Austrian Journal of Statistics
April 2014, Volume 43, 29–48.
http://www.ajs.or.at/

Macro-Integration for Solving Large Data Reconciliation Problems

Nino Mushkudiani
Statistics Netherlands

Jacco Daalmans
Statistics Netherlands

Jeroen Pannekoek
Statistics Netherlands

Abstract

Macro-integration technique is a well established method for reconciliation of large, high-dimensional tables, especially applied to macro-economic data at national statistical offices (NSO). This technique is mainly used when data obtained from different sources should be reconciled on a macro level. New areas of applications for this technique arise as new data sources become available to NSO's. Often these new data sources cannot be combined on a micro level, while macro integration could provide a solution for such problems. Yet, more research should be carried out to investigate if in such situations macro integration could indeed be applied. In this paper we propose two applications of macro-integration techniques in other domains than the traditional macro-economic applications. In particular: reconciliation of tables of a virtual census and reconciliation of monthly series of short term statistics figures with the quarterly figures of structural business statistics.

Keywords: Macro-integration, Data reconciliation, Census, Short term statistics.

1. Introduction

Macro-integration is widely used for 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, Champerowne, and Maede (1942)). Combining different data at 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 it 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 macro level. In this paper we propose new applications of macro-integration techniques in other domains than the traditional macro-economic applications.

In this paper we investigate the application of macro-integration techniques in the following areas: reconciliation of tables for the Census 2011 and reconciliation of monthly short term statistics figures with the quarterly structural business statistics figures.

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 (Denton 1971). The extended Denton method we use in this paper is defined in Bikker, Daalmans, and Mushkudiani (2013). In Section 3, we describe virtual Census 2011 data at Statistics Netherlands (SN) and the application of a macro-integration method for these data. In Section 4, we will do the same for the monthly series of the short term statistics figures. 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:

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

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:

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

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, e.g. Särndal, Swenson, and Wretman (1992). Following Boonstra (2004), we will compare in this subsection the GREG-estimator with the adjusted estimator given by Equation (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 $\boldsymbol{\Pi}$ 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

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

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 where 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 to 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 $\mathbf{1}_{n_j}$ 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

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

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 for the time series data. Suppose that our data consists of the N variables, each measured at T time points. We define these data x_{it}, $(i = 1, \ldots, N,\ t = 1, \ldots, T)$ as N time series, each of length 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

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

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

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

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

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

where I stands for the number of inequality constraints.

In Bikker *et al.* (2013) 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.* (2013) 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 linearized 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 inclusion of soft and ratio constraints in a model arises the possibility of handling macro-economic relations of data variables that were beyond the traditional linear (in)equality constraints. It opens up a possibility to 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 multi-dimensional cross-classifications (we will call these here hypercubes) about demographics and occupation. For each of these hypercubes 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 GBA (population register), 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. 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.

In different hypercubes, the same 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 correction method.

Currently at SN, the method of repeated weighting is used to combine variables from different sources and to make them consistent (Houbiers 2004). 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 margins of this table become consistent with the margins 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 mathematical optimization problems, very large problems can be handled. The software that has been built at Statistics Netherlands for the reconciliation of National Account tables is capable of dealing with a large number of variables (500 000) and restrictions (200 000). This software is built around the commercial optimization solver XPRESS.

We should emphasize that reconciliation should be applied on the macro level. First, 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 reliability of the entry should be defined. For example, an administrative source will in general have the most reliable information, and hence have a very high reliability. For the entries where no variance is available, a reliability weight can be defined using the knowledge and experience of the expert matter specialists. In our case the specialists group the data entries into different reliability classes and assign weights to each class, for a more detailed description see Bikker *et al.* (2013). During the reconciliation process, each entry of all tables will be adapted in such a way that the entries that are least reliable 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.

For step 4, we have identified a number of reconciliation problems for the census data:

I Some variables 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 the number of persons obtained from the hypercube with the variable Age with one year intervals for example from 10 to 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 set up to handle this problem.

II Before achieving consistency between all hypercubes we have to estimate each hypercube. We assume that an initial estimate for each hypercube can be made. However, this is not necessarily straightforward, especially in case of hypercubes that include variables from different data sources, for example a register and a sample. In Appendix A we will present a real data example of how one can estimate the hypercubes.

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. In Daalmans (2013) a solution for this problem is described in more detail.

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 all hypercubes should be adjusted so that the same margins are 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 following notation for census data. For $j = 1, \ldots, N$, a hypercube is defined by $H^{(j)}$. A marginal hypercube of $H^{(j)}$ will be defined by $M^{(j)}$. A variable in the hypercube $H^{(j)}$ is defined by $x_i^{(j)}$, where the subindex i denotes the variable, for example Province or Age and the super index (j) identifies the hypercube where the variable is included. 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

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

These marginal hypercubes can include some register variables. However, there is no register data available for the combination of the variables $x_i^{()}, x_k^{()}, \ldots, x_l^{()}$. On the other hand, for the marginal hypercubes that consist of a combination of variables for which register data is available, we will have the following restriction:

$$\widehat{M}^{(j_1)}_{p,q,\ldots,s} = \cdots = M^{(j_n)}_{p,q,\ldots,s}. \tag{26}$$

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 $\widehat{\mathbf{h}}^{(j)}$, $j = 1, ..., N$ 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 presented in Table 1. From these variables the only one that is observed in the survey is Occupation, the other

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

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 (population register). 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

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

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

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

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

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

$$w = \frac{16\,408\,487}{104\,674} = 156\,758.$$

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	80987	80073
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

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

II, the figures with smaller values are adjusted less and the figures with bigger values are adjusted more.

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 in the order of 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.

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

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 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.

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 $h_i^{(j)}$

Table 6: Weighted squared difference

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

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. Reconciliation of turnover figures

The second application of our macro-integration method is reconciliation of turnover figures for short term statistics (STS). Currently Statistics Netherlands is investigating the possibility of using a macro-integration method for this reconciliation. Monthly STS figures are partly based on a sample and partly on full-scale business reports. Small and middle sized businesses are included in the sample and all of the large businesses are approached. From these figures the business statistics department estimates the monthly turnover index for each sector for the Netherlands. On the other hand, we also have quarterly and yearly turnover figures of structural business statistics (SBS) based on register information. The monthly STS figures on a macro level should be consistent with the quarterly SBS figures. This condition should hold for the monthly and quarterly changes. For subject matter specialists the precise values of these figures are less important than the changes. Also, since the monthly and quarterly figures are obtained from different sources, the obvious choice is to consider the changes. In our application, quarterly figures are considered to be reliable and assumed to be fixed. In general, only those quarterly figures will be fixed that are already published.

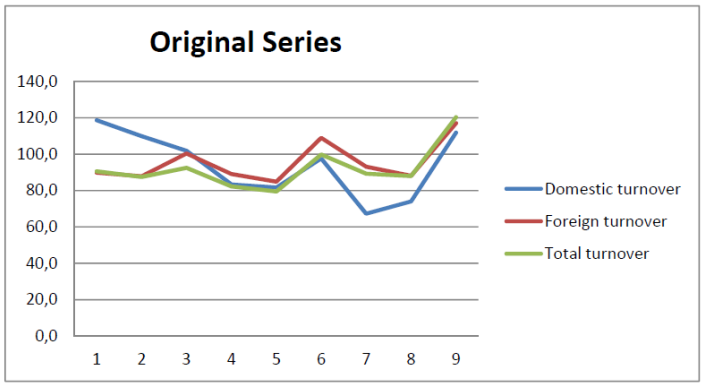

Figure 1: Monthly turnover figures for household appliances manufacture

Let us consider three STS monthly series of turnover indices for the industry "household appliances manufacture", see Figure 1. These monthly figures are: the index of total turnover $I^T_{m,i}$, the index of domestic turnover $I^D_{m,i}$ and the index of the foreign turnover $I^F_{m,i}$. We consider these figures for nine months, $(i = 1, \ldots, 9)$. For each series we have the corresponding quarterly SBS index, defined as $I^T_{q,k}$, $I^D_{q,k}$ and $I^F_{q,k}$, $(k = 1, \ldots, 3)$. These quarterly values are the benchmarks for our monthly series, since we will take these quarterly turnover indices as fixed.

On the other hand for the quarterly SBS turnover figures, subject matter specialists put the

following constraint on the three indices:

$$I^T_{q,k} = 0.375 \cdot I^D_{q,k} + 0.625 \cdot I^F_{q,k}, \quad \text{for} \quad k =, 1 \ldots, 3. \tag{33}$$

These equations reflect the relative share of the domestic and foreign turnover in total turnover in the base period.

Here we consider two different approaches for reconciliation of the STS series. In the first approach we assume that the monthly total turnover index has been adjusted pro rata already and we will apply a macro-integration method to reconcile domestic and foreign monthly indices. In the second approach we will take the original figures of the monthly total turnover index and apply a macro-integration method to reconcile the three time series (total, domestic and foreign turnover) simultaneously.

4.1. Pro rata approach

Suppose that the monthly total turnover figures $I^T_{m,i}$ are adjusted pro rata, and the pro rata estimate $\widetilde{I}^T_{m,i}$ satisfies the following constraint:

$$I^T_{q,k} = \frac{1}{3} \sum_{i=3(k-1)+1}^{3k} \widetilde{I}^T_{m,i}, \quad \text{for} \quad k = 1, \ldots, 3. \tag{34}$$

So we have original figures of monthly domestic and foreign turnover indices and pro rata adjusted figures of the total turnover indices, see Table 7. The quarterly figures are given in Table 8. We want to find the estimates $\widehat{I}^D_{m,i}$ and $\widehat{I}^F_{m,i}$ of our monthly series such that:

1. Monthly changes of domestic and foreign turnover indices are preserved as much as possible;

2. Average of monthly domestic and foreign turnover indices are equal to the corresponding quarterly turnover index;

$$I^A_{q,k} = \frac{1}{3} \sum_{i=3(k-1)+1}^{3k} \widehat{I}^A_{m,i}, \text{ for } \quad k = 1, \ldots, 3, \ A \in \{D, F\}. \tag{35}$$

3. All quarterly figures and monthly total turnover figures are fixed and monthly figures of domestic and foreign turnover indices can be adjusted;

4. For each month, the following constraints should hold:

$$\widetilde{I}^T_{m,i} = 0.375 \cdot \widehat{I}^D_{m,i} + 0.625 \cdot \widehat{I}^F_{m,i}, \quad \text{for} \quad i =, 1 \ldots, 9. \tag{36}$$

We can now specify the objective function for this problem. We assume here that the metric ς in (13) is the Euclidean metric:

$$\begin{aligned} \min_{\widehat{I}^D \widehat{I}^F} \sum_{i=2}^{9} & \frac{((\widehat{I}^D_{m,i} - \widehat{I}^D_{m,i-1}) - (I^D_{m,i} - I^D_{m,i-1}))^2}{v_D} \\ + & \frac{((\widehat{I}^F_{m,i} - \widehat{I}^F_{m,i-1}) - (I^F_{m,i} - I^F_{m,i-1}))^2}{v_F}, \end{aligned} \tag{37}$$

under the constraints defined in (35) and (36). Here v_D and v_F denote the weights of the series $\widehat{I}^D_{m,i}$ and $\widehat{I}^F_{m,i}$, respectively. In this example we have two kinds of hard constraint: within the same time period and over three time periods. We have no soft constraints. The first term in (37) will guarantee that the monthly changes $\widehat{I}^D_{m,i} - \widehat{I}^D_{m,i-1}$ is preserved as much as possible and the second term serves the same purpose for $\widehat{I}^F_{m,i}$ series.

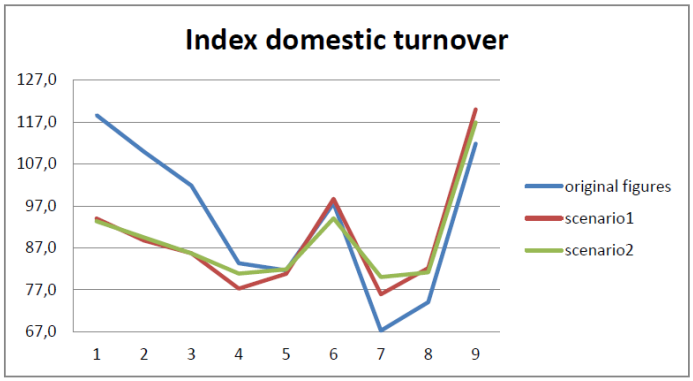

Figure 2: Original and reconciled domestic turnover figures

Table 7: Monthly turnover indices of industry "household appliances manufacture"

Month	Domestic turnover	Foreign turnover	Total turnover
1	118.64	89.84	90.6
2	109.92	87.85	87.5
3	101.85	100.34	92.5
4	83.35	89.08	82.3
5	81.65	84.90	79.5
6	97.63	108.97	99.8
7	67.26	93.10	89.3
8	74.06	88.07	88.0
9	111.83	117.04	120.3

Table 8: Quarterly turnover figures

Quarter	Domestic turnover	Foreign turnover	Total turnover
1	89.5	90.62	90.2
2	85.6	88.16	87.2
3	92.7	103.10	99.2

We consider two different pairs of weights for the monthly series. Accordingly, we have two different scenarios for the data integration problem. At first we assume that both series have the same weights equal to 1. In the second scenario we assume that the weights for the domestic turnover series is equal to 1 and the weight of the foreign turnover is equal to 0.1.

Using the statistical software package R we programmed an iterative algorithm described in e.g. De Waal, Pannekoek, and Scholtus (2011), Ch. 10 to solve the lineair optimization problem defined in (35)-(37), with the weights $v_D = v_F = 1$ for scenario 1 and $v_D = 1$ and $v_F = 0.1$ for scenario 2. To illustrate the preservation of changes we present the original and the reconciled series separately for domestic and foreign turnover in Figures 2 and 3. Observe that in scenario 1 both time series are equally reliable. However, in scenario 2 we assume that the foreign turnover figures are more reliable than the domestic turnover. As a result, the reconciled foreign turnover figures in scenario 2 have much better preserved monthly changes than the domestic turnover figures. Using the weights we can include extra information in the model. If in our example we know that the source for one series are more reliable than the other series, we can include this information in the model by adapting the weights.

4.2. Macro-integration approach

If instead of the pro rata adjusted series we consider the original figures for the total turnover and include these series in the objective function as well, we will obtain a new integration

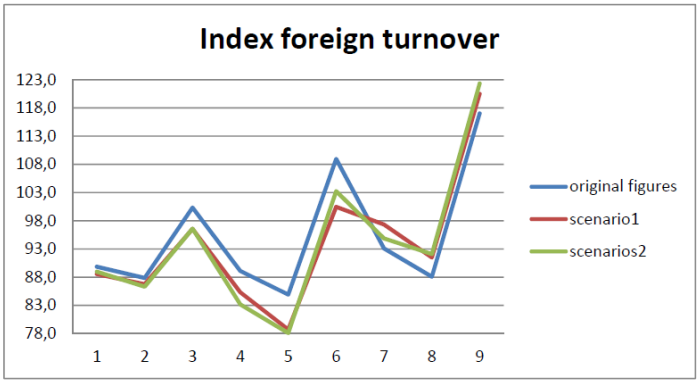

Figure 3: Original and reconciled foreign turnover figures

problem. In this case we will have three time series to benchmark. We want to find the estimates of the series $I^T_{m,i}$, $I^D_{m,i}$ and $I^F_{m,i}$ such that:

$$\begin{aligned}\min_{\widehat{I}^T\widehat{I}^D\widehat{I}^F} \sum_{i=2}^{9} & \frac{((\widehat{I}^T_{m,i} - \widehat{I}^T_{m,i-1}) - (I^T_{m,i} - I^T_{m,i-1}))^2}{v_T} \\ + & \frac{((\widehat{I}^D_{m,i} - \widehat{I}^D_{m,i-1}) - (I^D_{m,i} - I^D_{m,i-1}))^2}{v_D} \\ + & \frac{((\widehat{I}^F_{m,i} - \widehat{I}^F_{m,i-1}) - (I^F_{m,i} - I^F_{m,i-1}))^2}{v_F}.\end{aligned} \tag{38}$$

Here v_T denotes the weight of the series $I^T_{m,i}$. In the previous subsection we first adjusted the monthly total turnover figures pro rata. In the optimization problem (35)-(37) these figures were fixed. In this example we do not adjust the total turnover figures beforehand, we want to reconcile these simultaneously with the other figures. Therefore for this problem, (36) should change into constraints that include the estimates of the total turnover indices:

$$\widehat{I}^T_{m,i} = 0.375 \cdot \widehat{I}^D_{m,i} + 0.625 \cdot \widehat{I}^F_{m,i}, \quad \text{for} \quad i =, 1 \ldots, 9. \tag{39}$$

In addition, constraints in (35) should now also hold for $\widehat{I}^T_{m,i}$, the estimates of the total turnover indices:

$$I^A_{q,k} = \frac{1}{3} \sum_{i=3(k-1)+1}^{3k} \widehat{I}^A_{m,i}, \quad \text{for} \quad k = 1, \ldots, 3, \;\; A \in \{T, D, F\}. \tag{40}$$

For the macro-integration problem in (38)-(40) we defined two different scenarios according to the weights of the series. In the first scenario we consider the following weights:

$$v_T = 0.1 \quad \text{and} \quad v_D = v_F = 1.$$

And for the second scenario:

$$v_T = v_F = 0.1 \quad \text{and} \quad v_D = 1.$$

The estimates for these two scenarios were almost identical, see for example $\widehat{I}^{T1}$ and $\widehat{I}^{T2}$ in Table 9. It seems that the optimal estimates were found and the weight did not make much of a difference.

Remark In Figure 4 we compare the original figures of the total turnover with the adjusted figures from two different approaches described above. Adjusted figures are according to the macro-integration method, scenario 1 and the pro rata method. We can observe that

Table 9: Reconciled figures of total, domestic and foreign turnover for scenarios 1 and 2.

Month	$\widehat{I}^{T1}$	$\widehat{I}^{T2}$	$\widehat{I}^{D1}$	$\widehat{I}^{D2}$	$\widehat{I}^{F1}$	$\widehat{I}^{F2}$
1	90.43	90.59	93.89	93.30	88.36	88.96
2	87.12	87.09	88.49	88.63	86.30	86.16
3	93.05	92.93	86.12	86.57	97.20	96.74
4	79.35	79.43	75.20	74.88	81.84	82.16
5	78.41	78.49	80.04	79.77	77.44	77.72
6	103.84	103.68	101.56	102.15	105.20	104.60
7	86.81	86.69	74.17	74.60	94.39	93.95
8	89.18	89.07	83.02	83.41	92.87	92.47
9	121.61	121.83	120.91	120.09	122.04	122.88

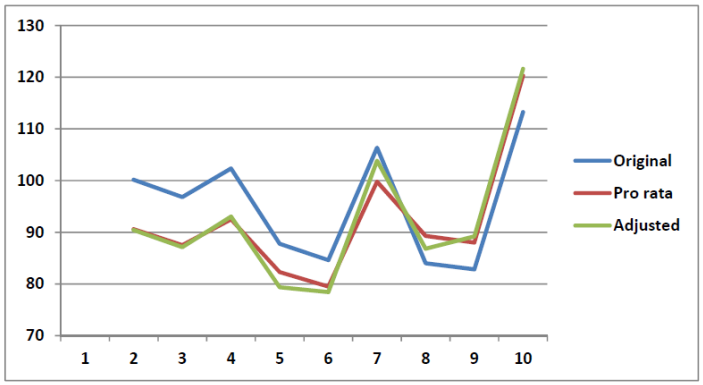

Figure 4: Original and adjusted monthly total turnover figures

the estimate obtained by the macro-integration method follows the monthly changes of the original time series better than the pro rata adjusted estimate, even though the difference between these estimates is minor. From the two methods described above, we would suggest to use the full macro-integration method. It has several advantages:

1. The estimated time series of the total turnover follow the monthly changes of the original series;

2. The original figures of the total turnover do not have to be adjusted beforehand, implying that the integration process incudes one step less.

3. The choice of the weights could become less important and this may lead to better estimates.

This example illustrates the use of a macro-integration method for time series STS data. SN is currently carrying out research on how to apply macro-integration of STS figures in the production process.

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 seems to be the only technique that is effective. 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 is very important, since the application of the repeated weighting method at SN is currently already hampered by its limitations. For this reason SN is using a combination of the weighting method and the macro-integration method. We feel that in the future macro-integration could be the only method used to ensure consistency of tables on macro level.

The second application is equally, if not more, important for SN. For the past couple of years, SN has had an additional data source for business statistics figures. The use of register data has increased considerably over the last years. Also the quality of data has improved, and through intensive communication between SN and the registers, our knowledge of the register variables has increased. At the same time, SN has taken measures to improve the quality of the surveys. Improving the quality of the monthly and quarterly data creates the possibility for reconciliation of the survey data with the register data. At this moment SN is taking steps to implement this reconciliation.

References

Bikker R, Daalmans J, Mushkudiani N (2013). "Benchmarking large accounting frameworks: a generalized multivariate model." *Economic Systems Research*.

Boonstra H (2004). "Calibration of tables of estimates." *Technical report*, Statistics Netherlands.

Daalmans J (2013). "A new micro-macro method for estimating Dutch census tables." *Presented at the Conference of European Statisticians, Group of Experts on Population and Housing Censuses. Geneva*.

De Waal T, Pannekoek J, Scholtus S (2011). *Handbook of statistical data editing and imputation*. Wiley handbooks in survey methodology. John Wiley & Sons.

Denton F (1971). "Adjustent of Monthly or Quarterly Series to Annual Totals: An Approach based Quadratic Minimization." *Journal of the American Statistical Association*, **66**, 99–102.

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

Särndal CE, Swenson B, Wretman J (1992). *Model assisted Survey Sampling*. Springer-Verlag, New York.

Stone J, Champerowne D, Maede J (1942). "The Precision of National Income Accounting Estimates." *The Review of Economic Studies*, **9**, 111–125.

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 (population register) 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 crosstable of variables, containing

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

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:

$$\sum_{k,l,h,i} \left(\widehat{x}_{1,k}^{(2)} \times \widehat{x}_{2,l}^{(2)} \times \widehat{x}_{3,h}^{(2)} \times \widehat{x}_{4,i}^{(2)} - x_{1,k}^{(2)} \times x_{2,l}^{(2)} \times x_{3,h}^{(2)} \times x_{4,i}^{(2)} \right)^2 \tag{41}$$

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

$$(\widehat{x}_{i,1}^{(2)}, \ldots, \widehat{x}_{i,r_i}^{(2)}) = (x_{i,1}^{(1)}, \ldots, x_{i,r_i}^{(1)}), \quad \text{for} \quad i = 1, 2, 3. \tag{42}$$

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 (42);

$$(\widehat{x}_{4,1}^{(2)}, \ldots, \widehat{x}_{4,r_4}^{(2)}) = (x_{4,1}^{(2)}, \ldots, x_{4,r_4}^{(2)}). \tag{43}$$

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

Affiliation:

Nino Mushkudiani

Department of Methodology
Statistics Netherlands
The Hague, The Netherlands
E-mail: n.mushkudiani@cbs.nl

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 43
April 2014

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2013-12-02
Accepted: 2014-03-03