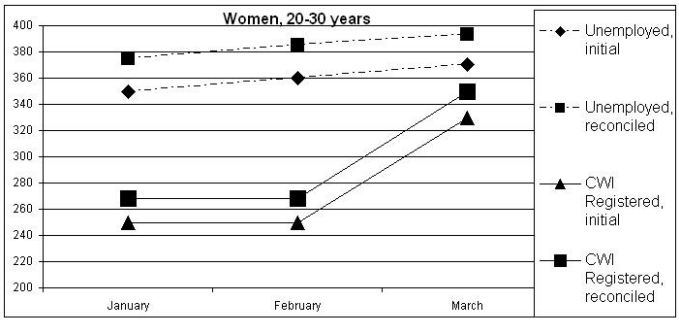

Figure 1. Women of 20-29 years; initial and reconciled unemployment and CWI Registered

Table 13. Women of 20-29 years; ratio between unemployed and CWI registered persons

	January	February	March
Initial data	1.400	1.440	1.121
Reconciled data 1	1.400	1.436	1.125

registered persons, which becomes clear by comparing the results in Table 14 with the results in Table 13.

Table 14. Women of 20-29 year; ratio between unemployed and CWI registered persons, scenario 2

	January	February	March
Initial data	1.400	1.440	1.121
Reconciled data 2	1.396	1.432	1.130

This is of course only a simple example where we only have two data sources and did not take into account many issues that will occur while combining different sources. However the method can be extended further.

5 Conclusions

Reconciliation of tables on a macro level can be very effective, especially when a large number of constraints should be fulfilled. Combining data sources of different structures, on a macro level is often easier to handle than on a micro-level. When data are very large and many sources should be combined macro-integration could be the only technique that can be used. Macro-integration is also more versatile than (re-)weighting techniques using GREG-estimation in the sense that inequality constraints and soft constraints can be incorporated easily.

The two examples considered in this paper are of great importance for SN: For the census, further developing the macro-integration approach can lead to solutions for consistency problems that could become very needed when application of the repeated weighting method is hampered by its limitations. Since macro-integration can be an alternative to (repeating) weighting methods it is very important to compare results obtained by these methods and to further investigate the applicability of both methods in various situations and their respective weak and strong points.

The second application is equally, if not more important for SN. For the past couple of years SN has an additional data source for labour force data, the Tax Office register. The use of this register has increased enormously over last years, as is the quality of data and the understanding of the variables that the tax office collects. At the same time, SN has taken means to improve the quality of the surveys, such as the labour force survey. Improving the quality of these data sources (especially of the Tax Office register data) creates the possibility for reconciliation of the survey data with the register data. Macro-integration could be applied to reconcile the register and survey data. Currently SN publishes variables obtained from each data source separately. Many of these variables are very similar but not entirely the same. Future research could show if, and how, it is possible to combine these sources in order to produce one set of figures.

References

R. Bikker and S. Buijtenhek. Alignment of quarterly sector accounts to annual data. Technical report, Statistics Netherlands, 2006.

R. Bikker, J. Daalmans, and N. Mushkudiani. A multivariate denton method for benchmarking large data sets. Technical report, Statistics Netherlands, 2010.

H.J. Boonstra. Calibration of tables of estimates. Technical report, Statistics Netherlands, 2004.

H.J. Boonstra, C. de Blois, and G.J. Linders. Macro-integration with inequality constraints-an application to the integration of transport and trade statistics. Technical report, Statistics Netherlands, 2010.

M. Houbiers. Towards a social statistical database and unified estimates at statistics netherlands. *Journal of official statistics*, 20:55–75, 2004.

C.-E. Särndal, B. Swenson, and J.H. Wretman. *Model assisted Survey Sampling*. Springer-Verlag, New York, 1992.

J.R.N. Stone, D.A. Champerowne, and J.E. Maede. The precision of national income accounting estimates. *Reviews of Economic Studies*, 9:111–125, 1942.

A Census data example

In this section we will construct a simple hypercube using two data sources. Consider two data sets: one is obtained from GBA (municipality data bases) register and the other is from LFS (labour force survey). The first data set consists of three variables: Province, Sex and Age and the second data set contains one additional variable: Occupation.

Table A.1. Categories of variable Province

Unknown	1
Groningen	2
Friesland	3
Drenthe	4
Overijssel	5
Flevoland	6
Gelderland	7
Utrecht	8
Noord-Holland	9
Zuid-Holland	10
Zeeland	11
Noord-Brabant	12
Limburg	13

For simplicity assume that the three common variables have the same categories in both data sets. Province has 13 categories, see Table A.1. The variable age is grouped in five year intervals and has 21 categories: $0- < 5, 5- < 10, ..., 95- < 100, 100+$. Sex has 2 categories and occupation 12 categories, see Table A.2.

Table A.2. Categories of variable occupation

Not stated	1
Armed forces occupations	2
Managers	3
Professionals	4
Technicians and associate professionals	5
Clerical support workers	6
Service and sales workers	7
Skilled agricultural, forestry, and fishery workers	8
Craft and related trades workers	9
Plant and machine operators, and assemblers	10
Elementary occupations	11
Not applicable	12

The data are initially available on the micro level. The total number of GBA persons is $N_{GBA} = 16\,408\,487$ and the total number of LFS persons is $N_{LFS} = 104\,674$. Both data sets were aggregated up to the publication level. The cross tables obtained are three and four dimensional hypercubes. The values

of hypercube obtained from the second sample is then adjusted using the same weights for each cell. The initial weight is then defined as follows:

$$w = \frac{16\,408\,487}{104\,674}.$$

We assume that the figures of the first data set (obtained from the GBA) are exogenous. That means these values will not be changed.

Suppose that in the variables defined by $x_i^{(j)}$ the subindex i will define the identity of the variable for example Province and the super index will define the data set where the variable will originate from. In our example we have two data sets, hence $j = 1$ or 2. For convenience, the variables Province, Sex and Age are numbered by 1, 2 and 3. In the first data set these variables are defined by $x_1^{(1)}, x_2^{(1)}$ and $x_3^{(1)}$. Similarly, in the second data set the variables Province, Sex, Age and Occupation are defined as $x_1^{(2)}, x_2^{(2)}$, $x_3^{(2)}$ and $x_4^{(2)}$. We define the marginal distribution of the variable $x_i^{(j)}$ as follows:

$$x_{i,1}^{(j)}, \ldots, x_{i,r_i}^{(j)},$$

the second index here defines the categories of the variable. For example, the variable Province x_1 has 13 categories, $r_1 = 13$. Each hypercube will have a

Table A.3. A part of the second hypercube

Province	Sex	age	Occupation	Number of persons
2	2	8	12	51
2	2	8	3	12
2	2	8	4	22
2	2	8	5	23
2	2	8	6	22
2	2	8	7	18
2	2	8	8	1
2	2	8	9	2
2	2	8	10	1
2	2	8	11	9

crosstable of variables, containing the values

$$x_{1,j}^{(1)} \times x_{2,k}^{(1)} \times x_{3,l}^{(1)}, \quad j = 1, \ldots, 13, \quad k = 1, 2, \quad l = 1, \ldots, 21.$$

For example, when $j = 2$, $k = 2$ and $l = 8$ we have that

$$x_{1,2}^{(1)} \times x_{2,2}^{(1)} \times x_{3,8}^{(1)} = 20422$$

this means that there live 20422 women of age between 35 and 40 in the province Groningen. In the second data set we also have the extra variable Occupation. In case when $j = 2$, $k = 2$ and $l = 8$ the number of persons in each category of the variable Occupation are presented in Table A.3. Note that it is the part of

the hypercube consisting of four variables. Observe that there are no persons in this hypercube with the categories 1 and 2 for the variable Occupation.

$$x_{1,2}^{(2)} \times x_{2,2}^{(2)} \times x_{3,8}^{(2)} \times \sum_{i=1}^{12} x_{4,i}^{(2)} = 161$$

We want to combine these two data sets into one. We can do this using the macro-integration method. For this simple example it is similar to post stratification methods. However, for the complete model, when we will have to make more than 60 hypercubes consistent with each other, the macro integration method is easier to generalize.

The reconciliation problem is defined as follows: We have variables $x_1^{(1)}, x_2^{(1)}$ and $x_3^{(1)}$ and $x_1^{(2)}, x_2^{(2)}, x_3^{(2)}$ and $x_4^{(2)}$. We want to find the estimates $\widehat{x}_1^{(2)}, \widehat{x}_2^{(2)}, \widehat{x}_3^{(2)}, \widehat{x}_4^{(2)}$ of $x_1^{(2)}, x_2^{(2)}$, $x_3^{(2)}$ and $x_4^{(2)}$, such that:

$$\sum_{k,l,h,i} \left(\widehat{x}_{1,k}^{(2)} \times \widehat{x}_{2,l}^{(2)} \times \widehat{x}_{3,h}^{(2)} \times \widehat{x}_{4,i}^{(2)} - x_{1,k}^{(2)} \times x_{2,l}^{(2)} \times x_{3,h}^{(2)} \times x_{4,i}^{(2)} \right)^2 \tag{A.1}$$

is minimized, under the restriction that the marginal distributions of the same variables from the sets 1 and 2 are the same:

$$(\widehat{x}_{i,1}^{(2)}, \ldots, \widehat{x}_{i,r_i}^{(2)}) = (x_{i,1}^{(1)}, \ldots, x_{i,r_i}^{(1)}), \quad \text{for} \quad i = 1, 2, 3. \tag{A.2}$$

Here we only require that the estimates $\widehat{x}_1^{(2)}, \widehat{x}_2^{(2)}, \widehat{x}_3^{(2)}, \widehat{x}_4^{(2)}$ should be as close as possible to the original values for each cell of the hypercube and the marginal distributions of the first three variables should be equal to the marginal distributions of these variables obtained from the first hypercube (register data).

We could make the set of restrictions heavier if we would add the restriction on the marginal distribution of the fourth variable to (A.2);

$$(\widehat{x}_{4,1}^{(2)}, \ldots, \widehat{x}_{4,r_4}^{(2)}) = (x_{4,1}^{(2)}, \ldots, x_{4,r_4}^{(2)}). \tag{A.3}$$

By this restriction we want to keep the marginal distribution of the variable occupation as it was observed in LFS.