Statistical Journal of the IAOS 31 (2015) 257–262
DOI 10.3233/SJI-150898
IOS Press

Reconciliation of labour market statistics using macro-integration

Nino Mushkudiani, Jacco Daalmans and Jeroen Pannekoek*
Department of Methodology, Statistics Netherlands, The Netherlands

Abstract. Macro-integration techniques are 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 to become very versatile techniques for integration of data from different sources at the macro level. Applications in other domains than macro-economics seem promising. In this paper we present an application to labour market data from two sources, an administrative one and a survey, with slightly different definitions and different frequencies of reporting (monthly, quarterly). The purpose is to combine these estimates to form a single monthly estimate. Depending on the specification of the macro-integration model several alternatives for obtaining such estimates are derived.

Keywords: Multiple sources, data integration, reconciliation of estimates, labour force survey, public employment service 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 supply and use tables (SUT) to new margins (see, e.g. [7]). 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.

At Statistics Netherlands (SN) formal macro-integration methods are applied more and more often during the last decade at Statistics Netherlands. One of the first large scale application is the reconciliation of quarterly and annual figures of supply and use tables at the National Accounts department. These tables contain over 500 thousand estimated figures which are the variables of the reconciliation problem. A new generalized multivariate Denton method was defined specifically for this application. The method replaced the informal methods that were used before. To include all of the numerous relations and constraints, the SUT model includes reliability weights and hard, soft, inequality and ratio constraints. The model is applied in a time series setting where it is important to preserve movements as well as possible. These movements can be defined as additive first-order differences or as proportional first-order differences in the sense that in a multivariate setting the different time series can be reconciled using either the additive or the proportional method (see [1]).

Another application of macro-integration at SN is the integration at the macro level of business statistics [2] introduced a method for the reconciliation of trade and transport statistics. Currently, there are also plans to reconcile monthly, survey-based production statistics with quarterly statistics that are based on tax office data. Finally, the macro-integration method is also used to solve some remaining consistency problems arising in the estimation (by repeated "calibration" weighting) of the many tables of the population and housing census (see [3,6]).

In this paper we investigate another application of macro-integration techniques, namely for labour mar-

[*] Corresponding author: Jeroen Pannekoek, Department of Methodology, Statistics Netherlands, The Netherlands. E-mail: j.pannekoek@cbs.nl.

ket statistics. The variables that form the labour market statistics are (estimates of) totals of the labour force, non labour force, unemployed labour force, employees, jobs, vacancies and social benefits. This is a complex reconciliation problem, mainly caused by the variety of data sources, that contain labour market variables with almost equal, but not exactly the same definitions. Moreover, data collected from these sources have different frequencies and different population coverage. In the example presented in this paper, we illustrate the methods by using one variable, unemployment, and reconcile the figures from two different sources: the Dutch labour force survey (LFS) and the Public Employment Service (PES) register data. The register data are usually updated on a quarterly basis. The survey (LFS) is based on a rotating panel design producing monthly figures. Our goal is to combine these data in order to produce one set of figures.

The paper is organized as follows: in Section 2 we describe our example data. In Section 3, we define the macro-integration model for the example. The results of the reconciliation process are presented in Section 4, and in Section 5 some conclusions are summarized.

2. The reconciliation problem for labour market estimates

Let us consider a simple example of two different sources. Suppose we have a labour force population of 60000 persons and we want to conduct a monthly labour force survey to estimate 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. For convenience we assume that the total number of respondents in the survey we want to conduct is 6000.

In the survey we observe two variables: whether a person has a job and if not whether she/he is registered at the PES. Suppose for simplicity that we do not have nonresponse and we consider results of the survey during the three months: January, February and March. From these figures we can estimate the number of unemployed persons in each group of the population (by multiplying the survey figures by 10). Denote these population figures by x_{ijt}, $t = 1, 2, 3$, $i = 1, \ldots, 4$ and $j = 1, 2$. Here t stands for the month, i and j denote the entries of the matrix Age×Sex, see Table 1. In parenthesis are the numbers of persons registered at the PES, denoted by y_{ijt}. Observe that there are less persons registered at the PES than the unemployed persons. This is due to the fact that part of the unemployment labour force do not register themselves at the PES, as they are not eligible for the social unemployment benefits.

On the other hand we have the PES register data. From these register data we can derive the number of persons that were registered as unemployed labour force at the end of each quarter. Denote these by R_{ijk}, where, as above, i and j denote the entries of the matrix Age×Sex and k defines the index for the quarter, see Table 2. Suppose that we do not have timeliness issues for the survey and register and both data are available for us at around the same time. In the ideal case the values of $y_{ij\text{ March}}$ and $R_{ij\text{March}}$ should be the same. However this is not the case. For example in March there were 330 women of age $20 - 29$ registered at PES according to the survey, $y_{113} = 330$, and 350 according to the PES register data, $R_{111} = 350$. Here the register data will be considered highly reliable and hence the figures R_{ijk} are fixed.

The reconciliation problem now is to find estimates $\widehat{x}_{ijt}$ and $\widehat{y}_{ijt}$ for x_{ijt} and y_{ijt} that satisfy constraints defined by the features of our data sources, which can be summarized as:

1. The figures from the survey have a certain variance, while we assumed that the PES register data are fixed;
2. The numbers of persons registered at the PES in March according to the survey and the numbers in the PES register should be the same ($\widehat{y}_{ij\text{ March}} = R_{ij\text{March}}$). This constraint is justified when the two sources have the same population coverage, have no timeliness problems and the definitions of the variable "register unemployed person", are the same.
3. Next, we want to preserve monthly changes of the series x_{ijt} and y_{ijt}, since we know that the absolute values of these series are much less reliable.
4. We want to preserve the ratios between the unemployment numbers x_{ijt} and the numbers of persons registered at the PES, y_{ijt}. For instance, for women of age 20–29 this ratio is $37/33$ in March. These ratios do not have to hold exactly, as they are observed in the sample, again we assume that these ratios are more reliable, than the absolute values of our series.

In the next section we consider the macro-integration model for this problem.

Table 1
Weighted unemployment data, $x_{ijt}(y_{ijt})$

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20–29	350 (250)	340 (270)	360 (250)	350 (290)	370 (330)	330 (300)
30–39	400 (380)	350 (320)	420 (370)	360 (320)	420 (350)	370 (350)
40–49	600 (500)	560 (500)	580 (510)	560 (490)	610 (580)	580 (550)
⩾ 50	420 (300)	380 (250)	420 (310)	400 (310)	430 (350)	400 (380)

Table 2
PES register data at the end of the first quarter, R_{ijk}

Age	Sex	
	Woman	Man
20–29	350	330
30–39	390	360
40–49	600	570
⩾ 50	370	395

3. Reconciliation model

The macro-integration approach to the reconciliation problem from the previous section is to view it as a constrained optimization problem. Several forms of the objective function for this optimization problem are commonly used in the literature. One of the widely applied objective functions is the Denton proportional first difference (PFD) function (see [4]). Using the notations from the previous section the PFD function, for (x_{ij}) time series, in a multivariate setup, is defined as follows:

$$\min_{x_{ijt}} \sum_{t=2}^{T} \sum_{ij} \left(\frac{\widehat{x}_{ijt}}{x_{ijt}} - \frac{\widehat{x}_{ijt-1}}{x_{ijt-1}} \right)^2 . \tag{1}$$

In this function the ratio between the estimated and the original figures will be preserved. When the growth rate of the time series should be preserved another objective function, the growth rate preservation (GRP) function, is used:

$$\min_{x_{ijt}} \sum_{t=2}^{T} \sum_{ij} \left(\frac{\widehat{x}_{ijt}}{\widehat{x}_{ijt-1}} - \frac{x_{ijt}}{x_{ijt-1}} \right)^2 . \tag{2}$$

Reconciliation based on GRP function is also extensively studied in the literature, see e.g. [5]. The PFD and GRP functions are closely related to each other:

$$\sum_{t=2}^{T} \sum_{ij} \left(\frac{\widehat{x}_{ijt}}{x_{ijt}} - \frac{\widehat{x}_{ijt-1}}{x_{ijt-1}} \right)^2 = \\ \sum_{t=2}^{T} \sum_{ij} \left[\frac{\widehat{x}_{ijt-1}}{x_{ijt}} \left(\frac{\widehat{x}_{ijt}}{\widehat{x}_{ijt-1}} - \frac{x_{ijt}}{x_{ijt-1}} \right) \right]^2 . \tag{3}$$

The other version of PFD is when the additive difference should be preserved instead of the proportional difference. This function is called the additive first difference (AFD) function. In the optimization problem defined by the constraints 1–4 in Section 2, we will use the additive first difference function. The main reason for using AFD is that constraint 3 states that monthly additive changes for our estimates $\widehat{x}_{ijt}$ and $\widehat{y}_{ijt}$ should be as close as possible (in some sense) to the monthly changes of their initial values x_{ijt} and y_{ijt}. Hence we want to find $\widehat{x}_{ijt}$ and $\widehat{y}_{ijt}$, such that the function

$$\sum_{t=2}^{3} \sum_{ij} \left[\frac{1}{v_{xij}} ((\widehat{x}_{ijt} - \widehat{x}_{ijt-1}) - (x_{ijt} - x_{ijt-1}))^2 \right. \\ \left. + \frac{1}{v_{yij}} ((\widehat{y}_{ijt} - \widehat{y}_{ijt-1}) - (y_{ijt} - y_{ijt-1}))^2 \right] \tag{4}$$

reaches its minimum. Here v_{xij} denotes the reliability measure of the series $(x_{ijt} - x_{ijt-1})$ and v_{yij} the reliability measure of $(y_{ijt} - y_{ijt-1})$. In the literature these reliability measures are often called variances. In our practical application we do not know the variance of our time series. Often we have just one realization. Therefore we want to avoid defining the random model, within a strict, formal frame. Instead the weights will serve as a reliability measure in our application. We assume that the weights for both series are the same for each time period and are proportional to the values of the series.

The first constraint states that we want to find the estimates $\widehat{x}_{ijt}$ and $\widehat{y}_{ijt}$ for x_{ijt} and y_{ijt} and that the values R_{ijk} are fixed. While from the second constraint follows that the estimates $\widehat{y}_{ijt}$ for y_{ijt} from the last month of the quarter should exactly be equal to R_{ijk}:

$$\widehat{y}_{ijt} = R_{ijk}, \text{for all } i, j \text{ and } t = 3, k = 1. \tag{5}$$

Finally, in constraint 4 we state that the ratios x_{ijt}/y_{ijt} should be preserved as much as possible. We ensure this by defining the soft ratio constraints

$$\widehat{x}_{ijt}/\widehat{y}_{ijt} \sim d_{ijt}, \text{ for all } i, j, t, \tag{6}$$

where $d_{ijt} = x_{ijt}/y_{ijt}$. Here the sign $\sim$ stands for approximation. We want the ratio $\widehat{x}_{ijt}/\widehat{y}_{ijt}$ to be close to

Table 3
Reconciled unemployment data $\widehat{x}_{ijt}(\widehat{y}_{ijt})$ for equal weights

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20–29	375.1 (268.0)	375.2 (298.3)	385.0 (268.2)	393.7 (319.0)	393.7 (350.0)	363.8 (330.0)
30–39	445.2 (422.0)	360.8 (329.9)	466.3 (410.9)	467.1 (329.8)	467.1 (390.0)	380.7 (360.0)
40–49	622.4 (519.0)	581.9 (519.5)	602.0 (529.4)	631.5 (509.5)	631.5 (600.0)	601.5 (570.0)
$\geqslant 50$	446.0 (318.8)	398.3 (262.5)	445.7 (329.2)	419.2 (323.7)	455.2 (370.0)	416.7 (395.0)

value d_{ijt} and the weight v_{dij} for the ratio x_{ijt}/y_{ijt} will define how close. For example if $v_{dij} = 0$ we will have that $\widehat{x}_{ijt}/\widehat{y}_{ijt} = d_{ijt}$ and the constraint in (6) will become a hard constraint. To include the soft ratio constraints into our objective function we first linearize (6):

$$\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt} \sim 0. \tag{7}$$

The weight of this linearized ratio v^*_{dij}, can be derived from v_{dij}, x_{ijt} and y_{ijt}, see [1]. Soft linearized ratios can be incorporated in the model by adding the following term to the objective function

$$+\sum_{t=1}^{T}\sum_{ij}\frac{1}{v^*_{dij}}(\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt})^2. \tag{8}$$

Now we can write out the objective function for our example: we want to find $\widehat{x}_{ijt}$ and $\widehat{y}_{ijt}$, such that

$$\min_{\widehat{y}\widehat{x}}\sum_{t=2}^{3}\sum_{ij}\left[\frac{1}{v_{xij}}((\widehat{x}_{ijt} - \widehat{x}_{ijt-1}) -(x_{ijt} - x_{ijt-1}))^2 + \frac{1}{v_{yij}}((\widehat{y}_{ijt} - \widehat{y}_{ijt-1}) -(y_{ijt} - y_{ijt-1}))^2 + \frac{1}{v^*_{dij}}(\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt})^2\right], \tag{9}$$

and

$$\widehat{y}_{ijt} = R_{ijk}, \text{ for all } i,j \text{ and } t = 3, k = 1. \tag{10}$$

Note that the quarterly unemployment numbers, R_{ijt}, are included in the model as parameters. These are not specified as free variables, because these figures are fixed (following the first constraint). Note as well that we can include the hard constraint (10) into the objective function in (9) by replacing $\widehat{y}_{ij3}$ by R_{ij1}. The constraint in (10) is hard because we have assumed that the PES register figures have variance equal to zero. Note that, in practice this might not be the case. Then the hard constraint in (10) will become a soft constraint:

$$\widehat{y}_{ijt} \sim R_{ijk}, \tag{11}$$

with some reliability weight for the PES register figures.

Table 4
Women of 20–29 years; ratio's for equal weights

	January	February	March
Initial ratio	1.400	1.440	1.121
Reconciled ratio	1.400	1.436	1.125

The model defined here is quite simple. An extended model which includes proportional and additive functions, hard and soft and equality and inequality constraints, reliability weights and allows some missingness in data is described in [1].

In the next section we solve the optimization problems (9)–(10) for the figures given in Tables 1 and 2.

4. Results

In order to solve the problem given by (9)–(10), we first need to define the weights $v_{xij}, v_{yij}, v^*_{dij}$. We assume that all weights $v_{xij}, v_{yij}, v^*_{dij}$ are equal to 300. Table 3 contains the estimated figures of $\widehat{x}_{ijt}$ and $\widehat{y}_{ijt}$. These figures show that the number of persons registered at PES (the numbers between parenthesis) in March are indeed consistent with the PES register figures (as depicted in Table 2).

To illustrate the preservation of changes and ratios d_{ijt}, we focus 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 4 it can be seen that the same holds true for the ratios d_{ijt}. Figure 1 also shows that the data reconciliation increases both the number of persons registered at PES and the number of unemployed people at each time period. The explanation is that, the number of persons registered at PES in the survey in month 3 is much smaller than the corresponding register figure of the first quarter. Since the survey figures have to match the register figures exactly and since all monthly changes of the survey figures should be preserved as much as possible, all monthly survey figures on the number of persons registered at the PES are increased. The same occurs to the number of unemployed persons, which can be

Table 5
Reconciled unemployment data $\widehat{x}_{ijt}(\widehat{y}_{ijt})$ for unequal weights

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20–29	376.1 (269.3)	375.52 (299.5)	385.9 (269.5)	385.34 (319.7)	395.6 (350.0)	365.1 (330.0)
30–39	445.2 (420.6)	360.90 (330.0)	465.5 (410.3)	370.90 (330.0)	465.7 (390.0)	380.9 (360.0)
40–49	622.5 (519.7)	582.07 (519.9)	602.4 (529.8)	582.05 (509.9)	632.3 (600.0)	602.0 (570.0)
$\geqslant 50$	446.4 (319.6)	398.97 (264.1)	446.3 (329.8)	418.69 (324.6)	456.1 (370.0)	418.4 (395.0)

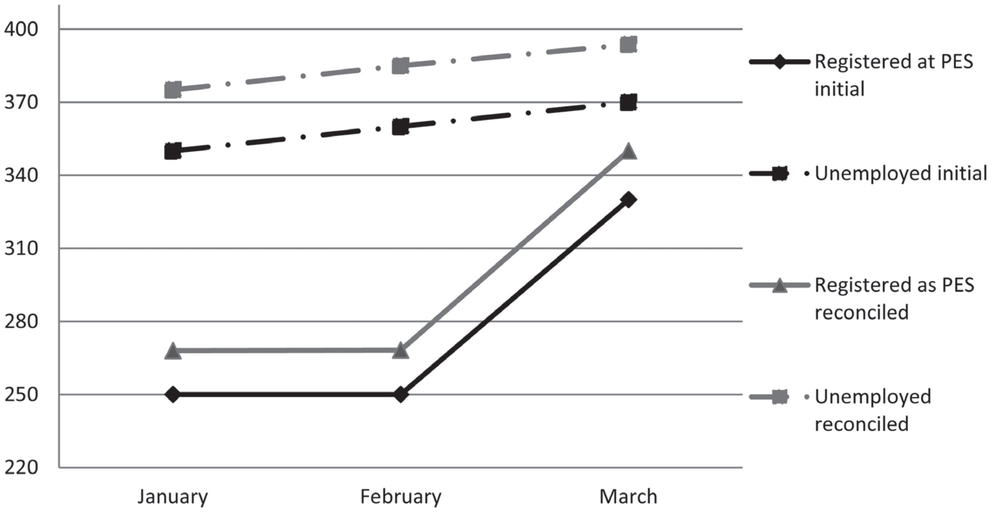

Fig. 1. Women of 20–29 years; initial and reconciled figures.

explained from the preservation of the ratios between the number of unemployed and the number of persons registered at PES at each time period. In Table 4 reconciliation data gives the ratio estimates for this scenario.

Now, suppose that we decrease the weight of the monthly changes from 300 to 100, but we do not change the weight of the ratios between unemployed and registered unemployed persons. As a result, the initial quarterly changes are preserved better at the expense of the ratio between the number of unemployed and persons registered at PES. The reconciled ratio figures for this scenario are 1.396, 1.432 and 1.130, respectively for January, February and March. The results from this reconciliation problem are presented in Table 5.

This is of course only a simple example where we have two data sources and did not take into account many issues. As we already mentioned above these issues could be: more data sources, different population coverages, different variable definitions, timeliness problems and reliability of data. Resolving of these issues will be impossible without the thorough investigation of the data structures.

5. Conclusions

The generalized multivariate Denton model used in this paper for the labour data example, was developed for the reconciliation of Dutch supply and use tables (see [1]). This model has many attractive properties for the practical applications; The model can combine proportional and additive methods, can handle applications of extremely large multivariate data sets and can include a wide range of relationships, by using hard and soft constraints, equality and inequality constraints and reliability weights; The method is implemented in the production process at Statistics Netherlands, and made reconciliation of SUT more transparant and efficient.

On a simple example of labour force data we tried to illustrate that the use of the multivariate Denton method can be effective also for other applications. Currently for labour market variables often micro-integration techniques are used. In certain cases macro-integration will have advantages over micro-integration especially when data from multiple sources of different structures, should be combined. By aggregating data, the number of figures that have to be made consistent decreases and as a consequence a smaller reconciliation problem can be obtained. Anomalies in the micro data may cancel out on a macro level and data linkage problems that occur at micro level will be avoided. Also the correction for differences in variable definitions, population coverage and reporting periods may be more easily achieved at the macro level than at the micro level. Macro-integration technique is currently used for reconciliation of population Census high dimensional tables at SN. For this application the macro-integration method is combined with the repeated weighting method. The reconciliation of Census data is an example where the micro-integration technique did not satisfy the requirements on the outcome and had to be combined with a macro-integration.

For the past couple of years SN have better access to a register data for labour market variables. Since the use of the register data has increased enormously over last years, as did the quality of data and the understanding of the variables. At the same time, SN has taken means to improve the quality of the surveys, such as the labour force survey. Improving the quality and understanding of register data creates the possibility for reconciliation of the survey data with the register data. Future research could show if, and how, it is possible to combine these sources in order to produce one set of figures.

When data are very large and many sources should be combined macro-integration could be the only technique that can be used. The research on applications of macro-integration methods is therefore of great importance.

References

[1] R. Bikker, J. Daalmans and N. Mushkudiani, Benchmarking large accounting frameworks: A generalized multivariate model, *J Economic Systems Research*, 2013.

[2] H.J. Boonstra, C. de Blois and G. Linders, Macro-integration with inequality constraints: An application to the integration of transport and trade statistics, *Statistica Neerlandica* **65** (2011), 407–431.

[3] J. Daalmans, 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*, 2013.

[4] F. Denton, Adjustent of monthly or quarterly series to annual totals: An approach based quadratic minimization, *Journal of the American Statistical Association* **66** (1971), 99–102.

[5] T. Di Fonzo and M. Marini, Benchmarking time series according to a growth rates preservation principle, *Journal of economic and social measurement* **37** (2012), 225–252.

[6] N. Mushkudiani, J. Daalmans and J. Pannekoek, Macro-integration techniques with applications to census tables and labour market statistics, Discussion paper, Statistics Netherlands, 2012.

[7] J. Stone, D. Champerowne and J. Maede, The precision of national income accounting estimates, *Reviews of Economic Studies* **9** (1942), 111–125.