Solving large data consistency problems at Statistics Netherlands using macro integration techniques

N. Mushkudiani, J. Daalmans and R. Bikker
Statistics Netherlands

22 July 2015

Outline

- 1 The model
- 2 Supply and use tables
- 3 Short term and structural business statistics
- 4 Census 2011
- 5 Labour force survey

Goal: reconciliation of macro figures; come in the form of large multi-dimensional tabulations, obtained from different sources. Combining different data at macro level, while taking all possible relations between variables into account, is the main objective. Combining different data sources also makes it possible to detect and correct flaws in data and to improve the accuracy of estimates.

Traditionally macro-integration techniques have been extensively applied in the area of macro-economics; compilation of the National Accounts, to adjust supply and use tables to new margins (see, e.g. Stone 1941).

Application of macro-integration techniques at SN:

- Supply and use tables;
- Short term statistics figures;
- Census data;

Denton 1971, Di Fonzo and Marini 2003, 2005, Bikker and Buijtenhek 2006, In Bikker, Daalmans and Mushkudiani, 2013 we developed generalized multivariate Denton method:

Goal:

Reconciliation of macro figures of large multi-dimensional tabulations obtained from different sources,
taking all possible relations between variables into account

Properties of generalized multivariate Denton model:

- Movement preservation principle; avoid a step problem;
- Take all possible relations between variables into account
- Include ratio constraints
- Include reliability weights
- Include soft and hard constraints

Objective function for the optimization model

Data x_{it} , ($i = 1, \dots, N$, $t = 1, \dots, T$) as N time series, each of length T . the total number of the variables x_{it} is $N \cdot T$

Find adjusted values $\hat{x}_{it}$:

w_{it}^2 denotes the variance of the i^{th} time series at time t

$$\begin{aligned}
& \sum_{i=1}^N \sum_{t=2}^T A_{it} \left(\frac{(\hat{X}_{it} - x_{it}) - (\hat{X}_{it-1} - x_{it-1})}{w_{it}} \right)^2 \\
& + \sum_{i=1}^N \sum_{t=2}^T (1 - A_{it}) \left(\frac{1}{w_{it}} \left(\frac{\hat{X}_{it}}{x_{it}} - \frac{\hat{X}_{it-1}}{x_{it-1}} \right) \right)^2 \\
& + \sum_{r=1}^C \frac{1}{w_r} \left(b_r - \sum_{i=1}^N \sum_{t=1}^T c_{rit} \hat{X}_{it} \right)^2 + \sum_{n,d=1}^N \sum_{t=1}^T \frac{(\hat{X}_{nt} - v_{ndt} \hat{X}_{dt})^2}{w_{ndt}^*}
\end{aligned}$$

We minimize this function over all $\hat{X}_{it}$ satisfying the constraints

$$\sum_{i=1}^N \sum_{t=1}^T c_{rit} \hat{X}_{it} = b_r, \quad r = 1, \dots, C,$$

$$\sum_{i=1}^N \sum_{t=1}^T a_{rit} \hat{X}_{it} \leq z_r, \quad r = 1, \dots, I,$$

Software tool XPRESS and R

This tool is used for:

- Matching quarterly and annual National Accounts; it replaced the informal methods that were used before;
- Plans for applying for reconciliation of STS statistics.
- Macro-integration method is used for estimating the population and housing census (see Daalmans 2014, Mushkudiani, et al 2014);

Basic Supply-Use reconciliation

Statistics Netherlands uses a semi-automated process for the reconciliation of SU tables in current and previous year's prices simultaneously.

Stone's model, extended with ratio constraints and inequality constraints.

The problems we solve typically consist of about 34.000 variables for quarterly SU tables and about 80.000 for the final estimates of the annual SU tables. This difference arises in the classifications of goods and services.

We use a set of subjective reliability weights $A, \dots, G$ that determine the amount a variable can deviate from its input value.

Special care must be given to parts of the SU tables that must align with counterparts in the Institutional Sector Accounts. This applies especially to the government sector and the financial institutions. Specific constraints should be defined.

Basic Supply-Use Benchmarking

The annual SU-tables of the last three years are revised each year. That leads to a small time series of twelve quarterly SU-tables that must be realigned with the annual figures. We combined the annual estimation and quarterly benchmarking processes into a single process, where partial new information on annuals is combined directly with the benchmarked quarterly series of the previous year's production cycle.

Also process of a manual correction before applying the model is carried out. The problem size is very large. The SU tables consist of about 17.000 nonzero entries, multiplied by three price levels and 12 quarters, leading to a problem with about 600.000 variables.

Figure : Reconciliation of twelve consecutive quarters

ESA revisions of SU

About every ten years a revision of the accounting rules.

Estimate a time series of annual SU tables according to the new definitions.

- Apply revision corrections independently to so called benchmark years, which are 8 to 10 years apart.
- For these years we use the automated reconciliation process.
- For the intervening years we apply our benchmarking model, this model will generally spread the changes in growth rates over all intervening years.
- Price relations are defined by ratio's and all price levels are benchmarked simultaneously.

Also large benchmarking problem. With a block of ten years, 17.000 nonzero table entries, in three price levels each, about 510.000 variables.

Apart from the three machine mentioned above, Statistics Netherlands uses several other machines:

- Gross fixed capital formation
- Reconciliation of the Institutional Sector Accounts
- Benchmarking of the Institutional Sector Accounts
- A machine for updating the taxes and subsidies on production and imports, trade margins and transport margins divided over branches of industry and products
- National Production index (IO tables per product type)
- Other machines for ESA revisions of time series

Redesigning production process of business statistics; Resulted in getting high quality quarterly turnover figures based on VAT data and questionnaire data for very complex businesses.
On the other hand we have monthly figures of turnover based on survey data.

Goal: Develop a method for reconciling monthly figures with quarterly figures.

Assume quarterly figures to be reliable and fixed.

Figure : Monthly turnover figures for household appliances manufacture

Original Series

We want to find the estimates of the series $I_{m,i}^T$, $I_{m,i}^D$ and $I_{m,i}^F$ such that:

$$\min_{\widehat{I}^T \widehat{I}^D \widehat{I}^F} \sum_{i=2}^9 \frac{((\widehat{I}_{m,i}^T - \widehat{I}_{m,i-1}^T) - (I_{m,i}^T - I_{m,i-1}^T))^2}{v_T} \quad (1)$$

$$+ \frac{((\widehat{I}_{m,i}^D - \widehat{I}_{m,i-1}^D) - (I_{m,i}^D - I_{m,i-1}^D))^2}{v_D}$$

$$+ \frac{((\widehat{I}_{m,i}^F - \widehat{I}_{m,i-1}^F) - (I_{m,i}^F - I_{m,i-1}^F))^2}{v_F}.$$

Here v_T denotes the weight of the series $I_{m,i}^T$.

The following relationship should hold:

$$\hat{l}_{m,i}^T = 0.375 \cdot \hat{l}_{m,i}^D + 0.625 \cdot \hat{l}_{m,i}^F, \text{ for } i = 1, \dots, 9. \quad (2)$$

In addition all indices should be benchmarked:

$$I_{q,k}^A = \frac{1}{3} \sum_{i=3(k-1)+1}^{3k} \hat{l}_{m,i}^A, \text{ for } k = 1, \dots, 3, A \in \{T, D, F\}. \quad (3)$$

Figure : 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.

We suggested to use the full macro-integration method.

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

Census 2011 had to produce 60 multi-dimensional cross-tabulations about demographics and occupation.

For the whole Dutch population, for each province and for each municipality.

Large number of hypercubes.

Data from many different sources and different structures are combined.

Mainly from the GBA (population register), also quite a few other sources (sample surveys and registers) are used, for example the labour force survey (LFS).

Each table consists of up to 10 variables. Most of the variables are included in many cross-tabulation. All tables and all common marginals should be consistent with each other.

Many inconsistencies;

The cause of this varies: different sources, differences in population coverage, different time periods of data collection, nonresponse correction method.

The method of repeated weighting (see Houbiers 2004) is used to combine variables from different sources and to make them consistent. Using repeated weighting, tables are reconciled in steps. 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 . When one table can not be estimates consistently directly the consistency is achieved by readjusting the table weights.

2001 weighted least square method, 2011 iterative proportional fitting.

Issues:

- Computational problems; IPF;
- Edit rules; Extending table containing edit variables; extended epsilon method;
- Zero-cell problem; extended epsilon method;
- Conflicting marginals; less detailed method; different order; merging tables;
- Order dependence; splitting up method;

Macro-integration method similar to Mushkudiani 2014 can be an alternative method;

Other applications of macro-integration methods to achieve consistency of high dimensional cross-tabulation:

Two sources Labour force survey and Tax office register data
Two sets of tables of jobs and wages and number of working persons.

Different population coverage, difference in definitions; time periods.
Results in large differences.