# Macro-integration techniques to reconcile labour market statistics from different sources

Nino Mushkudiani [1], Jacco Daalmans [1] and Jeroen Pannekoek [1,2]

[1]Department of Methodology, Statistics Netherlands, The Netherlands
[2]Corresponding author; Jeroen Pannekoek, e-mail: j.pannekoek@cbs.nl

**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 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 and their properties will be discussed.

Keywords: Multiple sources, data integration, reconciliation of estimates, labour force survey, administrative 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.

Formal macro-integration methods are applied increasingly often at Statistics Netherlands. The first application was for matching quarterly and annual National Accounts. A new multivariate Denton method (see Bikker et al. (2010)) replaced the informal methods that were used before. Integration at the macro level is also applied to business statistics: Boonstra et al. (2010) 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 data. Finally, macro integration is used for estimating the population and housing census (see Mushkudiani et al. (2012)).

In this paper we investigate another application for the macro-integration techniques, namely the labour market statistics: labour force, non labour force, unemployed labour force, employees, jobs, vacancies, 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. Data collected from these sources have different frequency and different population coverage. In the example presented in this paper, we only consider one variable, unemployment and reconcile the variables from only two different sources: the Dutch labour Force Survey (LFS) and the Unemployment Office Register (UOR). The register data are usually updated on a quarterly basis and the survey (LFS) is a rotating panel design producing monthly figures. Our goal is to combine these data in order to produce a single set of figures.

The paper is organized as follows: in Section 2 we introduce the reconciliation problem by an example; In Section 3, we define the macro-integration model for our example, and in Section 4 we present the results of this example. Some conclusions are summarized in Section 5.

## 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 a monthly labour force survey is conducted to estimate the unemployment rate by Age and Sex. The population totals for the Age by Sex combinations are known. For convenience we assume that the total number of respondents in the equal probability sample 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 Unemployment Office (UO). Suppose for simplicity that we do not have nonresponse and we consider results of the survey during 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 $x_{ijt}$, here $t$ stands for the month, $i$ and $j$ denote the entries of the matrix Age$\times$Sex, see Table 1. In parenthesis are the numbers

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

| 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) |

of persons registered at the UO, that are denoted by $y_{ijt}$. Observe that there are less persons registered at the UO than the unemployed persons. This is due to the fact that part of the unemployed persons do not register themselves at the UO as they are not eligible for unemployment benefits.

On the other hand we have the UOR. From this register we can derive the number of persons that were registered as unemployed and belong to the 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$\times$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 at around the same time. In the ideal case the values of $y_{ij\,\text{March}}$ are 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 -<30 registered at UO according to the survey, $y_{113} = 330$, and 350 according to the UOR, $R_{111} = 350$. Thus, there is inconsistency between the data from the survey and the UOR. Our reconciliation task is to find new estimates $\widehat{x_{ijt}}$ and $\widehat{y_{ijt}}$ for $x_{ijt}$ and $y_{ijt}$, that are consistent with the UOR data displayed in table 2 and reflect as well as possible the information available in the survey data displayed in table 1. For this purpose we

Table 2: UOR data at the end of the first quarter, $R_{ij1}$.

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

assume that the register data are highly reliable and are assigned a variance equal to zero. This implies that the figures $R_{ijk}$ are fixed. The survey data on the other hand have non-zero variances and the final estimates may deviate to some extend from the estimates displayed in table 1.

To accomplish this reconciliation task we consider the following constraints that the final estimates of the unemployment figures should satisfy, and that reflect the features of our data sources.

1. The figures from the survey have a non-zero variance, while we assumed that the UOR data are fixed;

2. The numbers of persons registered at the UO in March according to survey and the numbers in the UOR should be the same ($\widehat{y}_{ij\,\text{March}} = R_{ij\text{March}}$). This constraint will be relaxed later on (see, (8)).

3. Next, we want to preserve monthly changes of $x_{ijt}$ and $y_{ijt}$ series, since we know that the estimates of changes are much more reliable than the absolute levels of $x_{ijt}$ and $y_{ijt}$. These monthly changes do not have to be preserved exactly since they are estimated from a sample.

4. We want to preserve the ratios between the unemployment numbers $x_{ijt}$ and the numbers of persons registered at the UO, $y_{ijt}$. For instance, for women of age $20-29$ this ratio is $37/33$ in March. These ratios are assumed to be more reliable than the absolute values of our series. Again, these ratios do not have to hold exactly for the final estimates.

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

## 3. Reconciliation model

The macro-integration approach to the reconciliation problem from the previous section is to view it as a constrained optimization problem. Here we define the objective function for the optimization problem satisfying constraints 1-4 defined in Section 2.

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

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

Next, constraint 3 states that the monthly changes of the 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 $y_{ijt}$. For simplicity here we consider an Euclidean metric; hence we want to find

$\widehat{x}_{ijt}$ and $\widehat{y}_{ijt}$ such that the objective function

$$\sum_{t=2}^{T}\sum_{ij} \quad \frac{((\widehat{x}_{ijt}-\widehat{x}_{ijt-1})-(x_{ijt}-x_{ijt-1}))^2}{v_{xij}} + \frac{((\widehat{y}_{ijt}-\widehat{y}_{ijt-1})-(y_{ijt}-y_{ijt-1}))^2}{v_{yij}}, \tag{2}$$

reaches its minimum. Here $v_{xij}$ denotes the variance of $x_{ijt}$ and $v_{yij}$ the variance of $y_{ijt}$. We assume that the variance for both series is the same for each time period. Notice that the variances appear as weights in the objective function; larger weights imply that the final estimates may deviate more from the original estimates. Alternative weights could also be chosen. For instance weights proportional to the values of the series, implying that larger values will deviate more from the original estimates than smaller ones.

At last, in constraint 4 we state that the ratios $x_{ijt}/y_{ijt}$ should be preserved as much as possible. Hence we want that

$$\widehat{x}_{ijt}/\widehat{y}_{ijt} \sim d_{ijt}, \tag{3}$$

where $d_{ijt} = x_{ijt}/y_{ijt}$. As above, we define the weight $v_{dij}$ for the ratio $x_{ijt}/y_{ijt}$. This weight determines how much the ratio $\widehat{x}_{ijt}/\widehat{y}_{ijt}$ may deviate from $d_{ijt}$. For example if $v_{dij} = 0$ we will have that $\widehat{x}_{ijt}/\widehat{y}_{ijt} = d_{ijt}$ and this constraint will become a hard constraint. We call constraints in (3) soft ratio constraints. To include the soft ratio constraints into our objective function we first linearize (3):

$$\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt} \sim 0. \tag{4}$$

The weight of this constraint will be different, denote it by $v^*_{dij}$. Soft linearized ratios can be incorporated in the model by adding the following term to the objective function

$$+\sum_{t=1}^{T}\sum_{ij}\frac{(\widehat{x}_{ijt}-d_{ijt}\widehat{y}_{ijt})^2}{v^*_{dij}}. \tag{5}$$

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

$$\min_{\widehat{y}\widehat{x}}\sum_{t=2}^{T}\sum_{ij} \quad \frac{((\widehat{x}_{ijt}-\widehat{x}_{ijt-1})-(x_{ijt}-x_{ijt-1}))^2}{v_{xij}} + \frac{((\widehat{y}_{ijt}-\widehat{y}_{ijt-1})-(y_{ijt}-y_{ijt-1}))^2}{v_{yij}} + \frac{(\widehat{x}_{ijt}-d_{ijt}\widehat{y}_{ijt})^2}{v^*_{dij}}, \tag{6}$$

and

$$\widehat{y}_{ijt} = R_{ijk}, \quad \text{for all} \quad i,j \ \text{ and } \quad t = 3, k = 1. \tag{7}$$

Note that the quarterly unemployment numbers, $R_{ijt}$, are included in the model as fixed values in the hard constraint (7). These numbers are not specified as free variables, because these figures were considered as highly reliable reference figures from which the final estimates should not deviate. This strict assumption may be relaxed and the hard constraint in (7) will then become a soft constraint:

$$\widehat{y}_{ijt} \sim R_{ijk}, \tag{8}$$

with some reliability weight for the UOR figures.

The model defined here is quite simple. Extensions that allow for constraints with different functional forms, inequality constraints and some missing data are described in Bikker et al. (2010).

# 4. Results

In order to solve the optimization problem given by (6) - (7) for the figures given in Tables 1 and 2, we first need to define the variances $v_{xij}, v_{yij}, v^*_{dij}$. Let us assume that all variances $v_{xij}, v_{yij}, v^*_{dij}$ 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 UO (the numbers between parenthesis) in March are indeed consistent with the UOR figures (as depicted in Table 2).

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

| 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 $d_{ijt}$, 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 UO and the number of unemployed people at each time period. The explanation is that number of persons registered at UO in the survey in month 3

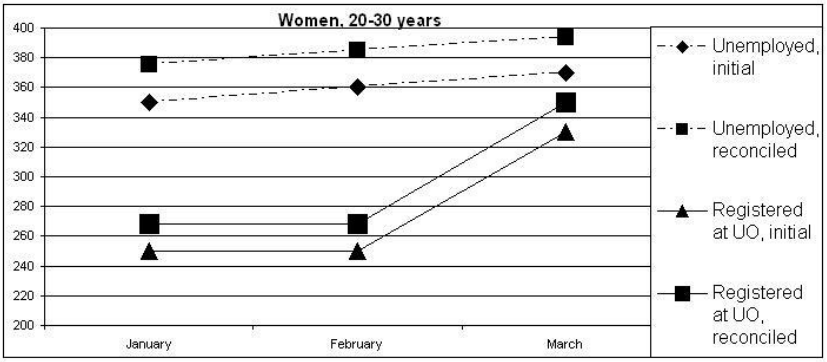


Figure 1: Women of 20-29 years; initial and reconciled figures

is much smaller than the corresponding register figure of the first quarter. Since the survey figures have to match exactly the register figures 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 UO 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 the number of persons registered at UO people at each time period.

Table 4: Women of 20-29 years; ratio's $x_{ijt}/y_{ijt}$ and $\widehat{x}_{ijt}/\widehat{y}_{ijt}$

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

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 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 UO, which becomes clear by comparing the results in Table 5 with the results in Table 4.

Table 5: Women of 20-29 year; ratio's $x_{ijt}/y_{ijt}$ and $\widehat{x}_{ijt}/\widehat{y}_{ijt}$, scenario 2

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

## 5. Conclusions

The use of register data at Statistics Netherlands (SN) has increased greatly over the last years, as is the quality of the 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 of these data sources creates the possibility for reconciliation of survey data with register data. The research reported here has shown ways to combine multiple sources in order to produce a single set of figures.

Macro-integration has some advantages over micro integration for making consistent statistics from multiple sources. By aggregating the data, the number of figures that have to be made consistent decreases and as a consequence a smaller reconciliation problem is obtained. Anomalies in the micro data may cancel out on a macro level and data linkage problems at the micro level will be avoided. The correction for differences in variable definition, population coverage and reporting periods may be more easily achieved at the macro level than at the micro level.

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

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

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

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

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